The Impact of Dependence on Unobservable Queues

Jamol Pender
School of Operations Research and Information Engineering
Cornell University
228 Rhodes Hall Ithaca, NY 14853
jjp274@cornell.edu

August 17, 2017

Abstract

In unobservable queueing systems, customers have several chances to leave the system and forgo their service from an agent. Typically customers may choose to leave by balking, which is based on the queue length or they may leave by reneging from the queue, which is based on the virtual waiting time and the patience of the customer. However, the current literature on unobservable queues assumes that the sequences of balking and reneging random variables are independent and this implies that a customer's decision to join the line is independent of the their willingness to wait for service. In this paper, we relax the independence assumption and assess the impact of dependence of the balking and reneging distributions on the stochastic behavior of the queue length and workload processes. We find that the dependence of the balking and reneging distributions significantly affects the distribution of the queue length through decreasing the strength of the state dependent drift towards the origin. Lastly, we show that the joint density near the origin and **not the correlation** of the balking and reneging random variables describes the impact of the queue length and workload processes.

1 Introduction

Performance analysis for service systems such as unobservable queues usually assume that the random variables that determine whether or not a customer will balk or renege from the queue are mutually independent. However, in a variety of real world applications, the balking and and reneging random variables may in fact be jointly dependent. For example, in a grocery store, customers that make the decision to not balk and join a long line of customers will most likely have enough patience to stay in the line for a longer period of time. Thus, there is a positive dependence on a customer's willingness to join a long queue and their willingness to wait a long time to receive service.

The current literature does not take into account the dependence of the balking and reneging distributions and as we will show, this can have a significant impact on the distribution of the queue length and workload processes. We will show that positive dependence among the balking and reneging distributions produces additional congestion in the queue. This additional congestion is created by the fact that customers that join the line are more likely to wait in the line longer, therefore reducing the number of customers that renege from the queue. However, what is suprising is that the correlation plays no obvious role in the increased congestion and instead the increased congestion is influenced by the joint distribution of the balking and reneging random variables near the origin.

We are motivated to study the impact of dependence on unobservable queues since there are several novel applications of unobservable queues. One exciting area that uses unobservable queues in its daily operations is *virtual queueing*. Virtual or mobile queueing allows customers to wait for service via their mobile devices or cell phones. By waiting in line on one's mobile device, the idea of virtual queueing allows customers to spend the time that they are waiting for service more efficiently and gives the customer more flexibilty to use their time more productively. From a psychological perspective, customers no longer agonize about waiting a long time to see a server since they are spending their *waiting time* occupied with other activities of *their*

choice. However, this increased flexibilty offered by virtual queueing is not without a cost. One of the main difficulties experienced in a ticket queue is customer abandonment. More often than not, when customers decide to renege from the queue, they often leave unnoticed to the service manager or operator. As a result of this unobservable feature of the system, what is listed as the actual queue length through an app on their phone is in fact an upper bound for the actual number of customers waiting and is often not the actual number customers to be served ahead of them.

1.1 Relevant and Related Literature

1.1.1 Queues with Correlated Random Variables

The study of queues with correlated parameters has a long history see for example, [1, 4, 16, 7, 19, 8, 9, 12, 13] and references therein. One of the most interesting papers on the subject is [14] where the arrival and service processes are assumed to be positively dependent. In [14], the author assumes that the arrival and service processes in a single server queue follow a bivariate negative exponential distribution, which they show leads to an analysis of the Kibble bivariate gamma distribution. Using the bivariate gamma distribution, they also derive an integral equation for the waiting time and derive important properties of the waiting time in terms of the parameters of the bivariate gamma distribution. In a follow-up paper [15], the author also derives closed form expressions for the moments of the waiting time when the arrival and service processes are dependent.

In other work by [6], the author studies the effect of dependency of the arrival and service processes on the steady state waiting time and queue length. The author shows under the bivariate Marshall-Olkin exponential distribution that the customer waiting time is monotonically decreasing in the dependency in the increasing convex ordering sense. [22] generalizes this result to other bivariate distributions for the arrival and service processes. Moreover, [23] further generalizes the result beyond queueing processes.

Another major body of research analyzes infinite server queues with dependent sequences of random variables. For example the work of [24, 25, 26] explores the impact of dependence on the nonstationary infinite server queue. More specifically, they study when the service times are correlated with each other. This type of dependence is typically seen in recalls or inquiries about consumer products since customers call the call center with the same issues and questions. The authors provide an approximate analysis of the mean and variance of the queue length as a function of the dependence between service times or arrival times. They also show that the correlation significantly impacts the variance but not the mean behavior. Thus, in the infinite server setting, the current literature is only limited to exploring dependence between arrival and service random variables.

One important fact is that none of the previous research explores queues with reneging customers or types of dependence that is not between the arrival and service processes. In this work, we not only explore the impact of dependence on the balking and reneging processes, but we also explore this dependence feature for unobservable queues, which are relatively new in the queueing literature.

1.1.2 Ticket Queues as Models for Unobservable Queues

The study of ticket queues is quite recent. The first paper that explores the nuances of ticket queues is by [31] and it shows that for heavily loaded systems with relatively *impatient* customers the ticket queue experiences a higher percentage of balking. This research analyzes ticket queues with the assumption that the distributions for balking and reneging are exponential. However, work by [17], extends the analysis to more general distributions for the abandonment distribution and balking distributions. Since general distributions are explored in [17], they appeal to heavy traffic theory instead of Markov process theory to derive useful approximations of performance. The main insight provided by [17] is to illustrate the theoretical and practical differences between observable and unobservable or standard and ticket queues in the critically loaded heavy traffic regime. One result is that the standard and ticket queues converge to the same limiting diffusion process, a regulated Ornstein-Uhlenbeck process. In fact a strong notion of similarity is proved by [17] where it shows that the standard and ticket queues are asymptotically coupled in heavy traffic. More recently, there is work by [20] that studies and derives optimal staffing levels for ticket queues with multiple servers. Their analysis is also in the Markovian setting and it does not consider general distributions nor does it derive insight from heavy traffic limit theory.

Like some of the literature in queueing theory such as [30, 28, 17], this paper exploits heavy traffic limit theory in order to derive simple and accurate approximations for the unobservable queue that managers can use to operate their service systems better. Using our heavy traffic theory, we construct estimates of the distribution of the queueing processes by deriving heuristic approximations of our heavy traffic limit theorem based on steady state limits. We show that we are able to approximate various performance metrics such as, the mean, variance, the fraction of balking customers, fraction of reneging customers, and the impact of dependence on these quantities. Moreover, we are also able to calculate the fraction of customers currently in the queue that will eventually renege from the queue. This provides customers and managers with approximations on how to adjust one's perception of how many customers will actually stay long enough to receive service.

1.2 Contributions of Paper

Unlike the previous research on unobservable queues, this paper attempts to assess the impact of the dependence of the balking and abandonment random variables by providing answers to the following questions.

- How does the dependence of the balking and reneging distributions impact the mean, variance, and distribution of the queue length and workload processes?
- How does the dependence of the balking and abandonment random variables impact the fraction of customers that decide to balk or renege from the queue?
- Is there a simple relationship between the correlation of the random variables and the queue length process?
- Are there simple and closed form approximations for modelling the unobservable queue with dependence?

1.3 Notation of Paper

In this section, we provide the reader with a guide of the notation that will be used throughout the rest of the paper. First, we let $\mathbb{R}$ denote the set of all real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers, and $\mathbb{N}_+$ the set of strictly positive integers. For a Polish space $\mathcal{S}$, let $\mathbf{D}(\mathbb{R}, \mathcal{S})$ denote the space of right continuous functions with left limits from $\mathbb{R}_+$ into the Polish space $\mathcal{S}$. The Polish spaces that we consider in this paper are $\mathbb{R}_+$ and $\mathbb{R}_+^2$. Moreover, we define $\varphi(x)$ and $\Phi(x)$ to be the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution where

$$\varphi(x) = \frac{1}{\sqrt{2 \cdot \pi}} e^{-x^2/2} \quad \text{and} \quad \Phi(x) = 1 - \overline{\Phi}(x) = \int_{-\infty}^{x} \varphi(y) dy. \tag{1.1}$$

We also define the standard hazard rate function $h : \mathbb{R} \mapsto \mathbb{R}_+$ to be the ratio of the density of the standard normal distribution to the tail of the standard normal distribution i.e.

$$h(x) = \frac{\varphi(x)}{\overline{\Phi}(x)} \text{ for every } x \in \mathbb{R}. \tag{1.2}$$

1.4 Outline of Paper

The remainder of the paper proceeds as follows. In Section 2, we define the unobservable queue model and show how it can be modelled via a ticket queue. We also provide a decomposition of the queueing model in terms of primitive random variables and processes. In Section 3, we demonstrate how to construct various sequences of dependent random variables as this is important for simulating and generating the queue length and workload processes. In Section 4, we provide the conditions under which our heavy traffic limit theorem is valid. We also provide the main results of the paper in this section. We also develop approximations for performance measures using knowledge of the steady state distribution of the regulated Ornstein-Uhlenbeck process. In Section 5, we also provide extensive numerical results using dependent random variables for the balking and reneging processes. Concluding remarks follow in Section 6 and finally in the Appendix, we provide all of the ingredients needed to prove the main results of the paper.

