The Power of Two in Queue Scheduling

Yan Chen, Jing Dong
Columbia University, New York, NY 10027

Utilizing customer service time information, we study a scheduling policy with two priority classes. In the single-server heavy-traffic regime, our scheduling policy achieves similar scaling for the queue length processes as the shortest remaining processing time first policy. Our analysis quantifies on how the tail of the service time distribution will affect the benefit one can gain from smart scheduling policies. When the service time information is estimated/predicted, we further analyze how prediction error will affect the performance our scheduling policy. Our results provide insights on the interplay between the service time distribution and the estimation error distribution on system performance. The imperfect information analysis also demonstrates the robustness of our scheduling policy.

Key words: Priority queue, Scheduling, Heavy-traffic diffusion limit

1. Introduction

In this paper, we propose a two-class priority scheduling policy that achieves essentially the same scaling as the shortest remaining process time first (SRPT) policy in the single-server heavy-traffic asymptotic regime. The two classes are defined by a single threshold where customers whose service time is below the threshold get higher priority. When the service time is unbounded, by properly scaling up the threshold as the traffic intensity of the system, \( \rho \), approaches one, the queue length process scales as \( o(1/(1-\rho)) \).

In a single server queue, it is known that the SRPT scheduling policy is optimal with respect to minimizing the steady-state average sojourn time and number of customers in the system (by Little’s law) (Schrage and Miler 1966). Later, Schrage (1968) also showed that SRPT minimizes the number of customers in the system at each point in time. Our result indicates that most of the benefit of SRPT can be captured by smartly dividing the customers into two priority classes when the traffic intensity is high, which is also when smart scheduling is needed the most. Notwithstanding the superiority of SPRT, it requires detailed job size information. In this respect, our policy has the advantage of requiring less job size information and easier to manage, i.e. we only need to manage two first-come-first-served (FCFS) priority queues.

We start by providing some intuition behind our development. In the heavy-traffic asymptotic regime, under the FCFS scheduling, the queue scales as \( 1/(1-\rho) \), while under the SRPT scheduling, when the service time has infinite support, the queue scales as \( o(1/(1-\rho)) \) as \( \rho \to 1 \) (see for example Lin et al. (2011) and Puha (2015)). This indicates that SRPT provides order-of-magnitude improvement over FCFS. To see how this is achieved, we notice that if we divide customers into
different priority classes according to their service time, i.e. the longer the service time, the lower the priority is, then the SRPT policy can be viewed as the limit of a sequence multi-class priority scheduling policies where we send the number of classes to infinity (Schrage and Miler 1966). Now, for multi-class queues in heavy-traffic, the queue under proper scaling only contains the jobs in the lowest priority class in the limit (see for example Reiman (1984), Van Mieghem (1995)). This result is known as a special case of the state space collapse. On the other hand, under the work-conserving policy, the workload scales as $1/(1 - \rho)$. Thus, if the average service time of the lowest priority class goes to infinity as $\rho$ approaches 1, then the queue size scales smaller than $1/(1 - \rho)$.

We make two important observations from the above intuition. First, we only need to focus on how to define the lowest priority class. Second, the average service time of the lowest priority class need to go to infinity as the traffic intensity approaches 1. The challenge here is that existing state space collapse result requires the class division to be fixed (does not change with the traffic intensity). When the class division is fixed, the queue actually scales as $1/(1 - \rho)$. Thus, it is a priori not clear when we change the class division with the traffic intensity of the systems, whether the property that “the scaled queue only contains jobs in the lowest priority class” remains true. Moreover, if we send the threshold for the lowest priority class to infinity as the traffic intensity approaches 1, we no longer know what would be the right scaling for the queue length process. The main technical contribution of this paper is providing rigorous analysis that addresses these challenges. The policy we propose requires sending the threshold for the low priority class to infinity as the traffic intensity approaches 1. We characterized the appropriate scaling and derive the diffusion limit of the queue length process under that scaling. The diffusion limit is a reflected Brownian motion with constant drift. Thus, its stationary distribution can be easily characterized. We also establish the interchange-of-limits result.

For the SRPT policy, the scaling depends on the service time distribution, its tail behavior in particular. Similarly, the scaling for our two-class scheduling policy also depends on the service time distribution. Specifically, the service time distribution will affect how we choose the threshold, which in turn affect the average service time of the low priority class. We show that with the same average service time, systems with heavier service time distributions benefit more (achieve order of magnitude smaller queues) from our scheduling policy.

In real applications, the service time requirements (job sizes) may not be exactly known to us. However, we may have statistical models to estimate/predict their values. We thus analyze the effect of prediction error on the performance of our scheduling policy. We focus on two types of errors: one is a classification error, the other is independent and identically distributed (iid) estimation error. There’re three main insights from our analysis. First, we characterize cases where we are still able to achieve the $o(1/(1 - \rho))$ scaling for the queue length process. These cases are
fairly general, indicating that our scheduling policy is quite robust to prediction errors. Second, for the classification error, it is more important not to wrongly classify the high priority class customers into low priority class than the other way around. Third, for the estimation error, the scaling depends on the tail distribution of the error term relative to the tail distribution of the service time. Achieving the $o(1/(1 - \rho))$ scaling in this case relying on the error distribution having a lighter tail than the service time distribution.

As a sanity check, we provide extensive numerical experiments to illustrate the performance our two-class priority scheduling policy as the traffic intensity increases. The numerical experiments provide important insights on the pre-limit performance of our policy. In particular, we compare the pre-limit performance of our scheduling policy to various other scheduling policies studied in the literature. We also look into the differences between preemption and non-preemption, the impact of the tail of the service time distribution, the effect of the variance of the estimation errors, etc. Lastly, we study the performance our two-class scheduling policy in the multi-server setting through simulation.

We conclude this section with a brief review of the literature. The goal is to put our work in the right context.

1.1. Related Literature

With the development of statistical learning techniques and growing availability of data, we are able to get more and more customer side information. From the queueing perspective, one valuable information is size of the job (service time) (Gans et al. 2010, Ibrahim et al. 2016)). With job-size information, smart scheduling can be made possible (see recent development in Emadi et al. (2019)).

The literature on scheduling using customer service time information has a long history. Schrage and Miler (1966) is among the first to study the SRPT policy and various other scheduling polices for the $M/G/1$ queue. As these scheduling policies require perfect job-size information, we see most applications in scheduling of jobs for machines, and computers in particular (Harchol-Balter et al. 2003). In addition to SRPT, popular scheduling policies include processor sharing (PS), shortest job first (SJF), forground-background etc. Recently, Scully et al. (2018) develope a unified framework to analyze these scheduling policies. Relatively few works look at the imperfect information setting (Wierman and Nuyens 2008, Dell’Amico et al. 2014). Our work contributes to this line of work by proposing a simple two-class priority-based scheduling policy that achieves superior performance and is robust to inaccurate job-size information.

Our result builds on heavy traffic asymptotic analysis. Process level asymptotic analysis of job-size based scheduling policies is very challenging, as it often requires keeping track of the remaining
processing time of each job in the system. Gromoll (2004) establishes the diffusion limit for PS queues. Gromoll et al. (2011) and Puha (2015) develop the diffusion limit for processes related to SRPT queues. Similar to these works, we establish the diffusion limit of the queue length processes for our scheduling policy. Our scaling parameter is similar to that established in Puha (2015). The heavy-traffic limit of mean response time of the SRPT policy is studied in Lin et al. (2011). The main insight from this line of work is that the performance of the SRPT policy depends heavily on the tail property of the service time distribution. Our work confirms this insight. In addition, we also study the interplay between the tail distribution of the actual service time and the tail distribution of the prediction error.

Our work is also closely related to the line of research on scheduling/prioritizing policies for single server queues with multiple classes of customers. Cox and Smith (1961) is among the earliest to analyze the optimality of an important class of index-based scheduling policies, known as $c\mu$-rule, where $c$ is the per unit holding cost per customer and $\mu$ is the service rate (the inverse of the mean service time). Under the $c\mu$-rule, we would give priority go the class with a larger value of $c\mu$. Van Mieghem (1995) extend the $c\mu$-rule to systems with convex waiting cost, and established the asymptotic optimality of a generalized $c\mu$-rule in the single-server heavy-traffic regime. Mandelbaum and Stolyar (2004) further extend the generalized $c\mu$-rule to multiple types of servers in addition to multiple class of customers under heavy-traffic. The main insight from this line of work is that under the diffusion scaling, the queue only contains jobs in the lowest priority class. In our model, we can think of everyone having the same linear holding cost. Thus, $c\mu$-rule reduces to the one where we’re prioritizing the class with a shorter average service time. We derive similar state-space collapse results, i.e. only the job with the lower priority class waits. The main difference of our work from the literature is that our classes are not fixed. The lower priority class is shrinking as the traffic intensity increases. The diffusion scaling we apply is also smaller than the ones in the $c\mu$-rule literature.

As only the jobs in the lowest priority class wait, there is certain fairness issue with the job-size based scheduling policies (Wierman and Harchol-Balter 2003). We do not study the fairness issue in this paper but it would be an interesting direction for future research. In recent years, there’re also works studying $c\mu$-rule under imperfect information (Argon and Ziya 2009, Sun et al. 2018). Our work expands this line of literature. In particular, we show how the tail distribution of the prediction error together with the tail distribution of the service time will affect the scaling of the queue.

It is also worth pointing out that the optimality of scheduling policies with simple structures, such as SRPT or $c\mu$-rule, has only been established for single server queues. In the asymptotic
sense, we can extend the optimality results to multiple servers under the conventional heavy-traffic regime where the number of servers is held fixed. For example, Grosof et al. (2018) recently establish the optimality of SRPT policy for multiserver queues in the conventional heavy-traffic regime. In many-server heavy-traffic regime where the number of servers is send to infinity with the arrival rate, the asymptotic optimality of $c\mu$ type of scheduling policies may no longer hold (see for example Harrison and Zeevi (2004)).

There is also a line of literature studying scheduling policies using customer patience information (Wein 1991, Mandelbaum and Momcilovic 2017). The general rule of thumb is to give priority to customers that are least patient. When patience information is not known, only distributional information is known, Bassamboo and Randhawa (2015) study a scheduling rule based on customers’ waiting history. When combining with service times and holding costs, there is also an index-based scheduling policy known as the $c\mu/\theta$ rule, where $\theta$ is the abandonment rate. When patience time follows an exponential distribution, the asymptotic optimality of this policy is established in the many-server overloaded regime (Atar et al. 2010). We note that with abandonment, the optimal scheduling policy can be highly nontrivial and depends on the patience time distribution (Puha and Ward 2019). A lot of the scheduling policies utilizing customer patience time information requires the system to operate in the overloaded regime, where waiting time is substantial. There, fluid approximation is often employed. We also note that when the system is critically loaded, the optimal diffusion control may no longer follow the simple $c\mu/\theta$ rule (Kim and Ward 2013, Kim et al. 2018). In this paper, we do not consider abandonment behavior, and our system is in the critically loaded regime where diffusion approximation is employed.