2 Description of Unobservable Queueing Model with Dependence

In this section, we provide the primitive random variables and the initial conditions for the ticket queue with dependence of the balking and abandonment random variables. For the reader's convenience, we use the terms ticket queue and unobservable queue interchangably throughout the rest of the paper.

2.1 Model Primitives

For each customer, we assume that they have an interarrival time, (potential) service time, initial balking tolerance and (potential) reneging tolerance. For the ith customer these times are captured in the quadruple (u_i, v_i, b_i, r_i). The sequences $\{u_i : i \geq 1\}$, $\{v_i : i \geq 1\}$, $\{b_i : i \geq 1\}$, $\{r_i : i \geq 1\}$ are all defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Since we relax the independence assumption for the balking and reneging random variables, we have that the interarrival times and service times are independent of each other and the balking and reneging random variables, however, we assume that the balking and reneging random variables are dependent on each other and have a joint distribution. The external mean arrival rate of jobs is λ, the mean service rate is μ. The sequences $\{u_i : i \geq 1\}$, $\{v_i : i \geq 1\}$ all have unitary means and the interarrival and service times of the ith customer are u_i/λ and v_i/μ, respectively. The arrival time of the ith job occurs at time

$$t_i = (1/\lambda) \sum_{j=1}^{i} u_j.$$

We assume that the unitary interarrival and service times have finite variances denoted by σ_a^2 and σ_s^2 respectively. The quantities b_i and r_i represent the random variables associated with balking and reneging, respectively. However, unlike the work of Jennings and Pender [17], the balking and reneging random variables are not independent. To describe the balking and reneging random variables, we define G_b and G_r to be their respective marginal cumulative distribution functions and we define $G(x, y)$ to be the joint distribution of the balking and reneging random variables. This implies that $G_b(\cdot) = G(\cdot, \infty)$ and $G_r(\cdot) = G(\infty, \cdot)$. We also assume that the marginal and joint cumulative distribution functions of the balking and reneging random variables vanish at zero. This means that $G_b(0) = G_r(0) = G(0, 0) = 0$. Not only do we assume that the derivatives of the marginal distributions exist at zero and are strictly positive, but we also assume that the sum of the partial derivatives of the joint balking and reneging distribution exist and are strictly positive. When there is no dependence and the balking and reneging distributions are independent, we define the sum of the derivatives of the balking and reneging distributions as

$$\gamma_{ind} = G_b'(0) + G_r'(0). \tag{2.1}$$

This quantity γ_{ind} is the rate at which the queue length process is pushed towards the origin due to balking and reneging when they are independent. However, when the balking and reneging distributions are dependent, the rate is different and becomes

$$\gamma \equiv G_x(0, \infty) + G_y(\infty, 0) - G_x(0, 0) - G_y(0, 0) = G_b'(0) + G_r'(0) - G_x(0, 0) - G_y(0, 0). \tag{2.2}$$

One should note that since $G_x(0, 0)$ and $G_y(0, 0)$ are non-negative, the effect of the dependence is to move the distribution further away from the origin, which increases the queue length and workload. It is also important to note that the sum of $G_x(0, 0)$ and $G_y(0, 0)$ should be non-negative. Suppose that it were not and $G(0, 0) = 0$, then it would be able to move in a direction that would make the joint probability negative and violate the assumptions about probabilitiy distributions.

2.2 The Queue Length and Related Processes

In this section, we define the process $Q = \{Q(t), t \geq 0\}$ to be the queue length process. The queue length process is constructed out of several simpler stochastic processes that describe the complicated dynamics of the queue length process. One such process is the arrival process is $A = \{A(t), t \geq 0\}$, which counts the number of customers that have arrived to the system in the interval (0,t]. Another process is $B = \{B(t), t \geq 0\}$ which is the balking process and serves to track as a function of time the number of arriving

customers who leave immediately upon arrival. The reneging process $R = \{R(t), t \geq 0\}$ tracks the number of people who have abandoned and whose abandonment has also been detected in the system. Unlike an observable queueing process where an abandoned customer is immediately detected, customers that abandon in an unobservable or ticket queue can only be detected when their service would have begun. The process $S = \{S(t), t \geq 0\}$ tracks the number of service completions that the server has done as a function of time. Moreover, we also define the busy time process $T = \{T(t), t \geq 0\}$, which calculates how much time has been spent by the server processing jobs as a function of time t. The idle time process $I = \{I(t), t \geq 0\}$ is complementary to $T(t)$: $I(t) = t - T(t)$ for each $t \geq 0$ since it records the time that the server has remained idle up to time t. Lastly, the workload process $W = \{W(t), t \geq 0\}$ reports as function of time the amount of effort required by the server to process those customers currently in-queue that will not abandon. For simplicity, we also assume that the initial value of the queue length and workload are equal to zero. Since non-zero queue lengths and workload distributions do not provide any additional insight to the main conclusions of the paper, we decide to initialize the processes at zero. Using the above stochastic processes, it is now possible to construct and describe the queue length process with the following equations given below. For each $t \geq 0$, we have that

$$Q(t) = A(t) - B(t) - R(t) - S(T(t)), \tag{2.3}$$

$$A(t) = \sup\left\{ j \geq 0 : \sum_{i=1}^{j} u_i/\lambda \leq t \right\}, \tag{2.4}$$

$$B(t) = \sum_{i=1}^{A(t)} 1(b_i \leq Q(t_i-)/\mu), \tag{2.5}$$

$$R(t) = \sum_{i=1}^{A(t)} 1(b_i > Q(t_i-)/\mu) \cdot 1(r_i \leq W(t_i-)) \cdot 1(W(t_i-) \leq t - t_i), \tag{2.6}$$

$$T(t) = \int_0^t 1(Q(s) > 0) ds, \tag{2.7}$$

$$I(t) = t - T(t), \tag{2.8}$$

$$S(t) = \sup\left\{ k \geq 0 : (1/\mu) \sum_{i=1}^{k} \tilde{v}(i) \leq t \right\}, \tag{2.9}$$

where

$$\tilde{v}(i) = v_i \cdot 1(b_i > Q(t_i-)/\mu) \cdot 1(r_i \geq W(t_i-)), \tag{2.10}$$

and

$$W(t) \quad = \quad I(t) - t + \sum_{i=1}^{A(t)} (v_i/\mu) \cdot 1(b_i > Q(t_i-)/\mu) \cdot 1(r_i > W(t_i-)). \tag{2.11}$$

$Q(t)$ represents the total number of customers currently in the queue and in service at time t. $A(t)$ represents the total number of arrivals up to time t. $B(t)$ is the total number of customers that have balked from the queue up to time t. $R(t)$ is the total number of customers that have reneged and have also been detected by the server up to time t. $T(t)$ represents the time that the server is busy and $I(t)$ is the amount of time that the server is idle. $W(t)$ is the current workload of the server at time t, which is the amount of time it would take for the server to become idle the arrival process were shut off. Lastly, $S(t)$ is the total number of customers that have departed the queue for receiving service by the server up to time t. All of these processes are important and necessary for the construction of the queue length and workload process and will be scaled in order to prove our main limit theorem. However, before we prove the main result, we show how to construct dependent random variables in the sequel.

3 Simulation of Dependent Random Variables

Since we are using dependent random variables for the balking and reneging processes, it is necessary to understand how to construct bivariate sequences of random variables that are dependent. In this section, we give examples of some joint bivariate distributions and how to simulate from them for the balking and reneging random variables.