2. Problem formulation and main results

We consider a sequence of $M/GI/1$ queues indexed by $n$, starting from empty at time 0. The service time distribution is fixed with mean $1/\mu = 1$. We denote $F$ as the cumulative distribution function (cdf) of the service time, and $\bar{F}$ as its tail CDF. We assume the service time distribution is continuous and denote $f$ as its probability density function (pdf). The arrival rate for the $n$-th system is $\lambda_n = 1 - \beta/\sqrt{n}$ for some $\beta > 0$. This implies that the traffic intensity of the sequence of systems approaches 1 at rate $1/\sqrt{n}$, i.e.

$$\rho^n := \lambda^n/\mu = 1 - \beta/\sqrt{n}.$$ 

We comment that our results hold as long as $\sqrt{n}(1 - \rho_n) \to \beta$ for some $\beta > 0$ as $n \to \infty$. We choose this specific form of $\lambda_n$ and $\mu$ to keep the exposition concise.

Let $A^n = \{A^n(t) : t \geq 0\}$ denote the arrival process of the $n$-th system. We also write $v^n(k)$ as the service time of the $k$-th arrival in system $n$. As the service time does not change for different scales of systems, we shall write $v$ as a generic random variable following the service time distribution $F$. 

We impose the following assumption on $v$.

**Assumption 1.** There exists $C > 0$, such that $f(x) > 0$ for any $x > C$. There exists $\delta > 0$, such that $E[v^{2+\delta}] < \infty$.

The first part of Assumption 1 essentially requires that the service time distribution has infinite support (unbounded). We know from the literature of SRPT that when the service time has bounded support, the queue scales as $1/(1 - \rho)$ (Lin et al. 2011). In this case, there’s no order-of-magnitude gain from “smart” scheduling. The moment condition in Assumption 1 is standard to establish the diffusion limit of queueing models (see for example Whitt (2002)). Under Assumption 1, we denote $\sigma^2 := \text{Var}(v)$.

Throughout the paper, we refer to waiting time as total time spent in the system. This include both the time spent in queue and the time spent in service, and is also referred to as the sojourn time or flow time in the literature. We refer to the number of jobs in the queue as all the customers in the system. We use the words jobs and customers interchangeably. We also define $\eta(t) := t$ and $\zeta(t) := 0$ for all $t \geq 0$.

### 2.1. A scheduling policy with two classes

We consider a scheduling policy where we divide customers into two classes based on their service time requirement. For the $n$-th system, we define $K^n$ as a threshold value. How we choose $K^n$ will be specified later. Customers with service time less than or equal to $K^n$ is classified into the higher priority class, class 1; Customer with service time greater than $K^n$ is classified into the lower priority class, class 2. For simplicity of exposition, we assume class 1 customers have preemptive priority over class 2 customers.

We start by introducing a few more notations. Denote $\lambda^n_1 = \lambda^n F(K^n)$, $\lambda^n_2 = \lambda^n \bar{F}(K^n)$ as the arrival rates of Class 1 and Class 2 customers in the $n$-th system. We also write

$$\mu^n_1 = \left( \frac{1}{F(K^n)} \int_0^{K^n} xf(x)dx \right)^{-1}, \quad \mu^n_2 = \left( \frac{1}{F(K^n)} \int_{K^n}^{\infty} xf(x)dx \right)^{-1}$$

as the service rates of the two classes respectively, and $\rho^n_i = \lambda^n_i / \mu^n_i$ as the traffic intensity for Class $i$, $i = 1, 2$. Let $A^n_i = \{ A^n_i(t) : t \geq 0 \}$ denote the arrival process of Class $i$ customers and $v^n_i(k)$ denote the service time of the $k$-th Class $i$ arrival in system $n$. Define $Q^n_i = \{ Q^n_i(t) : t \geq 0 \}$ as the queue length process of Class $i$, and $Q^n = Q^n_1 + Q^n_2$ as the total queue length. We also define

$$V^n(t) = \sum_{k=1}^{A^n_i(t)} v^n_i(k) - t$$

$$U^n(t) = V^n(t) - \inf_{0 \leq s \leq t} V^n(s) \wedge 0$$
\[ V_1^n(t) = \sum_{k=1}^{A_1^n(t)} v_1^n(k) - t \]
\[ U_1^n(t) = V_1^n(t) - \inf_{0 \leq s \leq t} V_1^n(s) \wedge 0 \]
\[ U_2^n(t) = U^n(t) - U_1^n(t) \]

Note that \( U^n \) is the total unfinished workload process, \( U_i^n \) is the unfinished workload process of class \( i \) jobs. To facilitate the asymptotic analysis, we define the diffusion scaled workload processes as \( \hat{U}_1^n(t) = U_1^n(nt)/\sqrt{n} \) and \( \hat{U}_2^n(t) = U_2^n(nt)/\sqrt{n} \). We also write \( \hat{A}_1^n(t) = (A_1^n(nt) - \lambda_1^n nt)/\sqrt{n} \) and \( \hat{A}_2^n(t) = (A_2^n(nt) - \lambda_2^n nt)/\sqrt{\lambda_2^n n} \) as the scaled arrival processes.

Before we present the main results, we shall introduce the general idea underlying our development. From the existing results for priority queues (Reiman 1984), we make the following three important observations:

1. The Class 1 queue scales as \((1 - \rho_1^n)^{-1}\). Note that if Class 1 customers have preemptive priority over Class 2 customers, then Class 1 queue evolves as if there is only Class 1 customers in the system.

2. The total workload process scales as \((1 - \rho^n)^{-1} = O(\sqrt{n})\). Indeed, this is true for any non-idling scheduling policies.

3. Under proper scaling, the limiting queue contains only Class 2 jobs (the lower priority class).

If the “limiting” queue only contains Class 2 jobs, then the queue length process scales as \( \mu_2^n \sqrt{n} \). To achieve a smaller scaling than \( \sqrt{n} \), we need \( \mu_2^n \to 0 \) as \( n \to \infty \). However, this need to be done with delicacy. Specifically, we need to make sure that under the scaling \( \mu_2^n \sqrt{n} \), we still see no Class 1 customers in the queue, i.e. \( \mu_2^n \sqrt{n} \) is of a larger scale than \((1 - \rho_1^n)^{-1}\).

Above all, the key is to choose the threshold \( K^n \) in an appropriate way. Our choice of \( K^n \) satisfies the following assumption.

**Assumption 2.** \( K^n \to \infty \), and there exists \( \delta \in (0, 1/2) \), such that \( n^{1/2 - \delta} \bar{F}(K^n) \to \infty \) as \( n \to \infty \).

We next provide some comments about Assumption 2. Let \( \gamma^n := K^n \mu_2^n \). Under Assumption 1, \( 1 \leq \gamma^n \leq 2 \). Then Assumption 2 implies that \( \mu_2^n \to 0 \) as \( n \to \infty \). We also notice that

\[ \rho_1^n = \lambda_1^n/\mu_1^n = \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^{K^n} x f(x) dx = 1 - \frac{\bar{F}(K^n)}{\mu_2^n} + O(1/\sqrt{n}), \]

Thus, under Assumption 2, there exists \( \delta \in (0, 1/2) \), such that

\[ \frac{n^{1/2 - \delta}}{K^n} (1 - \rho_1^n) \to \infty \text{ as } n \to \infty, \quad (1) \]
or equivalently,
\[ n^{1/2-\delta} \mu_2^n (1 - \rho^n) \to \infty \text{ as } n \to \infty, \]

Under work-conserving service policy and Assumption 1, it has been shown that (see for example Reiman (1984))
\[ \hat{U}^n \Rightarrow \text{RBM}(-\beta, 1 + \sigma^2) \text{ in } D[0, \infty) \text{ as } n \to \infty, \]

where \( \text{RBM}(-\beta, 1 + \sigma^2) \) denote a one-dimensional reflected Brownian motion with drift \(-\beta\) and variance \(1 + \sigma^2\). The following theorem is the main result of this paper. It shows that under the two-class priority scheduling policy with a proper choice of \( K^n \), the scaled unfinished workload will contain only Class 2 jobs. More importantly, the queue length process will scale as \( \sqrt{n}/K^n \) asymptotically.

Let \( \gamma := \lim_{n \to \infty} K^n \mu_2^n. \) Define \( \hat{Q} = \{\hat{Q}(t) : t \geq 0\} \) a reflected Brownian motion with drift \(-\gamma\beta\) and variance \(\gamma^2(1 + \sigma^2)\), i.e. \( \hat{Q} = \gamma \text{RBM}(-\beta, 1 + \sigma^2) \).

**Theorem 1.** Under Assumption 1 and 2,
\[ \frac{K^n}{\sqrt{n}} Q^n_1(nt) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty \]

and
\[ \frac{K^n}{\sqrt{n}} Q^n_2(nt) \Rightarrow \hat{Q}(t) \text{ in } D[0, \infty) \text{ as } n \to \infty. \]

Theorem 1 implies that
\[ \frac{K^n}{\sqrt{n}} Q^n(nt) \Rightarrow \hat{Q}(t) \text{ in } D[0, \infty) \text{ as } n \to \infty. \]

In addition, as \( \hat{\mu}_2^n K^n \to \gamma \) as \( n \to \infty \), we also have
\[ \frac{1}{\sqrt{n} \hat{\mu}_2^n} Q^n(nt) \Rightarrow \text{RMB}(-\beta, 1 + \sigma^2) \text{ in } D[0, \infty) \text{ as } n \to \infty. \]

This alternative expression for the scaling of the queue length process may be more natural in some settings.

From Assumption 2 and Theorem 1, we note that as \( K^n \) depends on the service time distribution, the scaling also depends on the service time distribution. In particular, the heavier the tail of the service time distribution, the smaller the queue scales under our two-class priority scheduling policy with a properly chosen \( K^n \). For example, when the service time follows Pareto distribution, \( \bar{F}(x) = \left(\frac{m}{x}\right)^\alpha \) for \( \alpha > 2 \), we can set \( K^n = mn^{(1/2-\delta)/\alpha} \). We observe that the smaller \( \alpha \) is, the heavier the tail of \( \bar{F} \) is, and the larger \( K^n \) (order of magnitude) is, which in turn implies that the smaller the queue would be. This is a very appealing property. Note that under the FCFS policy, the steady-state average queue length of an \( M/G/1 \) queue with \( \mu = 1 \) takes the form
\[ \rho + \frac{\lambda^2}{1 - \rho} \frac{1 + \sigma^2}{2}. \]

(3)
From (3), the heavier the tail of the service time distribution is, the larger the value of $\sigma^2$ is, and hence the larger the queue would be. While when applying smart scheduling here, the heavier tails actually lead to smaller queues.

We denote $\hat{Q}(\infty)$ as the stationary distribution of $\hat{Q}$. Then $\hat{Q}(\infty)$ follows an exponential distribution with rate $2\beta/(\gamma(1+\sigma^2))$ (Harrison and Williams 1987). We also establish the interchange-of-limits result for the queueing length process.

**Theorem 2.** Under Assumption 1 and 2, for any $x \in \mathbb{R}$,

$$\lim_{n \to \infty} \lim_{t \to \infty} P\left(\frac{K^n}{\sqrt{n}} Q^n(nt) \leq x\right) = \lim_{t \to \infty} \lim_{n \to \infty} P\left(\frac{K^n}{\sqrt{n}} Q^n(nt) \leq x\right) = 1 - \exp\left(-\frac{2\beta}{\gamma(1+\sigma^2)} x\right).$$

The proof of Theorem 1 and 2 is delayed to Section 3, where we also present several other results about the asymptotic behavior of the workload processes and virtual waiting time processes for the two priority classes.

2.2. Comparison to existing results

In this section, we compare our main results to results established in the literature. The purpose is twofold. First, it serves as a sanity check of our results. Second, it establishes the near optimality of our scheduling policy in the asymptotic sense.

For $M/G/1$ queues, we compare five scheduling policies. The first three of them focus on two-class priority queues with threshold $K^n$ for the $n$-th system and the last two are shortest job size based policies. In particular, the first one is *two-class nonpreemptive*. The second one is *two-class preemptive*, where the priority of the job only depends on its initial size. The third one is *two-class semi-preemptive*, where the priority of the job depends on its remaining size, i.e. if the remaining size of a job is smaller than $K^n$, then it is classified as a Class 1 job; otherwise, it is classified as a Class 2 job. The fourth one is shortest-job-first (SJF) where the job size refers to its initial size and the policy is nonpreemptive. The fifth one is SRPT. We first summarize results for the stationary average queue length under these scheduling policies. The detailed derivations of these results can be found in Schrage and Miler (1966).

We use $Q^n(\infty; i)$ to denote the generic stationary number of customers in system under scheduling policy $i$. Under the two-class nonpreemptive scheduling,

$$E[Q^n(\infty; 1)] = \rho^n + \frac{\lambda^n \lambda_1^n E[v^2]}{2(1-\rho_1^n)} + \frac{\lambda^n \lambda_2^n E[v^2]}{2(1-\rho_1^n)(1-\rho_2^n)}. \tag{4}$$

Under the two-class preemptive scheduling,

$$E[Q^n(\infty; 2)] = \rho_1^n + \frac{\rho_2^n}{1-\rho_1^n} + \frac{\lambda_1^{n,2} E[v_1^{n,2}]}{2(1-\rho_1^n)} + \frac{\lambda_2^n E[v^2]}{2(1-\rho_1^n)(1-\rho_2^n)}. \tag{5}$$
Under the two-class semi-preemptive scheduling,
\[
E[Q^n(\infty; 3)] = \rho^n_1 + \frac{\rho^n_2}{1 - \rho^n_2} - \frac{\lambda^n \rho^n_1 K^n}{1 - \rho^n_1} + \frac{\lambda^n \rho^n_2 K^{n,2}}{2(1 - \rho^n_2)} + \frac{\lambda^n \rho^n_1 E[v^n_1]}{2(1 - \rho^n_1)} + \frac{\lambda^n \rho^n_2 E[v^n_2]}{2(1 - \rho^n_2)(1 - \rho^n)}.
\]
Under the SJF scheduling,
\[
E[Q^n(\infty; 4)] = \rho^n + \frac{\lambda^n E[v^n]}{2} \int_0^\infty \frac{1}{(1 - \lambda^n \int_0^y ydF(y))^2} dF(x).
\]
Under the SRPT scheduling,
\[
E[Q^n(\infty; 5)] = \lambda^n \int_0^\infty \int_0^x \frac{1 - \lambda^n \int_0^y ydF(u)^2} {1 - \lambda^n \int_0^y ydF(y)} dydF(x) + \frac{\lambda^n E[v^n]}{2} \int_0^\infty \int_0^y ydF(y) + \bar{F}(x) x^2 \int_0^\infty \frac{1 - \lambda^n \int_0^y ydF(y)} (1 - \lambda^n \int_0^y ydF(y))^2 dF(x).
\]

For the first three scheduling policies, \(i = 1, 2, 3\), under Assumption 2, simple calculation reveals that
\[
\frac{K^n}{\sqrt{n}} E[Q^n(\infty; i)] \to \frac{\gamma(1 + \sigma^2)}{2\beta} \quad \text{as} \quad n \to \infty,
\]
or equivalently
\[
\frac{1}{\sqrt{n\mu^2_2}} E[Q^n(\infty; i)] \to \frac{(1 + \sigma^2)}{2\beta} \quad \text{as} \quad n \to \infty.
\]
The limit is the same as the expectation of \(\hat{Q}(\infty)\) derived in Theorem 2. It also indicates that the first three scheduling policies achieve the same asymptotic performance, i.e. preemption doesn’t matter in the limit.

Next, we notice that
\[
E[Q^n(\infty; 1)] \geq E[Q^n(\infty; 4)] \geq E[Q^n(\infty; 5)]
\]
We shall analyze the difference between them under different service time distributions in the heavy-traffic limit. To utilize results established in the literature, we consider two classes of service time distributions: Weibull and Pareto.

If the service time follows a Weibull distribution, \(\tilde{F}(x) = \exp(-(\xi x)^\alpha)\), Lin et al. (2011) establish that
\[
\frac{\log(\sqrt{n})}{\sqrt{n}} E[Q^n(\infty; 5)] \to \frac{\xi(1 + \sigma^2)}{2\beta} \quad \text{as} \quad n \to \infty.
\]
In this case, we can set \(K^n = \frac{1}{\xi}(\log(n^{1/2 - \delta}))^{1/\alpha}\), for any \(\delta > 0\), and \(\gamma = 1/\xi\). Then
\[
\frac{\log(n^{1/2 - \delta})}{\sqrt{n}} E[Q^n(\infty; 1)] \to \frac{\xi(1 + \sigma^2)}{2\beta} \quad \text{as} \quad n \to \infty.
\]
As \(\delta\) can be chosen arbitrarily small, this indicates that \(E[Q^n(\infty; i)]\), for \(i = 1, 4, 5\), achieves basically the same asymptotic performance.

If the service time follows a Pareto distribution, \(\tilde{F}(x) = (\frac{m}{x})^\alpha\) for \(\alpha > 2\), Lin et al. (2011) establish that
\[
n^{-\frac{2-\alpha}{2(\alpha-1)}} E[Q^n(\infty; 5)] \to \frac{\pi/(1 - \alpha)}{2\sin(\pi/(1 - \alpha))} \frac{(1 + \sigma^2)}{2m\beta} \quad \text{as} \quad n \to \infty.
\]
In this case, we can set \( K^n = mn^{1/2(1/2 - \delta)} \), for any \( \delta > 0 \), and we have \( \gamma = (\alpha - 1)/\alpha \). Then
\[
n^{-\frac{\alpha - 1/2\delta}{2\alpha}} E[Q^n(\infty; 1)] \to \frac{(1 + \sigma^2)}{2m\beta} \text{ as } n \to \infty.
\]
In this case, even though we can send \( \delta \) arbitrarily small, \( \frac{\alpha - 1}{2\alpha} \) is still larger than \( \frac{\alpha - 2}{2(\alpha - 1)} \). Thus, \( E[Q^n(\infty; 5)] \) in this case still gain a small order of magnitude improvement over \( E[Q^n(\infty; 1)] \).

Lastly, at the process level, under the SRPT policy, Puha (2015) establishes that if the service time distribution is Weibull,
\[
\log(\sqrt{n})^{1/\alpha} Q^n_1(nt) \Rightarrow \xi \text{RBM}(-\beta, 1 + \sigma^2) \text{ as } n \to \infty,
\]
which is basically the same as what we established in Theorem 1 for our two-class scheduling policy with e.g. \( K^n = \frac{1}{\xi}(\log(n^{1/2 - \delta}))^{1/\alpha} \).

3. Proof the Theorem 1 and 2

The proof of Theorem 1 is divided into two steps. We first establish the state-space collapse result for the workload process and queue length process for Class 1 customers (Proposition 1). We then establish the proper limit for the virtual waiting time process and queue length process for Class 2 customers (Proposition 2).

Recall that \( U^n_1(t) \) is the unfinished workload process for class 1. As class 1 customers has higher preemptive priority over class 2 customers, \( U^n_1(t) \) is also the virtual waiting time process for class 1. The following proposition establishes the state space collapse of \( K^n U^n_1(nt)/\sqrt{n} \) and \( K^n Q^n_1(nt)/\sqrt{n} \).

**Proposition 1.** Under Assumption 1 and 2,
\[
\frac{K^n}{\sqrt{n}} U^n_1(nt) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty,
\]
and
\[
\frac{K^n}{\sqrt{n}} Q^n_1(nt) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]

**Proof.** To simplify the notation, let
\[
\hat{U}^n_1(t) = \frac{K^n}{\sqrt{n}} U^n_1(nt) \text{ and } \hat{Q}^n_1(t) = \frac{K^n}{\sqrt{n}} Q^n_1(nt).
\]

We first establish an upper bound for \( U^n_1(t) \) which is based on the Lindley’s recursion for single server queue. Similar idea has been used in Blanchet et al. (2018).

Let \( W^n_k \) denote the waiting time of the \( k \)-th arriving class 1 customer in the \( n \)-th system. Then
\[
\sup_{0 \leq t \leq T} U^n_1(t) \leq \sup_{1 \leq k \leq A^n_1(T)} (W^n_k + v^n_1(k)).
\]
Let $X^n_k$ denote the interarrival time between the $(k - 1)$-th and $k$-th customer. We also write $\Delta^n_k := v^n_1(k - 1) - X^n_k$. Define $S^n_0 = 0$ and $S^n_k = S^n_{k-1} + \Delta^n_k$. Then

$$W^n_{k+1} = \max \{W^n_k + \Delta^n_{k+1}, 0\} = \max_{0 \leq i \leq k+1} \{S^n_{k+1} - S^n_i\} \overset{D}{=} \max_{0 \leq i \leq k+1} S^n_i.$$ Let $M^n = \sup_{k \geq 1} S^n_k$. Then

$$P \left( \max_{1 \leq k \leq N} W^n_k > a \right) = P \left( \max_{1 \leq k \leq N} W^n_k > a \right) \leq \sum_{k=1}^{N} P(W^n_k > a) \leq N P(M^n > a)$$

Notice that

$$E[\Delta^n_k] = \frac{1}{F(K^n)} \int_0^{K^n} x f(x) dx - \frac{1}{\lambda^n F(K^n)} = \frac{1}{\lambda^n F(K^n)} (\rho^n_1 - 1) < 0.$$ Thus, $M^n$ is the all time maximum of a negative drifted random walk. As $v^n_1(k - 1) \leq K^n$, i.e. bounded, $\Delta^n_k$ has finite moment generating function in a neighborhood of the original. We write $\psi_n(\theta) := \log E[\exp(\theta \Delta^n_k)]$. As $\psi'_n(0) = E[\Delta^n] < 0$ and $\psi''_n(0) = \text{Var}(\Delta^n_k) > 0$, there exists $\theta_n > 0$ such that $\psi_n(\theta_n) = 0$. By Cramér-Lundberg Theory, we have

$$P(M^n > c) \leq \exp(\theta_n c).$$

From Corollary 3 in Glynn and Whitt (1994), we also have

$$\theta_n = \frac{1}{\text{Var}(\Delta^n_k) \lambda^n F(K^n)} (1 - \rho^n_1) + o(E[\Delta^n]), \text{ as } n \to \infty.$$ As $\text{Var}(\Delta^n_k) \to 1 + \sigma^2$ and $\lambda^n F(K^n) \to 1$ as $n \to \infty$, $\lim_{n \to \infty} \theta_n/(1 - \rho^n_1) = 2/(1 + \sigma^2)$.