3.1 Randomly Repeated Random Variables

One way to model dependence between random variables is to use the same random variables with a certain probability. In the context of our queueing model, we suppose that b_i are distributed with cdf $G_b(x)$ and that the $r_i = b_i$ with probability p and the r_i random variables are independent of b_i, although with possibly a different distribution, with probability $(1-p)$. This yields the following joint cdf for (b_i, r_i)

$$\begin{aligned}
G(x,y) &= P(b_i \le x, r_i \le y) && (3.1)\\
&= P(b_i \le x, r_i \le y|\ b_i \perp\!\!\!\perp r_i) \cdot P(b_i \perp\!\!\!\perp r_i) + P(b_i \le x, r_i \le y | b_i = r_i) \cdot P(b_i = r_i) && (3.2)\\
&= P(b_i \le x, b_i \le y|\ r_i \perp\!\!\!\perp r_i) \cdot (1-p) + P(b_i \le x, r_i \le y | b_i = r_i) \cdot p && (3.3)\\
&= P(b_i \le x) \cdot P(r_i \le y) \cdot (1-p) + P(b_i \le x \wedge y) \cdot p && (3.4)\\
&= (1-p) \cdot G_b(x) \cdot F_b(y) + p \cdot G_b(x \wedge y). && (3.5)
\end{aligned}$$

This also implies that the sum of the partial derivatives of the joint distribution around zero is equal to

$$G_x(0,0) + G_y(0,0) = p \cdot G_b'(0). \tag{3.6}$$

This method of repeating the same random variables is not very useful if the balking and reneging distributions are not identical. However, the method is very easy to implement and yields a simple formula for the joint distribution at zero. Under this setting we have that state dependent drift parameter is equal to

$$\gamma = (2-p) \cdot G_b'(0). \tag{3.7}$$

3.2 Randomly Repeated and Scaled Random Variables

We have already seen that one way to model dependence between random variables is to use the same random variables with a certain probability. However, it is the case that the distribution might be a scaled version of the other random variable. In the context of our queueing model, we suppose that b_i are distributed with cdf $G_b(x)$ and that the $r_i = m \cdot b_i$ with probability p and the r_i random variables are independent of b_i, although with possibly a different distribution, with probability $(1-p)$. This yields the following joint cdf for (b_i, r_i)

$$\begin{aligned}
G(x,y) &= P(b_i \le x, r_i \le y \cdot m) && (3.8)\\
&= P(b_i \le x, r_i \le y \cdot m|\ b_i \perp\!\!\!\perp r_i) \cdot P(b_i \perp\!\!\!\perp r_i) + P(b_i \le x, r_i \le y \cdot m | b_i = r_i) \cdot P(b_i = r_i) && (3.9)\\
&= P(b_i \le x, b_i \le y|\ r_i \perp\!\!\!\perp r_i) \cdot (1-p) + P(b_i \le x, r_i \le y \cdot m | b_i = r_i) \cdot p && (3.10)\\
&= P(b_i \le x) \cdot P(r_i \le y \cdot m) \cdot (1-p) + P(b_i \le x \wedge y \cdot m) \cdot p && (3.11)\\
&= (1-p) \cdot G_b(x) \cdot F_b(y \cdot m) + p \cdot G_b(x \wedge y \cdot m). && (3.12)
\end{aligned}$$

This also implies that the sum of the partial derivatives of the joint distribution around zero is equal to

$$G_x(0,0) + G_y(0,0) = p \cdot G_b'(0) \cdot \{x < y \cdot m\} + m \cdot p \cdot G_b'(0) \cdot \{x \ge y \cdot m\}. \tag{3.13}$$

Unlike the unscaled setting, the sum of the partial derivatives of the joint distribution is not continuous at zero. This presents a significant challenge in terms of the theory, nonetheless, it is easy to simulate the random variables to understand the impact of the dependence. This is even more important when one process happens much faster than another.

3.3 Marshall Olkin Bivariate Exponential

Another method that is quite simple to implement is the method developed by Marshall and Olkin in [21]. This method generates correlated bivariate ($n = 2$) exponential random variables according to the specified parameters. The construction of this bivariate distribution is as follows. Suppose that we are analyzing a two-component system where the lifetimes of the components are denoted by X_1 and X_2. Let λ_1 and λ_2 be the rates of the Poisson processes governing the occurence of shocks in the first and second components respectively. Furthermore, let λ_3 represent the rate of shocks that affect both components simultaneously. Then the marginals are exponentially distributed with combined rates to account for both types of shocks, or $X_i \sim \text{Exp}(\lambda_i + \lambda_3)$. To simplify the notation throughout the this section, write $\bar{\lambda} = \lambda_1 + \lambda_2 + \lambda_3$. One way to simulate this bivariate distribution is to use the following procedure, which is also outlined in [11].

- Simulate $l = 3$ independent random variates v_1, v_2, v_3 from Uniform$(0, 1)$.
- Set $X_i = \min_{1\le k\le 3, i\in s_k, \lambda_k\neq 0}(-\ln v_k/\lambda_k), i = 1, 2$, where $s_k \in S$, the set of nonempty subsets of $\{1, 2\}$. Specifically, since $n = 2$ we have

$$X_1 = \min\left(-\ln\left(\frac{v_1}{\lambda_1}\right), -\ln\left(\frac{v_3}{\lambda_3}\right)\right)$$
$$X_2 = \min\left(-\ln\left(\frac{v_2}{\lambda_2}\right), -\ln\left(\frac{v_3}{\lambda_3}\right)\right).$$

As one can see, the random variables X_1 and X_2 are coupled through the joint shocks. It also follows that the joint and marginal densities of X_1 and X_2 are as follows

$$\begin{aligned}
G_b(x) &= 1 - \exp\left(-(\lambda_1 + \lambda_3)\cdot x\right) && (3.14)\\
G_r(y) &= 1 - \exp\left(-(\lambda_2 + \lambda_3)\cdot y\right) && (3.15)\\
\overline{G}(x, y) &= \exp\left(-(\lambda_1 + \lambda_3)\cdot x\right)\cdot \exp\left(-(\lambda_2 + \lambda_3)\cdot y\right)\cdot \exp\left(\lambda_3 \cdot \min(x, y)\right) && (3.16)\\
&= \overline{G}_b(x)\cdot \overline{G}_r(y)\cdot \exp\left(\lambda_3 \cdot \min(x, y)\right) && (3.17)\\
&= 1 - G_b(x) - G_r(y) + G(x, y) && (3.18)\\
\overline{G}_x(x, y) &= (\lambda_1 + \lambda_3)\cdot \lambda_3 \cdot \overline{G}_b(x)\cdot \overline{G}_r(y)\cdot \exp\left(\lambda_3 \cdot x\right)\cdot 1\{x < y\} && (3.19)\\
&+ (\lambda_1 + \lambda_3)\cdot \overline{G}_b(x)\cdot \overline{G}_r(y)\cdot \exp\left(\lambda_3 \cdot y\right)\cdot 1\{y < x\} && (3.20)\\
\overline{G}_y(x, y) &= (\lambda_2 + \lambda_3)\cdot \lambda_3 \cdot \overline{G}_b(x)\cdot \overline{G}_r(y)\cdot \exp\left(\lambda_3 \cdot x\right)\cdot 1\{y < x\} && (3.21)\\
&+ (\lambda_2 + \lambda_3)\cdot \overline{G}_b(x)\cdot \overline{G}_r(y)\cdot \exp\left(\lambda_3 \cdot x\right)\cdot 1\{x < y\} && (3.22)\\
& && (3.23)
\end{aligned}$$

The joint distribution also implies that random variables X_1 and X_2 have the following moments and correlation

$$E[X_1] = \frac{1}{\lambda_1 + \lambda_3} \tag{3.24}$$
$$E[X_2] = \frac{1}{\lambda_2 + \lambda_3} \tag{3.25}$$
$$\text{Var}[X_1] = \frac{1}{(\lambda_1 + \lambda_3)^2} \tag{3.26}$$
$$\text{Var}[X_2] = \frac{1}{(\lambda_2 + \lambda_3)^2} \tag{3.27}$$
$$\rho = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{\lambda_3}{\bar{\lambda}}. \tag{3.28}$$

In terms of computing important performance measures of the queueing process we also need to know the behavior of the joint distribution near the origin. Thus, we compute the following derivatives near the origin.

$$\overline{G}_x(0,0) = \begin{cases} \lambda_1 & \text{if } x < y \\ \lambda_1 + \lambda_3 & \text{if } x > y \\ \lambda_1 + \lambda_2 + \lambda_3 & \text{if } x = y \end{cases}$$

$$\overline{G}_y(0,0) = \begin{cases} \lambda_2 + \lambda_3 & \text{if } x < y \\ \lambda_2 & \text{if } x > y \\ \lambda_1 + \lambda_2 + \lambda_3 & \text{if } x = y \end{cases}$$

$$\overline{G}_x(0,0) + \overline{G}_y(0,0) = \begin{cases} \lambda_1 + \lambda_2 + \lambda_3 & \text{if } x < y \\ \lambda_1 + \lambda_2 + \lambda_3 & \text{if } x > y \\ \lambda_1 + \lambda_2 + \lambda_3 & \text{if } x = y \end{cases}$$

$$\overline{G}_x(0,0) + \overline{G}_y(0,0) = \lambda_1 + \lambda_2 + \lambda_3 = \bar{\lambda}$$

$$G_x(0,0) + G_y(0,0) = \lambda_3 = \rho \cdot \bar{\lambda}$$

This implies that for the Marshall-Olkin joint distribution, the parameter γ is equal to

$$\gamma = G_b'(0) + G_r'(0) - (G_x(0,0) + G_y(0,0)) = \lambda_1 + \lambda_2 + \lambda_3 = \bar{\lambda}.$$

Now that we have a better understanding of the joint distribution near zero, we can use this knowledge later to compute an approximate distribution of the queue length and workload processes.

3.4 Kibble's Bivariate Gamma

Another distribution that generalizes the Marshall-Olkin bivariate exponential distribution is the Kibble bivariate gamma distribution. Kibble's method for generating a bivariate gamma distribution with correlation $0 < \rho < 1$ is based on the well-known relationship between the chi-squared and normal distributions. In general, the sum of the square of n independent standard normal random variables $X_1, ..., X_n \sim N(0,1)$ has a chi-square distribution with n degrees of freedom i.e.

$$Z := \sum_{i=1}^{n} X_i^2 \ \sim \ \chi^2(n) \tag{3.29}$$

Furthermore the chi-squared distribution is a special case of the gamma distribution, such that $X \sim \Gamma\left(\frac{n}{2}, 2\right)$.

The procedure for generating instances of bivariate gamma random variables is outlined in [2, 24] and can be easily implemented in any numerical software package. We outline the procedure below.

- Generate n i.i.d. random samples $(X_1, Y_1), ..., (X_n, Y_n)$ from a bivariate normal with $\mu = 0$, correlation ρ_0, and
$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_0\sigma_1\sigma_2 \\ \rho_0\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$
- Let $X := \frac{1}{2\sigma_1^2}\sum_{i=1}^{n} X_i^2$ and $Y := \frac{1}{2\sigma_1^2}\sum_{i=1}^{n} Y_i^2$.
- Repeat the previous two steps N times. The resulting X and Y vectors are Kibble's bivariate gamma with $\rho = \rho_0^2$ with shape $\alpha = \frac{n}{2}$.

The joint probability density function for (X, Y) is

$$h(x,y) = \frac{1}{(1-\rho)\Gamma(\alpha)} \left(\frac{xy}{\rho}\right)^{\frac{\alpha-1}{2}} \exp\left(-\frac{x+y}{1-\rho}\right) I_{\alpha-1}\left(\frac{2\sqrt{xy\rho}}{1-\rho}\right) \tag{3.30}$$

where $x, y \geq 0$, $0 \leq \rho < 1$, $\Gamma(\cdot)$ is the gamma function, and $I_\nu(\cdot)$ is the modified Bessel function of the first kind and order ν. Given the joint distribution, we can also calculate the moments to find the mean, variance, and correlation of the random variables i.e.

$$E[X] = E[Y] = \text{Var}[X] = \text{Var}[Y] = \alpha$$

and

$$\text{Corr}[X, Y] = \rho_0^2 = \rho$$

3.5 The NORTA Method

The last method of generating bivariate sequences of random variables is the Normal to Anything (NORTA) method. This method was introduced by [5] and allows us to transform vectors of independent identically distributed i.i.d random variables into vectors with *arbitrary* marginal distributions and specific correlations. The idea is to first, generate a k-dimensional, standard multivariate normal vector $\mathbf{Z} = (Z_1, Z_2, ..., Z_k)'$ with correlation matrix Σ_Z. Let Φ be the univariate standard normal cumulative distribution function (cdf) and $F_{X_i}^{-1} \equiv \inf\{x : F_{X_i}(x) \geq u\}$ denote the inverse cdf of the desired marginal distribution of X_i. Then, the NORTA vector $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \begin{pmatrix} F_{X_1}^{-1}[\Phi(Z_1)] \\ F_{X_2}^{-1}[\Phi(Z_2)] \\ \vdots \\ F_{X_k}^{-1}[\Phi(Z_k)] \end{pmatrix}$$

$\mathbf{X}$ is a random vector with Corr[$\mathbf{X}$]=Σ_Z where $X_i \sim F_{X_i}, i = 1, 2, ..., k$ where each F_{X_i} is an arbitrary cdf. Although the NORTA method works for arbitrary dimensions we will only consider two dimensions in this work. The major advantage of the method is that this method is very simple to implement on a computer and can be used to construct sequences of *arbitrary* random variables. Moreover, some distributions like the uniform and log-normal distributions have explicit correlations that are well known, see for example [5]. Although we have explicit values for the subsequent correlation for some random variables using the NORTA method, the main disadvantage of this method is that it is almost impossible to calcuate the density near origin. We will see that the correlation is not as important in determining the stochastic behavior of the queueing process as the joint density of the balking and reneging distributions. The joint density near the origin is useful for our performance measure calculations, however, is unavailable in closed form using the NORTA method. Before we present the simulation results for dependent balking and reneging distributions, we will prove the main results of the paper, a heavy traffic limit theorem for the distribution of the queue length and workload processes.

4 Heavy Traffic Scaling and Limit Theorems

In this section, we prove the main result of the paper, a heavy traffic limit theorem. This heavy traffic limit theorem serves to approximate the queueing and workload processes of an unobservable queue with dependence between balking and reneging distributions. To derive this heavy traffic limit theorem, we need to appropriately scale the queueing and workload processes. To this end, we consider a sequence of queueing systems, indexed by scaling parameter n. The arrival and service rates of the nth system are given by λ^n and μ^n. Moreover, Equations (2.3)–(2.11) have straightforward analogues with the λ and μ replaced by λ^n and μ^n, respectively. It is important to notice that the balking and reneging rates are not scaled with the parameter n. Thus, for each $t \geq 0$, we have that

$$Q^n(t) = A^n(t) - B^n(t) - R^n(t) - S^n(T^n(t)), \tag{4.1}$$

$$A^n(t) = \sup\left\{ j \geq 0 : \sum_{i=1}^{j} u_i/\lambda^n \leq t \right\}, \tag{4.2}$$

$$B^n(t) = \sum_{i=1}^{A^n(t)} 1(b_i \le Q^n(t_i^n-)/\mu^n), \tag{4.3}$$

$$R^n(t) = \sum_{i=1}^{A^n(t)} 1(b_i > Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \le W^n(t_i^n-)) \cdot 1(W^n(t_i^n-) \le t - t_i), \tag{4.4}$$

$$T^n(t) = \int_0^t 1(Q^n(s) > 0)ds, \tag{4.5}$$

$$W^n(t) \quad = \quad I^n(t) - t + \sum_{i=1}^{A^n(t)} (v_i/\mu^n) \cdot 1(b_i > Q^n(t_i^n-)/\mu^n) \cdot 1(r_i > W^n(t_i^n-)). \tag{4.6}$$

and

$$I^n(t) = t - T^n(t), \tag{4.7}$$

where

$$t_i^n = (1/\lambda^n) \sum_{j=1}^{i} u_j.$$

The associated scaled processes are $Q^n = \{Q^n(t), t \ge 0\}$, $A^n = \{A^n(t), t \ge 0\}$, $B^n = \{B^n(t), t \ge 0\}$, $R^n = \{R^n(t), t \ge 0\}$, $S^n = \{S^n(t), t \ge 0\}$, $T^n = \{T^n(t), t \ge 0\}$, $I^n = \{I^n(t), t \ge 0\}$, and $W^n = \{W^n(t), t \ge 0\}$. Now we define the diffusion scaled queue length process $\tilde{Q}^n = \{\tilde{Q}^n(t), t \ge 0\}$ and the diffusion scaled workload process $\tilde{W}^n = \{\tilde{W}^n(t), t \ge 0\}$, where for each $t \ge 0$,

$$\tilde{Q}^n(t) = \frac{Q^n(t)}{\sqrt{n}} \qquad \text{and} \qquad \tilde{W}^n(t) = \sqrt{n} W^n(t).$$

In addition to not scaling the balking and reneging rates, one should note that we do not scale time since the arrival rates and service rates are already proportional to n. The main reason for not scaling the balking and abandonment times is that the demand for service may change and subsequently the service speed must adjust accordingly; however, customers will maintain the same willingness to wait and will not change their assessment of the service quality as demand increases.