For any $\delta > 0$, set

$$H^n := \min \{k \in \mathbb{Z} : P(A^n_1(nT) > k) \leq \delta/2\},$$

i.e. $H^n$ as smallest integer that is larger than the $\delta/2$-th upper quantile of $A^n_1(nT)$. As $A^n_1(nT) \sim \text{Poisson}(\lambda^n nT)$, $H^n = \Theta(n)$. In addition, from (1), we can pick $N(\epsilon, \delta)$ large enough, such that for any $n \geq N(\epsilon, \delta)$,

$$H^n \exp \left(-\frac{\sqrt{n} \theta_n \epsilon}{K_n}\right) < \delta/2.$$ Then for $n > N(\epsilon, \delta)$,

$$P \left( \sup_{0 \leq t \leq T} \hat{U}^n(t) > \epsilon \right) \leq P \left( \max_{1 \leq k \leq A^n_1(nT)} \frac{K^n}{\sqrt{n}} (W^n_k + v^n_1(k)) > \epsilon \right) \leq P(A^n_1(nT) > H^n) + P \left( \max_{1 \leq k \leq H^n} \frac{K^n}{\sqrt{n}} (W^n_k + v^n_1(k)) > \epsilon \right) \leq P(A^n_1(nT) > H^n) + P \left( \max_{1 \leq k \leq H^n} \frac{K^n}{\sqrt{n}} v^n_1(k) > \epsilon/2 \right) + H^n P \left( \frac{K^n}{\sqrt{n}} M^n > \epsilon/2 \right) \leq \frac{\delta}{2} + H^n \exp \left(-\frac{\sqrt{n} \theta_n \epsilon}{K_n}\right) < \delta.
This concludes the proof of state-space collapse of \( K^n U^n_t(\sqrt{nt})/\sqrt{n} \).

We next establish the state-space collapse of \( K^n Q^n_t(\sqrt{nt})/\sqrt{n} \). It follows similar lines of arguments as the proof of Theorem 4 in Section 3.2 of Reiman (1984). However, there are several extra technical difficulties we need to address along the way. We denote \( a^n_t(t) \) as the arrival time of the class 1 customer that is in service at time \( t \). If there’s no class 1 customer in service, we set \( a^n_t(t) = t \). We also write \( \bar{a}^n_t(t) = a^n_t(nt)/n \).

As

\[
A^n_t(t) - A^n_t(a^n_t(t)) \leq Q^n_t(t) \leq A^n_t(t) - A^n_t(a^n_t(t)) + 1,
\]

\[
K^n \left( \hat{A}^n_t(t) - \hat{A}^n_t(\bar{a}^n_t(t)) \right) + K^n \sqrt{n} \lambda^n_1(t - \bar{a}^n_t(t))
\]

\[
\leq \hat{Q}^n_t(t) \leq K^n \left( \hat{A}^n_t(t) - \hat{A}^n_t(\bar{a}^n_t(t)) \right) + K^n \sqrt{n} \lambda^n_1(t - \bar{a}^n_t(t)) + K^n/\sqrt{n}.
\]

We also notice that as

\[
U^n_t(a^n_t(t)) \leq t - a^n_t(t) \leq U^n_t(a^n_t(t)) + v^n_t(A^n_t(a^n_t(t))),
\]

\[
\hat{U}^n_t(\bar{a}^n_t(t)) \leq K^n \sqrt{n}(t - \bar{a}^n_t(t)) \leq \hat{U}^n_t(\bar{a}^n_t(t)) + \frac{K^n}{\sqrt{n}} v^n_t(A^n_t(a^n_t(nt))).
\]

Under Assumption 1 and 2,

\[
\frac{K^n}{\sqrt{n}} v^n_t(A^n_t(u^n_t(nt))) \leq \frac{K^{n,2}}{\sqrt{n}} \to 0 \text{ as } n \to \infty.
\]

Thus,

\[
\sup_{0 \leq t \leq T} \left| K^n \sqrt{n}(t - \bar{a}^n_t(t)) - \hat{U}^n_t(\bar{a}^n_t(t)) \right| \to 0 \text{ as } n \to \infty. \quad (10)
\]

This also implies that \( \bar{a}^n_t(t) \Rightarrow t \) as \( n \to \infty \).

Based on the above analysis we have

\[
\sup_{0 \leq t \leq T} \left| \hat{Q}^n_t(t) - \hat{U}^n_t(t) \right| \leq \sup_{0 \leq t \leq T} K^n \left[ \hat{A}^n_t(t) - \hat{A}^n_t(\bar{a}^n_t(t)) \right] + \sup_{0 \leq t \leq T} \left| \lambda^n_1 K^n \sqrt{n}(t - \bar{a}^n_t(t)) - \hat{U}^n_t(\bar{a}^n_t(t)) \right| + \sup_{0 \leq t \leq T} \left| \hat{U}^n_t(\bar{a}^n_t(t)) - \hat{U}^n_t(t) \right| + \frac{K^n}{\sqrt{n}}.
\]

We next analyze (A), (B) and (C) one by one. For (A), let \( \delta_n = \frac{\epsilon}{4(n-1) K^n} \), then

\[
P \left( \sup_{0 \leq t \leq T} K^n \left| \hat{A}^n_t(t) - \hat{A}^n_t(\bar{a}^n_t(t)) \right| > \epsilon \right)
\]
\[ P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| A^n_i(nt) - \lambda^n_i(nt) - (A^n_i(n\bar{a}^n_i(t)) - \lambda^n_i n\bar{a}^n_i(t)) \right| > \epsilon \right) \]
\[ \leq P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| A^n_i(nt) - A^1_i(n\bar{a}^1_i(t)) \right| + \left| \lambda^n_i nt - \lambda^1_i n\bar{a}^1_i(t) \right| > \epsilon \right) \]
\[ \leq P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| A^n_i(nt) - A^1_i(n\bar{a}^1_i(t)) \right| > \frac{\epsilon}{2} \right) + P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| nt - n\bar{a}^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ \leq P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| A^n_i(nt) - A^1_i(n\bar{a}^1_i(t)) \right| > \frac{\epsilon}{2} ; \sup_{0 \leq t \leq T} \left| nt - n\bar{a}^1_i(t) \right| > \delta_n \right) \]
\[ + P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| A^n_i(nt) - A^1_i(n\bar{a}^1_i(t)) \right| > \frac{\epsilon}{2} ; \sup_{0 \leq t \leq T} \left| nt - \bar{a}^1_i(t) \right| \leq \delta_n \right) \]
\[ + P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| nt - n\bar{a}^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ \leq P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| nt - n\bar{a}^1_i(t) \right| > \frac{K^n}{\sqrt{n}} \delta_n \right) + P \left( \sup_{0 \leq t \leq n T} \frac{K^n}{\sqrt{n}} \left| A^1_i(t + \delta_n) - A^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ + P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| nt - n\bar{a}^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ \leq P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| nt - n\bar{a}^1_i(t) \right| > \frac{\epsilon}{4(\epsilon - 1)} \right) + P \left( \sup_{0 \leq t \leq n T} \frac{K^n}{\sqrt{n}} \left| A^1_i(t + \delta_n) - A^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ + P \left( \sup_{0 \leq t \leq T} \frac{K^n}{\sqrt{n}} \left| nt - n\bar{a}^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ (a) \]
\[ (b) \]
\[ (c) \]

Form (10), we have \( \frac{K^n}{\sqrt{n}}(nt - n\bar{a}^1_i(t)) \Rightarrow \zeta(t) \) as \( n \to \infty \). Thus, both (a) and (c) goes to zero as \( n \to \infty \). For (b), let \( Z_i, i = 1, 2, \ldots \), denote i.i.d. Poisson random variables with rate \( \delta_n \). Then we can write
\[ P \left( \sup_{0 \leq t \leq n T} \frac{K^n}{\sqrt{n}} \left| A^1_i(t + \delta_n) - A^1_i(t) \right| > \frac{\epsilon}{2} \right) \]
\[ \leq P \left( \sup_{0 \leq t \leq n T} \frac{K^n}{\sqrt{n}} Z_i > \frac{\epsilon}{4} \right) \]
\[ \leq \frac{n T}{\delta_n} \exp \left( \delta_n (e - 1) - \frac{\epsilon \sqrt{n}}{2 K^n} \right) \] by Chernoff's bound
\[ = \frac{n T}{\delta_n} \exp \left( - \frac{\epsilon \sqrt{n}}{4 K^n} \right) \to 0 \text{ as } n \to \infty. \]

Thus (A) converges to zero as \( n \to \infty \).

For (B), under Assumption 2, \( \lambda^n_i \to 1 \) as \( n \to \infty \). Then by (10),
\[ \sup_{0 \leq t \leq T} \left| \lambda^n_i K^n \sqrt{n}(t - \bar{a}^n_i(t)) - \hat{U}^n_i(\bar{a}^n_i(t)) \right| \to 0 \text{ as } n \to \infty. \]

For (C), as \( \hat{U}^n_i \Rightarrow \zeta \) and \( \bar{a}^n_i \Rightarrow \eta \) as \( n \to \infty \),
\[ \sup_{0 \leq t \leq T} \left| \hat{U}^n_i(\bar{a}^n_i(t)) - \hat{U}^n_i(t) \right| \to 0 \text{ as } n \to \infty. \]
This completes the proof of the second part of Proposition 1.