4.1 Decomposition of Queue Length Process

Now that we have a complete representation of the queue length process in terms of its primitive parts, we can even further decompose the diffusion-scaled queue length processes for each $t \ge 0$ into more convenient probabilistic quantities i.e.

$$\begin{aligned} \tilde{Q}^n(t) \quad = \quad & \tilde{Q}^n(0) + \tilde{A}^n(t) - \tilde{M}_b^n(\bar{A}^n(t)) - \tilde{M}_r^n(\bar{A}^n(t)) - \tilde{\varepsilon}_q^n(t) - \tilde{\delta}_q^n(t) - \tilde{S}^n(T^n(t)) \\ & -\gamma \int_0^t \tilde{Q}^n(s)ds + \frac{(\lambda^n - \mu^n)}{\sqrt{n}} t + \tilde{Y}^n(t), \end{aligned} \tag{4.8}$$

where, $\tilde{Q}^n(0) = 0$ is the scaled initial queue length and, for each $t \ge 0$,

$$\tilde{A}^n(t) = \frac{1}{\sqrt{n}} \left(A^n(t) - \lambda^n t \right), \tag{4.9}$$

$$\bar{A}^n(t) = \frac{1}{n} A^n(t), \tag{4.10}$$

$$\tilde{M}_b^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left(1(b_i \le Q^n(t_i^n-)/\mu^n) - G_b(\sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n) \right), \tag{4.11}$$

$$\tilde{M}_r^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left(1(r_i \le Q^n(t_i^n-)/\mu^n) - G_r(\sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n) \right) \tag{4.12}$$
$$- \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left(1(b_i \le Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \le Q^n(t_i^n-)/\mu^n) - G(\sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n, \sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n) \right),$$

$$\varepsilon^n(t) = \left(R^n(t) - \sum_{i=1}^{A^n(t)} (1(r_i \le Q^n(T_i^n-)/\mu^n) - 1(b_i \le Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \le Q^n(t_i^n-)/\mu^n)) \right). \tag{4.13}$$

$$S^n(t) = \sup \left\{ k \ge 0 : (1/\mu^n) \sum_{i=1}^{k} \hat{v}^n(i) \le t \right\}, \tag{4.14}$$

where $\hat{v}^n(i)$ represents the ith person to be served by the server i.e.

$$\hat{v}^n(i) = \frac{v_i}{\mu^n} \cdot 1(b_i > Q^n(t_i^n-)/\mu^n) \cdot 1(r_i > W^n(t_i^n-). \tag{4.15}$$

Moreover, we have that

$$\tilde{\delta}_q^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \left(G_b(\sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n) + G_r(\sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n) \right) \tag{4.16}$$
$$- \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} G\left(\sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n, \sqrt{n}\tilde{Q}^n(t_i^n-)/\mu^n \right) - \gamma \int_0^t \tilde{Q}^n(s)ds,$$

$$\tilde{S}^n(t) = \frac{1}{\sqrt{n}} \left(S^n(t) - \mu^n t \right), \tag{4.17}$$

$$\tilde{\varepsilon}_q^n = \frac{1}{\sqrt{n}} \varepsilon^n(t), \tag{4.18}$$

and

$$\tilde{Y}^n(t) = \left(\frac{\mu^n}{n} \right) \tilde{I}^n(t) = \left(\frac{\mu^n}{\sqrt{n}} \right) I^n(t). \tag{4.19}$$

We refer to $\tilde{A}^n = \{\tilde{A}^n(t), t \ge 0\}$ as the diffusion-scaled arrival process and to $\bar{A}^n = \{\bar{A}^n(t), t \ge 0\}$ as its fluid-scaled analog. The reader may notice that the process $\tilde{\delta}_q^n = \{\tilde{\delta}_q^n(t), t \ge 0\}$ has what looks like an instantaneous drift that is proportionate to the value of the scaled queue length process. The remaining processes are centered and diffusion scaled versions of their original analogs: $\tilde{M}_b^n = \{\tilde{M}_b^n(t), t \ge 0\}$, $\tilde{M}_r^n = \{\tilde{M}_r^n(t), t \ge 0\}$, $\tilde{S}^n = \{\tilde{S}^n(t), t \ge 0\}$, $\tilde{\varepsilon}_q^n = \{\tilde{\varepsilon}_q^n(t), t \ge 0\}$, and $\tilde{Y}^n = \{Y^n(t), t \ge 0\}$. In fact, $\tilde{M}_b^n$ and $\tilde{M}_r^n$ are martingales under an appropriate filtration.

4.2 Heavy Traffic Setup and Limit Theorem

In order to prove a heavy traffic limit theorem, suppose that for our sequence of systems indexed by n, that arrival and service rates are order n quantities and are asymptotically identical; that is, as $n \to \infty$,

$$\lambda^n/n \to \mu \quad \text{and} \quad \mu^n/n \to \mu. \tag{4.20}$$

We also assume that the difference between the arrival and service rates should be an order $\sqrt{n}$ quantity such that as $n \to \infty$,

$$(\lambda^n - \mu^n)/\sqrt{n} = \beta^n \to \beta \in (-\infty, \infty). \tag{4.21}$$

One can refer to (4.20)-(4.21) as the heavy traffic conditions. The heavy traffic conditions also imply that

$$\lambda^n/\mu^n \to 1, \tag{4.22}$$

as $n \to \infty$. Moreover, we assume that the random variables associated with balking and abandonment are unaffected by the change in the index n. We also define

$$\sigma \equiv \mu\sqrt{\sigma_a^2 + \sigma_s^2} \tag{4.23}$$

the standard deviation associated with the arrival and service times. Lastly, define $B = \{B(t), t \geq 0\}$ as a Brownian motion with no drift and an infinitesimal variance of 1. Fortunately, we are aided by the framework developed in [28], which provides a theoretical justification for the queue length and workload process representations i.e (4.6) and (4.8).

$$(\tilde{Q}^n, \tilde{Y}^n) = (\Phi_\gamma, \Psi_\gamma)(\tilde{X}_q^n), \tag{4.24}$$

where $(\Phi_\gamma, \Psi_\gamma) : \mathbf{D}(\mathbb{R}_+, \mathbb{R}) \mapsto \mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^+)$ is a Lipschitz continuous map, and $\tilde{X}_q^n = \{\tilde{X}_q^n(t), t \geq 0\}$ for each $t \geq 0$

$$\tilde{X}_q^n(t) = \tilde{A}^n(t) - \tilde{M}_b^n(\bar{A}^n(t)) - \tilde{M}_r^n(\bar{A}^n(t)) - \tilde{\varepsilon}_q^n(t) - \tilde{\delta}_q^n(t) - \tilde{S}^n(T^n(t)) + \frac{(\lambda^n - \mu^n)}{\sqrt{n}}t. \tag{4.25}$$

Since the convergence of the departure process is quite difficult, we will resort to using the workload process for convergence. We will show that the individual parts of $\tilde{X}_q^n$ will converge to either an asymptotically negiglible quantity, a deterministic drift, or a Brownian motion. In fact, $\tilde{X}_q^n$ converges to

$$\tilde{X}_q = \beta e + \sigma B. \tag{4.26}$$

We now present our main result for the diffusion scaled workload and queue length processes.

Theorem 4.1. *If*

$$\tilde{Q}^n(0) \Rightarrow 0, \qquad \text{as } n \to \infty, \tag{4.27}$$

then

$$(\tilde{Q}^n, \tilde{Y}^n) \Rightarrow (\tilde{Q}, \tilde{Y}), \qquad \text{as } n \to \infty, \tag{4.28}$$

where $\tilde{Q} = \Phi_\gamma(\tilde{X}_q)$, $\tilde{Y} = \Psi_\gamma(\tilde{X}_q)$, and together $\tilde{Q}$ and $\tilde{Y}$ have the same law as the following stochastic differential equation

$$d\tilde{Q}(t) = \beta dt - \gamma \cdot \tilde{Q}(t)dt + \sigma dB(t) + d\tilde{Y}(t). \tag{4.29}$$

Proof. Theorem 4.1 Given our decomposition for the queue length term given in (4.8), we must show convergence for each part of the decomposition in order to prove our limit to the regulated Ornstein-Uhlenbeck process. We begin with the elements of the process $\tilde{X}_q^n$ (see (4.25)), which we will show converge to either zero, a drift, or a Brownian motion. By Proposition A.8 and the random time change theorem of [3], we have that

$$\tilde{M}_b^n \circ \bar{A}^n \to 0 \quad \text{and} \quad \tilde{M}_r^n \circ \bar{A}^n \to 0$$

in probabilitiy as $n \to \infty$. In order to show that reneging is equivalent to balking in the ticket queue, where customers are detected later upon potential service entry, we need to show that

$$\tilde{\varepsilon}_q^n \to 0,$$

in probability as $n \to \infty$. This proved in Proposition A.12. This proposition is also very critical to our analysis because in the ticket queue customers are not recorded as reneging customers until their service would have begun.