Proposition 1 implies that the virtual waiting time and queue length of Class 1 queue scales slower than \( \sqrt{n}/K^n = O(\sqrt{n} \mu_2^2) \). Moreover, from Proposition 1, we have

\[
\hat{U}_2^n - \hat{U}^n \Rightarrow \zeta \text{ in } D[0, \infty) \text{ as } n \to \infty,
\]

and thus,

\[
\hat{U}_2^n \Rightarrow \text{RBM}(-\beta, (1 + \sigma^2)) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]

For Class 2 customers, as it has lower priority, \( U_2^n(t) \) is smaller than its virtual waiting time at \( t \). To characterize the waiting time of Class 2 customers, following Reiman (1984), we define

\[
B^n(t, U^n) := \inf\{s \geq 0 : V_1^n(t + s) - V_1^n(t) + U^n \leq 0\}.
\]

Notice that \( B^n(t, U^n(t)) \) is virtual waiting time process for Class 2. We also define

\[
\hat{B}^n(t) := B(nt, U^n(nt))/\sqrt{n}.
\]

**Proposition 2.** Under Assumption 1 and 2,

\[
\rho_2^n \hat{B}^n(t) - \hat{U}^n(t) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty,
\]

and

\[
\frac{K^n}{\sqrt{n}} Q_2^n(nt) \Rightarrow \hat{Q}(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]

**Proof.** The proof of the first part of the proposition follows from the proof of Theorem 2 in Section 4 of Reiman (1984). However, here we have \( \rho_1^n \to 1 \text{ as } n \to \infty. \) Define

\[
\hat{V}_1^n(t) = \frac{1}{\sqrt{n}} (V_1^n(nt) - nt(1 - \rho_1^n))
\]

Then

\[
\hat{B}^n(t) := \inf\{s \geq 0 : \hat{V}_1^n(t + s/\sqrt{n}) - \hat{V}_1^n(t) + \hat{U}^n(t) \leq (1 - \rho_1^n)s\}.
\]

Under Assumption 1,

\[
\hat{V}_1^n \Rightarrow \text{BM}(0, 1 + \sigma^2) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]

This implies that

\[
P \left( \sup_{0 \leq t \leq T} \sup_{0 \leq s \leq n^{1/2 - \delta/2}/K^n} |\hat{V}_1^n(t + s/\sqrt{n}) - \hat{V}_1^n(t)| \leq \epsilon \right) \to 1 \text{ as } n \to \infty.
\]
where $\delta$ is chose according to (1), i.e. $(1 - \rho^n_\alpha)n^{1/2-\delta}/K^n \to \infty$ as $n \to \infty$. We also notice that
\[
\sup_{0 \leq s \leq n^{1/2-\delta}/K^n} |\hat{V}_n^H(t + s/\sqrt{n}) - \hat{V}_1(t)| \leq \epsilon
\]
implies that if $B^n(t) \leq n^{1/2-\delta}/K^n$,
\[
\sup_{0 \leq t \leq T} |(1 - \rho^n_\alpha)\hat{B}^n(t) - \hat{U}^n(t)| \leq \epsilon;
\]
if $B^n(t) > n^{1/2-\delta}/K^n$,
\[
\hat{U}^n(t) > (1 - \rho^n_\alpha)n^{\delta/2}n^{1/2-\delta}/K^n.
\]
As $\hat{U}^n \Rightarrow \text{RMB}(-\beta, 1 + \sigma^2)$ as $n \to \infty$ and (1),
\[
P\left(\hat{U}^n(t) > n^{\delta/2}(1 - \rho^n_\alpha)n^{1/2-\delta}/K^n\right) \to 0 \text{ as } n \to \infty.
\]
Thus,
\[
P\left(\sup_{0 \leq t \leq T} |(1 - \rho^n_\alpha)\hat{B}^n(t) - \hat{U}^n(t)| \leq \epsilon\right) \to 1 \text{ as } n \to \infty.
\]
In addition, as $\rho^n_\alpha/(1 - \rho^n_\alpha) \to 1$ as $n \to \infty$, we have
\[
\rho^n_\alpha\hat{B}^n(t) - \hat{U}^n(t) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]
We now prove the second part. Let $a^n_\alpha(t)$ denote the time of the oldest class 2 customer in the system at time $t$. We also write $\bar{a}^n_\alpha(t) = a^n_\alpha(nt)/n$. Then
\[
A^n_2(t) - A^n_2(a^n_\alpha(t)) \leq Q^n_2(t) \leq A^n_2(t) - A^n_2(a^n_\alpha(t)) + 1
\]
and
\[
B^n(a^n_\alpha(t), U_n(a^n_\alpha(t))-) \leq t - a^n_\alpha(t) \leq B^n(a^n_\alpha(t), U_n(a^n_\alpha(t))).
\]
Then
\[
\rho^n_\alpha\hat{B}^n(\bar{a}^n_\alpha(t))- \leq \rho^n_\alpha\sqrt{n}(t - \bar{a}^n_\alpha(t)) \leq \rho^n_\alpha\hat{B}^n(\bar{a}^n_\alpha(t)).
\]
From the first part of the proposition, we have $\rho^n_\alpha\hat{B}_n(t) \Rightarrow \text{RBM}(-\beta, 1 + \sigma^2)$. Therefore,
\[
\sup_{0 \leq t \leq T} \left|\rho^n_\alpha\hat{B}^n(\bar{a}^n_\alpha(t)) - \rho^n_\alpha\sqrt{n}(t - \bar{a}^n_\alpha(t))\right| \Rightarrow 0 \text{ as } n \to \infty. \tag{11}
\]
We also have $\bar{a}^n_\alpha \Rightarrow \eta$ in $D[0, \infty)$ as $n \to \infty$.

Let $\hat{Q}^n_2(t) = \frac{K^n}{\sqrt{n}}Q^n_2(nt)$. Then
\[
K^n\sqrt{\lambda^n_2} \left(\hat{A}^n_2(t) - \hat{A}^n_2(a^n_\alpha(t))\right) + \frac{\gamma^n\lambda^n_2}{\sqrt{n}}(t - a^n_\alpha(t))
\]
\[
\leq \hat{Q}^n_2(t) \leq K^n\sqrt{\lambda^n_2} \left(\hat{A}^n_2(t) - \hat{A}^n_2(a^n_\alpha(t))\right) + \frac{\gamma^n\lambda^n_2}{\sqrt{n}}(t - a^n_\alpha(t)) + \frac{1}{\sqrt{n}\mu_2(n)}.
\]
Recall that $\gamma^n = K^n \mu_2^n$. Thus,

$$
\sup_{0 \leq t \leq T} \left| \hat{Q}_n^2(t) - \gamma^n \hat{U}_n^2(t) \right| \leq K^n \sqrt{\frac{\lambda_2^n}{2}} \sup_{0 \leq t \leq T} \left| \hat{A}_2^n(t) - \hat{A}_2^n(\tilde{a}_2^n(t)) \right|

+ \sup_{0 \leq t \leq T} \gamma^n \left| \rho_2^n \hat{B}_n(\tilde{a}_2^n(t)) - \rho_2^n \sqrt{\bar{a}_2^n(t)} \right|

+ \gamma^n \sup_{0 \leq t \leq T} \left| \rho_2^n \hat{A}_n(t) - \frac{\gamma^n}{\gamma^n} \hat{A}_n(t) \right|

\quad + \sup_{0 \leq t \leq T} \left| \rho_2^n \hat{B}_n(\tilde{a}_2^n(t)) - \frac{\gamma^n}{\gamma^n} \hat{U}_n(t) \right|

\quad + \sup_{0 \leq t \leq T} \left| \rho_2^n \hat{A}_n(t) - \frac{\gamma^n}{\gamma^n} \hat{A}_n(t) \right|

\quad + \sup_{0 \leq t \leq T} \left| \rho_2^n \hat{B}_n(\tilde{a}_2^n(t)) - \frac{\gamma^n}{\gamma^n} \hat{U}_n(t) \right|

\quad + \frac{\gamma^n}{\gamma^n} \sup_{0 \leq t \leq T} \left| \rho_2^n \hat{B}_n(\tilde{a}_2^n(t)) - \frac{\gamma^n}{\gamma^n} \hat{U}_n(t) \right|

\quad \Rightarrow 0 \quad \text{as} \quad n \to \infty.
$$

For part (I), under Assumption 1 and 2, $K^n \sqrt{\frac{\lambda_2^n}{2}} \to 0$ as $n \to \infty$. We also have $\hat{A}_2^n \Rightarrow \text{BM}$ as $n \to \infty$. Then as $\tilde{a}_2^n \Rightarrow \eta$,

$$
K^n \sqrt{\frac{\lambda_2^n}{2}} \sup_{0 \leq t \leq T} \left| \hat{A}_2^n(t) - \hat{A}_2^n(\tilde{a}_2^n(t)) \right| \Rightarrow 0 \quad \text{as} \quad n \to \infty.
$$

As $\gamma^n \to \gamma$ as $n \to \infty$, the weak convergence to 0 of part (II) follows from (11), and the convergence of part (III) follows from the first part of Proposition 2. We have thus complete the proof of the second part of Proposition 2. □

Proposition 2 indicates that the virtual waiting time process Class 2 queue scales as

$$
\sqrt{n} \rho_2^n = O(\sqrt{n} K^n \tilde{F}(K^n)).
$$

And the queue length process for Class 2 queue scales as $K^n / \sqrt{n} = O(\sqrt{n} \mu_2^n)$. Combining Proposition 1 and Proposition 2, we have proved Theorem 1.

**Proof of Theorem 2** From Theorem 1, we have $\lim_{n \to \infty} \lim_{t \to \infty} K^n \sqrt{\frac{\lambda_2^n}{2}} Q^n(nt)$ has the same distribution as the stationary distribution of $\text{RBM}(-\gamma \beta, \gamma^2 (1 + \sigma^2))$, i.e. $\hat{Q}(\infty)$, which follows an exponential distribution with rate $2\beta / (\gamma (1 + \sigma^2))$.

For the $n$-th system, as $\rho^n < 1$ and the scheduling policy is non-idling, the stationary distribution of $Q^n(t)$ is well defined. Particularly, under the two-class preemptive priority rule, the steady-state queue length satisfies (Schrage and Miler 1966)

$$
E[Q^n(\infty)] = \rho^n_1 \frac{\rho_2^n}{1 - \rho_1^n} + \frac{\lambda_2^n \gamma^2 E[v_2^n]}{2(1 - \rho_1^n)} + \frac{\lambda^n \lambda_2^n \gamma^2 E[v]}{2(1 - \rho_1^n)(1 - \rho^n)}
$$

As $\frac{\sqrt{n}}{K^n} (1 - \rho_1^n) \to \infty$, $\sqrt{n}(1 - \rho^n) \to \beta$ and $K^n \mu_2^n \to \gamma$ as $n \to \infty$, we have

$$
\frac{K^n}{\sqrt{n}} E[Q^n(\infty)] = \frac{\gamma(1 + \sigma^2)}{2\beta} \text{ as } n \to \infty.
$$

This implies $\frac{K^n}{\sqrt{n}} Q^n(\infty)$ is tight. The interchange of limit result then follows from Gamarnik and Zeevi (2006). □
4. Imperfect service time information

In many real-life applications, especially in service systems, we may not have perfect information about the service time of arriving customers. However, with the advancement of statistical learning techniques and the availability of data, we may have high-quality predictions of their service times. In this section, we study how the quality of the prediction affects the performance of our scheduling policy. We consider two types of prediction models. One is a classification model; the other is predicting the actual service times with i.i.d. estimation errors.

4.1. Classification errors

We assume there is a classifier that can help predict which class an arriving customer belongs to. In this case, there are two types of errors: classifying a Class 1 customer as Class 2, and classifying a Class 2 customer as Class 1. In particular, for the $n$-th system, we denote

\[ p_{11}^n := P(\text{Classified as Class 1} | \text{Class 1}), \quad p_{12}^n := P(\text{Classified as Class 2} | \text{Class 1}), \]

\[ p_{22}^n := P(\text{Classified as Class 2} | \text{Class 2}), \quad p_{21}^n := P(\text{Classified as Class 1} | \text{Class 2}). \]

We next study our scheduling policy under the classification errors.

Recall that the real Class 1 customers are those whose service times are shorter than or equal to $K^n$ for the $n$-th system, and the real Class 2 customers are those with service times larger than $K^n$. For customers who are classified as Class $i$, we write $\hat{\lambda}_i^n$ as their arrival rate, $1/\hat{\mu}_i^n$ as their average service time, and $\hat{\rho}_i^n = \hat{\lambda}_i^n / \hat{\mu}_i^n$ as the corresponding traffic intensity. Then

\[ \hat{\lambda}_i^n = p_{11}^n \lambda_1^n + p_{12}^n \lambda_2^n, \]

\[ \hat{\mu}_i^n = \left( \frac{p_{11}^n \lambda_1^n / \mu_1^n + p_{21}^n \lambda_2^n / \mu_2^n}{p_{11}^n \lambda_1^n + p_{12}^n \lambda_2^n} \right)^{-1}. \]

We first note that the perfect information case can be viewed as a special case of our classification error setting with $p_{12}^n = 0$ and $p_{22}^n = 1$. In what follows, we consider the non-trivial cases where $p_{12}^n > 0$ or $p_{22}^n < 1$. Based on our discussion in Section 2, if the queue for the classified Class 1 customers is almost negligible, the size of the classified Class 2 queue scales as $\sqrt{n} \hat{\mu}_2^n$. To achieve a smaller scaling than $\sqrt{n}$ here, we need $\hat{\mu}_2^n \to 0$ as $n \to \infty$. Now, for the classified Class 1 queue to be negligible, we also need $n^{1/2-\delta} \hat{\mu}_2^n (1 - \hat{\rho}_1^n) \to \infty$ as $n \to \infty$. Based on these considerations and the analysis in Theorem 1, we having the following theorem that summarizes the cases where we can still achieve $o(\sqrt{n})$ scaling.

**Theorem 3.** Under Assumption 1 and 2,

(a) if $p_{12}^n = 0$ or $\bar{F}(K^n) p_{22}^n / p_{12}^n \to \infty$, and $p_{22}^n \bar{F}(K^n) n^{1/2-\delta} \to \infty$ as $n \to \infty$ for some $\delta > 0$ then

\[ \frac{K^n}{\sqrt{n}} \hat{Q}^n(nt) \Rightarrow \hat{Q}(t) \text{ in } D[0, \infty) \text{ as } n \to \infty; \]
(b) if \( \bar{F}(K^n)p_{22}^n/p_{12}^n \to c \in (0, \infty) \), and \( p_{22}^n \bar{F}(K^n)n^{1/2-\delta} \to \infty \) as \( n \to \infty \) for some \( \delta > 0 \), then

\[
\frac{K^n}{\sqrt{n}}Q^n(nt) \to \frac{1+c}{c} \hat{Q}(t) \text{ in } D[0, \infty) \text{ as } n \to \infty;
\]

(c) if \( \bar{F}(K^n)p_{22}^n/p_{12}^n \to 0 \) but \( K^n \bar{F}(K^n)p_{22}^n/p_{12}^n \to \infty \), and \( p_{12}^n n^{1/2-\delta} \to \infty \) as \( n \to \infty \) for some \( \delta > 0 \), then

\[
\frac{K^n \bar{F}(K^n)p_{22}^n/p_{12}^n}{\sqrt{n}}Q^n(nt) \to \hat{Q}(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]

Before we present the proof of Theorem 3, we first make a few important observations here. First, to achieve \( o(\sqrt{n}) \) scaling, \( p_{12}^n \) needs to decay to 0 sufficiently fast. As for \( p_{22}^n \), somewhat surprisingly, if \( p_{12}^n \) decays to zero fast enough, we do not have many restrictions for \( p_{22}^n \). For example, we even allow \( p_{22}^n \) to decay to zero as well, as long as it doesn’t decay too fast. Above all, it is very important not to wrongly classify Class 1 customers into Class 2, while we can be more tolerate about classifying Class 2 customers into Class 1. To see this, note that we need \( \lim_{n \to \infty} p_{12}^n = 0 \), while \( \inf_n p_{21}^n \) can be strictly larger than 0. Second, case (a) and (b) in Theorem 3 has the same scaling as the perfect information case, suggesting that the queue length processes in these two cases are of the same order as in the perfect information case. However, if we compare the constant terms, the steady-state average queue length in case (a) is smaller than that in case (b). In case (c), as \( \bar{F}(K^n)p_{22}^n/p_{12}^n \to 0 \), the scaling is larger than \( \sqrt{n}/K^n \), implying that the queue length is of larger order in this case than the perfect information case. Lastly, we comment that the condition that \( p_{22}^n \bar{F}(K^n)n^{1/2-\delta} \to \infty \) in case (a) and (b), and the condition that \( p_{12}^n n^{1/2-\delta} \to \infty \) in case (c) are to ensure that the Class 1 queue vanishes under the appropriate scaling. The other conditions under case (a), (b) and (c) are to ensure that \( \hat{\mu}_2^n \to 0 \) as \( n \to \infty \) accordingly. As we will see in the proof of Theorem 3, these conditions are quite tight, i.e. when they do not hold, we’re unlikely to achieve \( o(\sqrt{n}) \) scaling.

Proof of Theorem 3. We start by analyzing \( \hat{\mu}_2^n \) under the three cases. In particular, we shall start by checking if \( \hat{\mu}_2^n \to 0 \) as \( n \to \infty \), and how \( \hat{\mu}_2^n \) compares to \( \mu_2^n \). First notice that if \( p_{12}^n = 0 \), \( \hat{\mu}_2^n = \mu_2^n \). If \( p_{12}^n > 0 \),

\[
\frac{1}{\hat{\mu}_2^n} = \frac{p_{12}^n \rho_1^n + p_{22}^n \rho_2^n}{p_{12}^n \lambda_1^n + p_{22}^n \lambda_2^n} = \frac{\rho_1^n + p_{22}^n \lambda_2^n \bar{F}(K^n)/\mu_2^n}{\lambda_1^n + p_{22}^n \lambda_2^n \bar{F}(K^n)}.
\]

We also notice that \( \rho_1^n \) and \( \lambda_1^n \) converge to 1 as \( n \to \infty \). In addition, \( \mu_2^n \to 0 \) as \( n \to \infty \). Thus, in case (a), when \( p_{12}^n > 0 \) and \( \bar{F}(K^n)p_{22}^n/p_{12}^n \to \infty \), we have

\[
\mu_2^n / \hat{\mu}_2^n \to 1 \text{ as } n \to \infty.
\]

In case (b), \( \bar{F}(K^n)p_{22}^n/p_{12}^n \to c \in (0, \infty) \). In this case,

\[
\mu_2^n / \hat{\mu}_2^n \to c/(1+c) \text{ as } n \to \infty.
\]
In case (c), \( \bar{F}(K^n)p_{22}^n / p_{12}^n \to 0 \) but \( K^n \bar{F}(K^n)p_{22}^n / p_{12}^n \to \infty \). In this case, we have

\[
\frac{\mu_2^n p_{12}^n}{\bar{F}(K^n)p_{22}^n \mu_2^n} \to 1 \text{ as } n \to \infty.
\]

We particularly notice that as \( \bar{F}(K^n)p_{22}^n / p_{12}^n \to 0, \) \( \mu_2^n \gg \mu_2^n \). Lastly, note that case (a), (b), and (c) in the above analysis cover all the cases where \( \mu_2^n \to 0 \) as \( n \to \infty \).

We next show that in case (a) and (b), if \( p_{22}^n \bar{F}(K^n) n^{1/2-\delta} \to \infty, (1 - \hat{\rho}_1^n)n^{1/2-\delta} \bar{\mu}_2^n \to \infty \) as \( n \to \infty \), i.e. if the classified Class 1 queue is negligible under the scaling \( \sqrt{n} \mu_2^n \). In the analysis, we also observe the tightness of this condition. First notice that

\[
1 - \hat{\rho}_1^n = 1 - p_{11}^n \rho_1^n - p_{21}^n \rho_2^n = 1 - \lambda^n + \rho^n - p_{11}^n \lambda^n \bar{F}(K^n)/\mu_1^n - p_{21}^n \lambda^n \bar{F}(K^n)/\mu_2^n = \beta/\sqrt{n} + \lambda^n (p_{22}^n - (p_{22}^n - p_{12}^n)\bar{F}(K^n)/\mu_2^n).
\]

We only need to focus on the second term in (12) for subsequent analysis. In particular, we define

\[
\Xi := p_{12}^n + (p_{22}^n - p_{12}^n)\bar{F}(K^n)/\mu_2^n.
\]

We further rearrange \( \Xi n^{1/2-\delta} \bar{\mu}_2^n \) as

\[
\left( p_{12}^n + (p_{22}^n - p_{12}^n) \frac{\bar{F}(K^n)}{\mu_2^n} \right) n^{1/2-\delta} \bar{\mu}_2^n = p_{12}^n n^{1/2-\delta} (\bar{\mu}_2^n - \bar{F}(K^n)) + p_{22}^n \bar{F}(K^n) n^{1/2-\delta} \frac{\hat{\mu}_2^n}{\mu_2^n}.
\]

In case (a), as \( \hat{\mu}_2^n / \mu_2^n \to 1 \), if \( p_{22}^n \bar{F}(K^n) n^{1/2-\delta} \to \infty, \)

\[
\Xi n^{1/2-\delta} \bar{\mu}_2^n \to \infty \text{ as } n \to \infty.
\]

As for the tightness of the condition, we notice that when \( p_{12}^n > 0 \), as \( \bar{F}(K^n) p_{22}^n / p_{12}^n \to \infty \), \( p_{22}^n \bar{F}(K^n) \gg p_{12}^n \).

In case (b), as \( \bar{\mu}_2^n / \mu_2^n \to (1 + c) / c \), we have that if \( p_{22}^n \bar{F}(K^n) n^{1/2-\delta} \to \infty, (14) \) holds. Moreover, we notice that, as \( \bar{F}(K^n) p_{22}^n / p_{12}^n \to c, p_{12}^n \) and \( p_{22}^n \bar{F}(K^n) \) is of the same order. This implies the tightness of the condition.