We also show in Proposition A.14 that the balking and abandonment processes converge to a restortative drift term i.e.

$$\tilde{\delta}_q^n \to 0,$$

in probability as $n \to \infty$. Moreover, Proposition A.4 gives the convergence of the centered and scaled arrival and departure processes to scaled Brownian motions. Thus, we can show that

$$\tilde{S}^n + \tilde{A}^n \to \sigma_s B_s + \sigma_a B_a \stackrel{D}{=} \sigma B.$$

Finally, the heavy traffic condition i.e. (4.21) provides the deterministic drift of the queueing process i.e.

$$\frac{(\lambda^n - \mu^n)}{\sqrt{n}} \to \beta$$

as $n \to \infty$. Hence,

$$\tilde{X}_q^n \Rightarrow \tilde{X}_q.$$

From (4.30) we have that

$$(\tilde{Q}^n(0), \tilde{X}_q^n) \Rightarrow (0, \tilde{X}_q), \qquad \text{as } n \to \infty,$$

and hence, by the Continuous Mapping Theorem [3, 28],

$$(\tilde{Q}^n, \tilde{Y}^n) = (\Phi_\gamma, \Psi_\gamma)(0, \tilde{X}_q^n) \Rightarrow (\Phi_\gamma, \Psi_\gamma)(0, \tilde{X}_q) = (\tilde{Q}, \tilde{Y}), \qquad \text{as } n \to \infty.$$

This concludes the proof. □

Theorem 4.2. *If*

$$\tilde{W}^n(0) \Rightarrow 0, \qquad \textit{as } n \to \infty, \tag{4.30}$$

then

$$(\tilde{W}^n, \tilde{Y}^n) \Rightarrow (\tilde{Q}/\mu, \tilde{Y}), \qquad \textit{as } n \to \infty. \tag{4.31}$$

Proof. Theorem 4.2 Recall that from Theorem 4.1, we have that queue length process converges to a diffusion process given in Equation 4.29. Now by using the state space collapse result of Proposition A.9, we have that asymptotically the scaled workload and scaled queue lengths are scalar multiples of one another and therefore converge together to scalar multiples of the same stochastic diffusion process. Hence (4.31) holds. This concludes the proof. □

Remark. One thing to note is that the limit theorems do not depend explicitly on the correlation of the balking and reneging random variables. Thus, the correlation is not an important parameter in governing the dynamics of the queue length process. However, the distribution of the queue length and workload processes depend on the joint distribution of the balking and reneging random variables near the origin. The joint distribution near the origin serves to decrease the state dependent drift parameter of the regulated OU process and changes the dynamics of the process significantly.

Corollary 4.3. *Under the conditions of 4.1 and assuming that the balking and reneging distributions are independent, then the parameter γ of 4.29 can be replaced by γ_{ind} where*

$$\gamma_{ind} = G_b'(0) + G_r'(0). \tag{4.32}$$

Proof. This follows from the fact that the joint density can be written as $G(x, y) = G_b(x) \cdot G_r(y)$ and each partial derivative is equal to zero when evaluated at the origin since we assume that the balking and reneging distributions have no mass at the origin. □

The processes $\tilde{W}(t)$ and $\tilde{Q}(t)$ are refered to as reflected Ornstein-Uhlenbeck (ROU) processes. The steady state distribution of $\tilde{W}(t)$ or $\tilde{Q}(t)$ is a truncated (at zero) normal variable,

$$\lim_{t\to\infty} \tilde{Q}(t) = \tilde{Q}(\infty) \stackrel{D}{=} \text{Truncated Normal}\left(\frac{\beta}{\gamma}, \frac{\sigma^2}{2\gamma}, 0, \infty\right), \tag{4.33}$$

and

$$\lim_{t\to\infty} \tilde{W}(t) = \tilde{W}(\infty) \stackrel{D}{=} \text{Truncated Normal}\left(\frac{\beta}{\mu\gamma}, \frac{\sigma^2}{2\mu^2\gamma}, 0, \infty\right), \tag{4.34}$$

There is an extensive amount of literature on the ROU process. See for example [29] for a study of the ROU process when the balking and reneging parameters are small. When the balking and reneging parameters are small, the ROU process can be viewed as a perturbation of reflected Brownian motion, which is highly studied in the queueing and applied probability literature. Moreover, if one is interested in computing additional steady state moment approximations of the queue length process such as the skewness and kurtosis, see for example [27]. [27] contains explicit formulas for the cumulant moments of the truncated normal distribution in terms of Hermite polynomials and the hazard rate function of the normal distribution.

4.3 Heavy Traffic Based Approximations for Performance Measures

In this section, we use the main result Theorem 4.1 to approximate various performance measures and quantities of interest for the ticket queue with dependence. We will show how the dependent structure of the balking and reneging random variables impacts various performance meausre of the queue length and workload processes.

4.3.1 Moments of Queue Length and Workload Processes

The first performance measures we approximate are the mean and variance of the queue length and workload processes. We use our knowledge of the steady state distribution of the reflected OU process as a proxy for the distribution of the queue length and workload process. With this in mind, we derive the following approximations for the mean and variance of the queue length and workload processes.

$$E[\tilde{W}(\infty)] = \frac{\beta}{\mu\gamma} + \frac{\sigma}{\mu\sqrt{2\gamma}} \cdot h\left(-\frac{\beta}{\sigma\sqrt{\gamma/2}}\right), \tag{4.35}$$

and

$$E[\tilde{Q}(\infty)] = \frac{\beta}{\gamma} + \frac{\sigma}{\sqrt{2\gamma}} \cdot h\left(-\frac{\beta}{\sigma\sqrt{\gamma/2}}\right). \tag{4.36}$$

It is also possible to compute the variance of the workload and queue length processes using the explicit formulas of [27] as

$$\mathrm{Var}[\tilde{W}(\infty)] = \frac{\sigma^2}{2\mu^2\gamma} \cdot \left(1 + \left(\frac{\beta}{\sigma\sqrt{\gamma/2}}\right) \cdot h\left(-\frac{\beta}{\sigma\sqrt{\gamma/2}}\right) - h\left(-\frac{\beta}{\sigma\sqrt{\gamma/2}}\right)^2\right), \tag{4.37}$$

and

$$\mathrm{Var}[\tilde{Q}(\infty)] = \frac{\sigma^2}{2\gamma} \cdot \left(1 + \left(\frac{\beta}{\sigma\sqrt{\gamma/2}}\right) \cdot h\left(-\frac{\beta}{\sigma\sqrt{\gamma/2}}\right) - h\left(-\frac{\beta}{\sigma\sqrt{\gamma/2}}\right)^2\right). \tag{4.38}$$

4.3.2 Fraction of Balking Customers

Now we will show how to use the steady state approximations for the queue length to calculate the fraction of customers that balk from the queue.

$$\begin{aligned}
\frac{B(t)}{A(t)} &= \frac{\sum_{i=1}^{A(t)} 1(b_i < Q(t_i)/\mu)}{A(t)} && (4.39)\\
&\approx \frac{\lambda t \cdot G_b(Q(\infty)/\mu)}{\lambda \cdot t} && (4.40)\\
&\approx \frac{\lambda t \cdot G_b'(0) \cdot E[Q(\infty)]/\mu}{\lambda t} && (4.41)\\
&\approx \frac{G_b'(0) \cdot E[Q(\infty)]}{\lambda} && (4.42)\\
& && (4.43)
\end{aligned}$$

One subtle point that is worth mentioning is that the dependence between the balking and reneging random variables does not affect the fraction of balking customers directly. It can only be seen through the change in the mean queue length. This lack of a direct impact is because the dependence is somewhat directional and the impact of the dependence is not equally felt by the balking and reneging distributions. This is a result of the ordering of balking and reneging processes in which balking happens first and a customer can only renege if they did not balk initially.