Similarly, for case (c), we shall show that if \( p_{12}^n n^{1/2-\delta} \to \infty, (14) \) holds. For this case, we need to rearrange \( \Xi n^{1/2-\delta} \bar{\mu}_2^n \) as

\[
\left( p_{12}^n + (p_{22}^n - p_{12}^n) \frac{\bar{F}(K^n)}{\mu_2^n} \right) n^{1/2-\delta} \bar{\mu}_2^n = p_{12}^n n^{1/2-\delta} (\hat{\mu}_2^n - \bar{F}(K^n)) + n^{1/2-\delta} p_{22}^n \frac{\bar{F}(K^n)}{\mu_2^n} \frac{\mu_2^n p_{12}^n}{\bar{F}(K^n) p_{22}^n} n^{1/2-\delta} \frac{\hat{\mu}_2^n \bar{F}(K^n) p_{22}^n}{\mu_2^n p_{12}^n}.
\]
As \( \hat{\mu}_n^2 \bar{F}(K^n)p_{22}^n \rightarrow 1 \), we have that when \( p_{12}^n n^{1/2-\delta} \rightarrow \infty \), (14) holds. We also notice that as \( \hat{\mu}_2^n - \bar{F}(K^n) \rightarrow 0 \) as \( n \rightarrow \infty \), this condition is tight here. Lastly, we comment that in case (c), \( \bar{F}(K^n)p_{22}^n/p_{12}^n \rightarrow 0 \), implying that \( \bar{F}(K^n)p_{22}^n \ll p_{12}^n \). This is why the second condition in case (c) is different from the second condition in case (a) and (b).

The above analysis basically verifies the conditions under which the classified Class 1 queue has state-space collapse and the classified Class 2 customers scales slower than \( \sqrt{n} \). The rest of proof follows exactly the same line of argument as the proof of Theorem 1. We shall thus omit it here.

4.2. Estimation Errors

In this section, we study the effect of iid estimation errors on the performance of our two-class scheduling rule. Specifically, we assume the estimated job size, \( \hat{v} \), takes the form

\[
\hat{v} = v + \epsilon,
\]

where \( v \) is the actual job size, and \( \epsilon \) is an error term with mean zero and is independent of the \( v \). We denote \( \Phi \) as the cdf of \( \epsilon \), \( \bar{\Phi} \) as its tail cdf, and \( \phi \) as its pdf, assuming \( \epsilon \) has continuous distribution.

For the \( n \)-th system, if the estimated job size is larger than \( K^n \), we assign the job to Class 1, the high priority class. Otherwise, i.e. if \( \hat{v} > K^n \), we assign the job to Class 2, the low priority class. Following the notations in Section 4.1, we define \( \hat{\lambda}_i^n, \hat{\mu}_i^n \), and \( \hat{\rho}_i^n : = \lambda_i^n / \mu_i^n \) as the arrival rate, service rate, and traffic intensity, respectively, of the classified Class \( i \), \( i = 1, 2 \), jobs for the \( n \)-th system. Then we have

\[
\lambda_1^n = \lambda^n P(v + \epsilon < K^n) = \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^\infty f(t)\Phi(K^n - t)dt;
\]
\[
\lambda_2^n = \lambda^n P(v + \epsilon \geq K^n) = \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^\infty f(t)\bar{\Phi}(K^n - t)dt;
\]
\[
1/\hat{\mu}_1^n = E[v|v + \epsilon < K^n] = \frac{\int_0^\infty tf(t)\Phi(K^n - t)dt}{\int_0^\infty f(t)\Phi(K^n - t)dt};
\]
\[
1/\hat{\mu}_2^n = E[v|v + \epsilon \geq K^n] = \frac{\int_0^\infty tf(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}.
\]

Now, following the analysis in Section 2, for the Class 2 queue to scale as \( o(\sqrt{n}) \), we need \( \hat{\mu}_2^n \rightarrow 0 \) as \( n \rightarrow \infty \). For the Class 1 queue to be negligible under the scaling \( \sqrt{n}\hat{\mu}_2^n \), we need

\[
n^{1/2-\delta}\hat{\mu}_2^n (1 - \hat{\rho}_1^n) \rightarrow \infty \text{ as } n \rightarrow \infty \text{ for some } \delta > 0. \tag{15}
\]

We notice that as

\[
1 - \hat{\rho}_1^n = 1 - \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^\infty tf(t)\Phi(K^n - t)dt.
\]
\[
= \int_0^\infty tf(t)\Phi(K^n - t)dt + O(1/\sqrt{n}),
\]
(15) is equivalent to requiring
\[ n^{1/2-\delta} \int_0^\infty f(t) \Phi(K^n - t) dt \to \infty \text{ as } n \to \infty \text{ for some } \delta > 0. \] (16)

We divide the analysis into two cases depending on the support of the distribution of \( \epsilon \).
1. \( \epsilon \) has support on \([-M_l, M_u]\) for some \( M_l, M_u \in (0, \infty) \).
2. \( \epsilon \) has support on \((-\infty, \infty)\).

We start with the bounded support case. The following theorem shows that we’re still able to achieve \( o(\sqrt{n}) \) scaling in this case.

**Theorem 4.** For \( \epsilon \) with bounded support, \( \hat{\mu}_n^2 \to 0 \) and
\[
\frac{1}{\sqrt{n\hat{\mu}_n^2}} Q^n(nt) \Rightarrow RMB(-\beta, 1 + \sigma^2) \text{ in } D[0, \infty) \text{ as } n \to \infty.
\]

**Proof** Assume \( \epsilon \) has support on \([-M_l, M_u]\) for some \( M_l, M_u \in (0, \infty) \). We first note that for \( K^n > M_u \),
\[
n^{1/2-\delta} \int_0^\infty f(t) \Phi(K^n - t) dt = n^{1/2-\delta} \int_{K^n-M_u}^\infty f(t) \Phi(K^n - t) dt \geq n^{1/2-\delta} \int_{K^n}^\infty f(t) \Phi(K^n - t) dt \geq n^{1/2-\delta} \Phi(0) \bar{F}(K^n) \to \infty \text{ as } n \to \infty.
\]

Also note that,
\[
\frac{1}{\hat{\mu}_n^2} = \frac{\int_{K^n-M_u}^\infty t f(t) \Phi(K^n - t) dt}{\int_{K^n-M_u}^\infty f(t) \Phi(K^n - t) dt} \geq K^n - M_u.
\]

Thus, \( \hat{\mu}_n^2 \to 0 \) as \( n \to \infty \). The rest of proof follows exactly the same line of argument as the proof of Theorem 1.

□

We next study the case when \( \epsilon \) has support on \( \mathbb{R} \). Theorem 5 establishes several convergence results in this case. As we shall explain in Theorem 5, the convergence analysis depends on the tail behavior of the service time distribution versus the tail behavior of the estimation error. To facilitate the comparison of the tails, we define
\[ I(x) := \int_0^\infty \phi(x - t) f(t) dt \]

Let
\[ h(x) = -\frac{d}{dx} \log f(x), \ l(x) = -\frac{d}{dt} \log \phi(x) \text{ and } u(x) = \phi^{-1}(f(x)). \]

We also define regular oscillation following Berman (1992).
**Definition 1.** A positive continuous function $g$ is regularly oscillating if

$$
\lim_{x,x'\to \infty, x/x' \to 1} \frac{g(x)}{g(x')} = 1.
$$

**Theorem 5.** Under Assumption 1 and 2, and suppose that $h(x)$ and $l(x)$ are regularly oscillating, and $\phi$ has support on $\mathbb{R}$,

(a) if $\limsup_{x \to \infty} h(x) < \liminf_{x \to \infty} l(x),$

or

(b) if $\lim_{x \to \infty} h(x)/l(x) = 0$ and $\lim_{x \to \infty} u(x)h(x) = \infty,$

then $\hat{\mu}_2^n \to 0$ and

$$
\frac{1}{\sqrt{n\hat{\mu}_2^n}}Q^n(nt) \Rightarrow RM B(-\beta, 1 + \sigma^2) \text{ in } D[0, \infty) \text{ as } n \to \infty.
$$

Theorem 5 provides two sufficient conditions for the queue to achieve $o(\sqrt{n})$ scaling. We notice that both conditions essentially require that the error term has a lighter tail than the service time distribution. If, on the other hand, the error term has a heavier tail, then it is very likely to wrongly classify Class 1 jobs into Class 2, which, from our analysis in Section 4.1, may fail to achieve $o(\sqrt{n})$ scaling.

**Proof of Theorem 5.** We prove the two cases separately. We shall only verify the conditions for the Class 1 queue to be negligible, i.e. (16), and the the Class 2 queue to scale as $o(\sqrt{n})$, i.e. $\hat{\mu}_2^n \to 0$. The rest of proof then follows exactly the same line of argument as the proof of Theorem 1.

**Case (a)** We first introduce some properties of $I(x)$ that was established in Berman (1992). Under the condition of Case (a), Theorem 2.1 in Berman (1992) establishes that

$$
\lim_{x \to \infty} \frac{I(x)}{f(x)\int_{-\infty}^{\infty} \exp(th(x))\phi(t)dt} = 1,
$$

and there exits $C < \infty$ such that for all large values of $x$

$$
1 \leq \int_{-\infty}^{\infty} \exp(th(x))\phi(t)dt \leq C.
$$

We would like to point out that Theorem 2.1 in Berman (1992) assume $f(x)$ has support on $(-\infty, \infty)$. However, the proof can be straightforwardly extended to the case where $f(x)$ has support on $(c, \infty)$. Based on these results, we have for $K^n$ sufficiently large

$$
\int_{0}^{\infty} f(t)\Phi(K^n - t)dt = \int_{K^n}^{\infty} I(x)dx \geq \frac{1}{2} \int_{K^n}^{\infty} f(x)dx = \frac{1}{2}\tilde{F}(K^n).
$$
Thus, (16) holds.

On the other hand, for \( K^n \) sufficiently large

\[
\hat{\mu}_2^n = \frac{\int_0^\infty f(t) \tilde{\Phi}(K^n - t) dt}{\int_0^\infty tf(t) \tilde{\Phi}(K^n - t) dt} \leq \frac{\int_{K^n} f(x) dx}{\Phi(0) \int_{K^n} tf(t) dt} \leq \frac{2C \int_{K^n} f(x) dx}{\Phi(0) \int_{K^n} tf(t) dt} = \frac{2C}{\Phi(0)} \mu_2^n.
\]

Thus, \( \hat{\mu}_2^n \to 0 \) as \( n \to \infty \).

**Case (b)** We again start by introducing some properties of \( I(x) \) that was established in Berman (1992). Under the condition of Case (b), Theorem 3.2 in Berman (1992) establishes that

\[
\limsup_{x \to \infty} \frac{I(x)}{f(x - u(x))} \leq \tilde{\Phi}(0) \quad \text{and} \quad \liminf_{x \to \infty} \frac{I(x)}{f(x)} \geq \tilde{\Phi}(0).
\]

Based on these results, we have for \( K^n \) sufficiently large,

\[
\int_0^\infty f(t) \tilde{\Phi}(K^n - t) dt = \int_{K^n} I(x) dx \geq \frac{1}{2} \tilde{\Phi}(0) \int_{K^n} f(x) dx = \frac{1}{2} \tilde{\Phi}(0) \tilde{F}(K^n).
\]

Thus, (16) holds.