4.3.3 Fraction of Reneging Customers

Now we will show how to use the steady state approximations for the queue length to calculate the fraction of customers that renege from the queue. Since the balking customers actually has an affect on the customers who renege, we expect to see the dependence between balking and reneging take effect.

$$\begin{aligned}
\frac{R(t)}{A(t)} &\approx \frac{\sum_{i=1}^{A(t)} 1(b_i > Q(t_i)/\mu) \cdot 1(r_i < W(t_i))}{A(t)} && (4.44)\\
&\approx \frac{\sum_{i=1}^{A(t)} 1(r_i < Q(t_i)/\mu) - 1(b_i < Q(t_i)/\mu) \cdot 1(r_i < Q(t_i)/\mu)}{A(t)} && (4.45)\\
&\approx \frac{\lambda t \cdot (G_r(Q(\infty)/\mu) - G(Q(\infty)/\mu, Q(\infty)/\mu))}{\lambda \cdot t} && (4.46)\\
&\approx \frac{\lambda t \cdot (G_r'(0) - G_x(0,0) - G_y(0,0)) \cdot E[Q(\infty)]/\mu}{\lambda t} && (4.47)\\
&\approx \frac{(G_r'(0) - G_x(0,0) - G_y(0,0)) \cdot E[Q(\infty)]}{\lambda} && (4.48)
\end{aligned}$$

It is for the reneging customers where we see the dependence structure of the balking and reneging random variables. In fact the dependence structure actually reduces the number of customers that renege from the queue. This is not the case for the balk because of the ordering of balking and reneging. Moreover, one should notice that if the balking and reneging random variables are identical, then the reneging fraction should be near zero. Although the theory predicts a value of zero for reneging if it is identical to the balking random variables, in simulations this reneging fraction will be small, but non-zero. This is because in the limit, we have a state space collapse result, which shows that the normalized queue length and workload processes are identical. However, in reality where n is finite, the normalized workload and queue length values are not the identical and the number of customers that renege from the queue reflects the difference between the two processes. Furthermore, if the balking and reneging random variables were independent, then we have that

$$\frac{R(t)}{A(t)} \approx \frac{G_r'(0) \cdot E[Q(\infty)]}{\lambda}. \tag{4.49}$$

We can even construct a further refinement of the fraction of reneging customers. This refinement takes into account the fact that the reneging process does not include the customers who balked from the queue. Thus, we have the following refinement for the fraction of reneging customers

$$\frac{R(t)}{A(t)} \approx \frac{(G_r'(0) - G_x(0,0) - G_y(0,0)) \cdot E[Q(\infty)]}{\lambda} \cdot \left(1 - \frac{G_b'(0) \cdot E[Q(\infty)]}{\lambda}\right). \tag{4.50}$$

This differs from the original approximation by multiplying the fraction of people who do not balk, which is smaller than one.

4.3.4 Reneging Customers in Queue

Since we are working with ticket queues, it is important to have a good understanding of the customers waiting for service. Not all customers will receive service and it is necessary to know how many customers that are currently in the queue that going to renege from the queue. Given this number, one knows the

actual queue length and not the upper bound that can be calculated from the difference of the customer in service and the number of arrivals. From our heavy traffic limit theorem it is possible to derive an approximation for this number of reneging customers in queue. If we define the number of customers that have not been cleared and will eventually renege from the queue to be the stochastic process $Z(t)$, then we have the following approximation for the steady state mean of $Z(t)$ i.e. $Z(\infty)$

$$\begin{aligned}
\mathbb{E}[Z(t)] &\approx \frac{1}{2}E\left[G_r\left(W(t)\right)\right]\cdot E\left[Q(t)\right] && (4.51)\\
&\approx \frac{1}{2}E\left[G_r\left(\frac{Q(t)}{\mu}\right)\right]\cdot E\left[Q(t)\right] && (4.52)\\
&\approx \frac{1}{2}E\left[G_r'(0)\cdot\frac{Q(t)}{\mu}\right]\cdot E\left[Q(t)\right] && (4.53)\\
&= \frac{G_r'(0)}{2\mu}\cdot E\left[Q(t)\right]^2 && (4.54)\\
&\approx \frac{G_r'(0)}{2}\cdot E\left[\tilde{Q}(\infty)\right]^2. && (4.55)
\end{aligned}$$

This approximation is based on the fact that there are $Q(t)$ customers in the queue at time t. Moreover, we have that the probability that a customer reneges from the queue is $G_r\left(W(t)\right)$ and the workload is proportional to the queue length scaled by the mean service time. Using a simple Taylor expansion, we find that the overestimate of the queue length depends on the second moment of the queue length distribution. The scale factor of one half appears because the customer that we are measuring for can be chosen uniformly from the $Q(t)$ customers that are present in the queue at time t. This approximation gives a customer an approximation of how many customers will abandon from the queue and they might be able to adjust their patience accordingly.

5 Numerical results

In this section we provide numerical comparisions of our limit theorems with stochastic simulation. We compare our heavy traffic limit theorems with the following bivariate sequences: bivariate Marshall-Olkin, randomly repeated random variables using (uniform, exponential, hyperexponential) distributions, Kibble's bivariate gamma, and NORTA using (uniform, exponential, lognormal) distributions.

5.1 Bivariate Marshall-Olkin

In this section, we simulate the balking and reneging random variables with the bivariate Marshall-Olkin distribution and compare against our limit theorems see that our limit theorems are quite good at approximating the behavior of the stochastic system. We also see that the performance is better when there is less dependence.

Table 5.1: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Exp(θ_b), Exp(θ_r), Marshall-Olkin, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu}\cdot\beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	100	−0.5	95	8.00	4	4	0.00	$\sqrt{2}$	2.61	2.64 ± 0.021	4.08	5.87 ± 0.047	0.0261	0.0232 ± 0.0002	0.1097	0.0959 ± 0.0006	0.1097	0.0977	0.0708 ± 0.0004
0.25	100	−0.5	95	6.40	4	4	1.60	$\sqrt{2}$	2.89	2.80 ± 0.012	5.04	6.37 ± 0.048	0.0289	0.0255 ± 0.0001	0.1215	0.1015 ± 0.0007	0.0729	0.0641	0.0525 ± 0.0002
0.5	100	−0.5	95	5.33	4	4	2.67	$\sqrt{2}$	3.14	2.93 ± 0.017	5.98	6.84 ± 0.060	0.0314	0.0273 ± 0.0002	0.1320	0.1061 ± 0.0008	0.0440	0.0382	0.0381 ± 0.0002
0.75	100	−0.5	95	4.57	4	4	3.43	$\sqrt{2}$	3.36	3.06 ± 0.012	6.90	7.24 ± 0.086	0.0336	0.0291 ± 0.0002	0.1415	0.1101 ± 0.0004	0.0202	0.0174	0.0264 ± 0.0005
1	100	−0.5	95	4.00	4	4	4.00	$\sqrt{2}$	3.57	3.17 ± 0.021	7.81	7.68 ± 0.097	0.0357	0.0306 ± 0.0002	0.1502	0.1140 ± 0.0010	0.0000	0.0000	0.0165 ± 0.0002
0	100	0	100	8.00	4	4	0.00	$\sqrt{2}$	2.82	2.93 ± 0.020	4.54	6.65 ± 0.038	0.0282	0.0256 ± 0.0002	0.1128	0.1060 ± 0.0006	0.1128	0.1001	0.0770 ± 0.0007
0.25	100	0	100	6.40	4	4	1.60	$\sqrt{2}$	3.15	3.13 ± 0.010	5.68	7.28 ± 0.040	0.0315	0.0283 ± 0.0001	0.1262	0.1129 ± 0.0012	0.0757	0.0661	0.0569 ± 0.0004
0.5	100	0	100	5.33	4	4	2.67	$\sqrt{2}$	3.45	3.31 ± 0.016	6.81	7.92 ± 0.087	0.0345	0.0306 ± 0.0002	0.1382	0.1185 ± 0.0006	0.0461	0.0397	0.0413 ± 0.0002
0.75	100	0	100	4.57	4	4	3.43	$\sqrt{2}$	3.73	3.44 ± 0.021	7.95	8.30 ± 0.129	0.0373	0.0325 ± 0.0003	0.1493	0.1227 ± 0.0004	0.0213	0.0181	0.0283 ± 0.0003
1	100	0	100	4.00	4	4	4.00	$\sqrt{2}$	3.99	3.61 ± 0.026	9.08	8.94 ± 0.112	0.0399	0.0348 ± 0.0002	0.1596	0.1281 ± 0.0005	0.0000	0.0000	0.0175 ± 0.0002
0	100	0.5	105	8.00	4	4	0.00	$\sqrt{2}$	3.06	3.25 ± 0.016	5.05	7.47 ± 0.045	0.0306	0.0281 ± 0.0002	0.1166	0.1168 ± 0.0004	0.1166	0.1030	0.0826 ± 0.0006
0.25	100	0.5	105	6.40	4	4	1.60	$\sqrt{2}$	3.46	3.48 ± 0.021	6.39	8.19 ± 0.059	0.0346	0.0312 ± 0.0002	0.1316	0.1247 ± 0.0007	0.0790	0.0686	0.0615 ± 0.0006
0.5	100	0.5	105	5.33	4	4	2.67	$\sqrt{2}$	3.82	3.70 ± 0.016	7.75	8.91 ± 0.099	0.0382	0.0341 ± 0.0002	0.1455	0.1320 ± 0.0008	0.0485	0.0414	0.0444 ± 0.0003
0.75	100	0.5	105	4.57	4	4	3.43	$\sqrt{2}$	4.16	3.90 ± 0.051	9.13	9.57 ± 0.188	0.0416	0.0368 ± 0.0005	0.1584	0.1383 ± 0.0014	0.0226	0.0190	0.0304 ± 0.0004
1	100	0.5	105	4.00	4	4	4.00	$\sqrt{2}$	4.48	4.09 ± 0.024	10.54	10.27 ± 0.116	0.0448	0.0393 ± 0.0003	0.1706	0.1442 ± 0.0012	0.0000	0.0000	0.0183 ± 0.0001