We next study the asymptotic behavior of \( \hat{\mu}_2^n \). For any fixed \( y > 0 \), and \( K^n \) sufficiently large, we have

\[
\frac{1}{\hat{\mu}_2^n} = \frac{\int_0^\infty tf(t) \tilde{\Phi}(K^n - t) dt}{\int_0^\infty f(t) \tilde{\Phi}(K^n - t) dt} \geq \frac{2y \int_{2y}^\infty f(t) \tilde{\Phi}(K^n - t) dt}{\int_0^\infty f(t) \tilde{\Phi}(K^n - t) dt} \geq 2y \left( 1 - \frac{\int_0^{2y} f(t) \tilde{\Phi}(K^n - t) dt}{\int_0^\infty f(t) \tilde{\Phi}(K^n - t) dt} \right)
\]

We then notice that

\[
\frac{\int_0^{2y} f(t) \tilde{\Phi}(K^n - t) dt}{\int_0^\infty f(t) \tilde{\Phi}(K^n - t) dt} \leq \frac{\tilde{\Phi}(K^n - 2y)}{\frac{1}{2} \tilde{\Phi}(0) \tilde{F}(K^n)}.
\]

As \( \lim_{x \to \infty} h(x)/l(x) = 0 \),

\[
\frac{\tilde{\Phi}(K^n - 2y)}{\frac{1}{2} \tilde{\Phi}(0) \tilde{F}(K^n)} \to 0 \quad \text{as} \quad n \to \infty.
\]

Thus for \( K^n \) sufficiently large, \( \frac{1}{\hat{\mu}_2^n} \geq y \), or equivalently, \( \hat{\mu}_2^n \leq 1/y \). As \( y \) can be set arbitrarily large, we have \( \hat{\mu}_2^n \to 0 \) as \( n \to \infty \).

\( \square \)
5. Numerical experiments

In this section, we conduct some numerical experiments for different scheduling policies. We start with a sanity check for our two-class scheduling policy. The goal is to demonstrate the effect of different service time distributions. We then study the pre-limit performance of various scheduling policy. We highlight the difference between preemptive versus non-preemptive policies. We also conduct extensive analysis on the effect of estimation/prediction errors on the performance of our scheduling policy. Lastly, we extend the analysis to the multiserver setting.

Note that for single server queues, we already have closed-form expression for the steady-state queue length under commonly used scheduling policies (see (4) - (8)). Thus, the analysis in Section 5.1-5.3 utilize these close-from expressions and involves mostly numerical method to evaluate integrals. For the multi-server queue setting, less close-from expressions are known, we thus use Monte-Carlo simulation to approximate the performance in Section 5.4.

5.1. The effect of the service time distribution

We consider four different service time distributions: (i) Exponential, $\bar{F}(x) = \exp(-\mu x)$ with $\mu = 1$; (ii) Weibull, $\bar{F}(x) = \exp(-\xi x^\alpha)$ with $\alpha = 1.5$ and $\xi^\alpha \approx 0.86$; (iii) Weibull, $\bar{F}(x) = \exp(-\xi x^\alpha)$ with $\alpha = 2.5$ and $\xi^\alpha \approx 0.74$; (iv) Pareto, $\bar{F}(x) = (m/x)^\alpha$ with $\alpha = 3$, $m = 2/3$. Note that the four distributions have very different tail behavior: the two Weibull distributions have lighter tails than exponential, which has lighter tail than the Pareto. For (i), we set $K^n = 0.49 \log(n)$; for (ii) and (iii), we set $K^n = \frac{1}{\xi}(0.49 \log(n))^{1/\alpha}$, for (iv), we set $K^n = mn^{0.49/\alpha}$. Notice that in all four cases, we have $\bar{F}(K^n) = n^{-0.49}$. We set $\lambda_n = 1 - 1/\beta_n$. Figure 1 plots the steady-state average queue length scaled by $1/(1 - \rho^n) = \sqrt{n}$ for different values of the traffic intensity $\rho^n$, i.e. $n$, and different service time distributions.

![Figure 1](image)

We observe that when scaled by $1/(1 - \rho)$, the steady-state average queue length converges to zero as $\rho \to 1$, suggesting the queue scales as $o(1/(1 - \rho))$. We also observe that the heavier the tail
of the service time distribution, the smaller the queue is under our two-class priority scheduling policy. This indicates that systems with heavier tail service time distributions will benefit more from smart scheduling.

5.2. Different scheduling policies

In Figure 2, we compare the performance of our preemptive two-class priority policy with FCFS, SJF, and SRPT. We consider two service time distributions: one is Exponential with rate 1, the other is Pareto, \( \bar{F}(x) = \left( \frac{2}{3} x \right)^3 \). We set \( \lambda^n = 1 - 1/\sqrt{n} \) and plot the steady-state average queue length scaled by \( 1/(1 - \rho^n) = \sqrt{n} \). For our two-class priority policy, we choose \( K^n \) such that \( \bar{F}(K^n) = n^{-0.49} \).

We observe that all the smart scheduling polices are performing much better than FCFS. We also observe that although the three smart scheduling policies are similar in the limit, our two-class priority policy performs slight worse than SJF and SRPT in the pre-limit.

We next study the effect of preemption on the performance of our two class-class priority scheduling policy. To facilitate the comparison, we shall scale the queue length by \( \sqrt{n} \mu_2^n \). We know from our analysis in Section 2.2 that in the limit, the scaled steady-state queue all converge to \( (1 + \sigma^2)/2 \) (see (9)). Figure 3 compares the performance of our two-class priority scheduling policy with preemption, without preemption and with semi-preemption for different values of the traffic intensity \( \rho^n \). We show results for two service time distributions: exponential and Pareto as in Figure 2. We choose \( K^n \) such that \( \bar{F}(K^n) = n^{-0.2} \).
We observe that when scaled by \( \sqrt{n \mu_2^2} \), the steady-state average queue length indeed converges to \( (1 + \sigma^2)/2 \) as \( n \to \infty \), i.e. \( \rho \to 1 \). This is regardless of whether we allow preemption or not. However, in the pre-limit, preemption leads to better performance, i.e. shorter queue on average. We would also like to point out that preemption might be harder or more costly to implement in practice.

### 5.3. The effect of the estimation error

We consider the case with iid estimation errors. Specifically, we assume \( \epsilon \) follows iid Normal distribution with mean 0 and variance \( \tau^2 \), i.e. \( \epsilon \sim N(0, \tau^2) \). Figure 4 (a) and Figure 5 (a) plot the steady-state queue length scaled by \( \sqrt{n \hat{\mu}_2^n} \) for different values of traffic intensity \( \rho^n \), and estimation error variance \( \tau^2 \). Figure 4 (b) and Figure 5 (b) plot \( K^n \hat{\mu}_2^n \) for different values of \( \rho^n \) and \( \tau^2 \). We show results for two service times distributions: Pareto, \( \bar{F}(x) = \left(\frac{x}{x + 2}\right)^3 \) (Figure 4), which corresponds to Case (a) in Theorem 5; and Weibull, \( \bar{F}(x) = \exp(-0.86x^{1.5}) \) (Figure 5), which corresponds to Case (b) in Theorem 5.
Figure 4  Numerical comparison of the preemptive two class priority policy for Pareto service time under different variances of $\epsilon$

Figure 5  Numerical comparison of the preemptive two class priority policy for Weibull service time under different variances of $\epsilon$

Figure 4(a) and Figure 5 (a) confirm the results of Theorem 5. In particular, we see that as $\rho$ approaches 1, the scaled queue length converges to $(1 + \sigma^2)/2$. We also observe from Figure 4 (b) and Figure 5 (b) that as the estimation error increases ($\tau$ increases), the queue scales larger. This is because $\hat{\mu}_2^n$ increases with $\tau$, and so does $\sqrt{n}\hat{\mu}_2^n$.

5.4. Multiserver queues

As discussed in Section 1.1, even though we only prove the convergence results for single server queue, the same results should hold for multiserver queues under the conventional heavy-traffic limit. In this section, we extend the analysis to multiserver queues through numerical experiments.
In Table 1, we show the steady-state queue length for an M/M/10 queue under different scheduling policies. We set $\mu = 1$ and $\beta = 1$, and vary the traffic intensity, $\rho = \lambda / (c \mu)$ by varying the value of the arrival rate $\lambda$. For our two-class policy, we set $K_n = 0.4 \log(n)$. For the FCFS, non-preemptive two class (Two-class NP), and SJF policies, we utilized the closed-form expression to calculate the steady-state queue length (see online supplement in Argon and Ziya (2009) and Cobham (1954)). For the preemptive two class (Two-class P) and SRPT policies, we use stochastic simulation with a time-horizon $T = 5 \times 10^5$ and 20 independent replications. For these two policies, we provide both the long-run average queue length and the corresponding 95% confidence interval in Table 1.

<table>
<thead>
<tr>
<th>Traffic intensity</th>
<th>FCFS</th>
<th>Two-class NP</th>
<th>SJF</th>
<th>Two-class P</th>
<th>SPRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>9.637</td>
<td>1.777</td>
<td>1.570</td>
<td>1.181 $\pm$ 0.002</td>
<td>1.121 $\pm$ 0.002</td>
</tr>
<tr>
<td>0.85</td>
<td>11.503</td>
<td>2.481</td>
<td>2.102</td>
<td>1.681 $\pm$ 0.006</td>
<td>1.537 $\pm$ 0.004</td>
</tr>
<tr>
<td>0.90</td>
<td>15.019</td>
<td>3.790</td>
<td>3.038</td>
<td>2.796 $\pm$ 0.013</td>
<td>2.363 $\pm$ 0.007</td>
</tr>
<tr>
<td>0.95</td>
<td>25.186</td>
<td>7.178</td>
<td>5.296</td>
<td>6.368 $\pm$ 0.059</td>
<td>4.563 $\pm$ 0.026</td>
</tr>
</tbody>
</table>

We observe from Table 1 that using smart scheduling, we can reduce the steady-state average queue length substantially. Moreover, our two-class priority rule achieves similar, but in general slightly worse, performance than the SJF or SRPT policy. We also observe that preemption in general further improves the performance, but may be hard to administrate in applications.

6. Conclusion

In this paper, we propose a simple two-class priority scheduling policy that achieves similar performance as the SRPT policy. We characterize the process-level diffusion limit under the single-server heavy-traffic regime. The scaling is highly nonstandard and depends on the service time distribution. The heavier the service time distribution, the slower the queue scales as the traffic intensity $\rho$ approaches one.

We also study the effect of imperfect service time information on the performance of our scheduling policy. We show that when the tail of the error distribution is lighter than the tail of the service time distribution, we are still able to achieve $o(1/(1 - \rho))$ scaling with our scheduling policy.

References