Table 5.2: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Exp(θ_b), Exp(θ_r), Marshall-Olkin, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	400	−0.5	390	8.00	4	4	0.00	$\sqrt{2}$	5.21	5.25 ± 0.032	16.33	20.06 ± 0.225	0.0130	0.0122 ± 0.0001	0.0534	0.0501 ± 0.0002	0.0534	0.0506	0.0428 ± 0.0003
0.25	400	−0.5	390	6.40	4	4	1.60	$\sqrt{2}$	5.77	5.66 ± 0.049	20.16	22.89 ± 0.385	0.0144	0.0135 ± 0.0001	0.0592	0.0541 ± 0.0003	0.0355	0.0334	0.0309 ± 0.0002
0.5	400	−0.5	390	5.33	4	4	2.67	$\sqrt{2}$	6.27	6.00 ± 0.008	23.91	25.19 ± 0.104	0.0157	0.0145 ± 0.0000	0.0643	0.0571 ± 0.0003	0.0214	0.0201	0.0215 ± 0.0000
0.75	400	−0.5	390	4.57	4	4	3.43	$\sqrt{2}$	6.72	6.33 ± 0.029	27.61	27.56 ± 0.252	0.0168	0.0154 ± 0.0001	0.0689	0.0601 ± 0.0003	0.0098	0.0092	0.0139 ± 0.0002
1	400	−0.5	390	4.00	4	4	4.00	$\sqrt{2}$	7.14	6.58 ± 0.035	31.25	29.54 ± 0.650	0.0178	0.0162 ± 0.0001	0.0732	0.0623 ± 0.0006	0.0000	0.0000	0.0075 ± 0.0001
0	400	0	400	8.00	4	4	0.00	$\sqrt{2}$	5.64	5.76 ± 0.035	18.17	22.48 ± 0.241	0.0141	0.0133 ± 0.0001	0.0564	0.0548 ± 0.0003	0.0564	0.0532	0.0464 ± 0.0004
0.25	400	0	400	6.40	4	4	1.60	$\sqrt{2}$	6.31	6.26 ± 0.057	22.71	25.92 ± 0.433	0.0158	0.0148 ± 0.0001	0.0631	0.0596 ± 0.0004	0.0378	0.0355	0.0336 ± 0.0003
0.5	400	0	400	5.33	4	4	2.67	$\sqrt{2}$	6.91	6.67 ± 0.012	27.25	28.77 ± 0.155	0.0173	0.0160 ± 0.0000	0.0691	0.0633 ± 0.0003	0.0230	0.0214	0.0233 ± 0.0001
0.75	400	0	400	4.57	4	4	3.43	$\sqrt{2}$	7.46	7.08 ± 0.035	31.80	31.73 ± 0.291	0.0187	0.0173 ± 0.0001	0.0746	0.0670 ± 0.0003	0.0107	0.0099	0.0151 ± 0.0002
1	400	0	400	4.00	4	4	4.00	$\sqrt{2}$	7.98	7.42 ± 0.055	36.34	34.55 ± 0.610	0.0199	0.0183 ± 0.0001	0.0798	0.0699 ± 0.0005	0.0000	0.0000	0.0079 ± 0.0000
0	400	0.5	410	8.00	4	4	0.00	$\sqrt{2}$	6.12	6.31 ± 0.040	20.19	25.12 ± 0.296	0.0153	0.0145 ± 0.0001	0.0597	0.0599 ± 0.0004	0.0597	0.0562	0.0501 ± 0.0005
0.25	400	0.5	410	6.40	4	4	1.60	$\sqrt{2}$	6.91	6.91 ± 0.059	25.54	29.20 ± 0.464	0.0173	0.0163 ± 0.0001	0.0674	0.0655 ± 0.0004	0.0405	0.0377	0.0364 ± 0.0002
0.5	400	0.5	410	5.33	4	4	2.67	$\sqrt{2}$	7.64	7.42 ± 0.018	30.99	32.62 ± 0.121	0.0191	0.0178 ± 0.0001	0.0745	0.0700 ± 0.0002	0.0248	0.0230	0.0253 ± 0.0001
0.75	400	0.5	410	4.57	4	4	3.43	$\sqrt{2}$	8.32	7.93 ± 0.042	36.53	36.33 ± 0.371	0.0208	0.0193 ± 0.0001	0.0811	0.0747 ± 0.0004	0.0116	0.0107	0.0163 ± 0.0002
1	400	0.5	410	4.00	4	4	4.00	$\sqrt{2}$	8.96	8.33 ± 0.034	42.14	39.45 ± 0.803	0.0224	0.0205 ± 0.0001	0.0874	0.0782 ± 0.0005	0.0000	0.0000	0.0083 ± 0.0000

5.2 Randomly Repeated Random Variables

In this section, we simulate the balking and reneging random variables with randomly repeated random variables with (uniform, exponential, hyperexponential) distributions. Like in the previous case, we compare against our limit theorems see that our limit theorems are quite good at approximating the behavior of the stochastic system. We also see that the performance is better when there is less dependence.

5.2.1 Uniform Marginals

Table 5.3: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Unif(0, 1), Randomly Repeated, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	100	−0.5	95	2.00	1	1	0.00	$\sqrt{2}$	4.82	4.69 ± 0.066	14.67	16.99 ± 0.542	0.0482	0.0438 ± 0.0007	0.0508	0.0467 ± 0.0008	0.0508	0.0482	0.0403 ± 0.0011
0.25	100	−0.5	95	1.75	1	1	0.25	$\sqrt{2}$	5.10	4.88 ± 0.072	16.52	18.08 ± 0.573	0.0510	0.0461 ± 0.0008	0.0537	0.0488 ± 0.0005	0.0403	0.0381	0.0333 ± 0.0005
0.5	100	−0.5	95	1.50	1	1	0.50	$\sqrt{2}$	5.44	5.13 ± 0.077	18.92	19.74 ± 0.789	0.0544	0.0491 ± 0.0008	0.0573	0.0513 ± 0.0005	0.0286	0.0270	0.0257 ± 0.0003
0.75	100	−0.5	95	1.25	1	1	0.75	$\sqrt{2}$	5.86	5.36 ± 0.032	22.18	21.39 ± 0.298	0.0586	0.0520 ± 0.0003	0.0617	0.0536 ± 0.0005	0.0154	0.0145	0.0173 ± 0.0004
1	100	−0.5	95	1.00	1	1	1.00	$\sqrt{2}$	6.41	5.72 ± 0.069	26.85	23.90 ± 0.558	0.0641	0.0564 ± 0.0007	0.0675	0.0572 ± 0.0008	0.0000	0.0000	0.0076 ± 0.0002
0	100	0	100	2.00	1	1	0.00	$\sqrt{2}$	5.64	5.62 ± 0.068	18.17	21.11 ± 0.504	0.0564	0.0519 ± 0.0007	0.0564	0.0561 ± 0.0008	0.0564	0.0532	0.0470 ± 0.0004
0.25	100	0	100	1.75	1	1	0.25	$\sqrt{2}$	6.03	5.88 ± 0.057	20.76	22.75 ± 0.319	0.0603	0.0550 ± 0.0006	0.0603	0.0588 ± 0.0006	0.0452	0.0425	0.0390 ± 0.0004
0.5	100	0	100	1.50	1	1	0.50	$\sqrt{2}$	6.51	6.22 ± 0.072	24.23	25.08 ± 0.558	0.0651	0.0591 ± 0.0007	0.0651	0.0620 ± 0.0007	0.0326	0.0305	0.0302 ± 0.0004
0.75	100	0	100	1.25	1	1	0.75	$\sqrt{2}$	7.14	6.67 ± 0.088	29.07	27.69 ± 0.598	0.0714	0.0645 ± 0.0009	0.0714	0.0668 ± 0.0007	0.0178	0.0166	0.0202 ± 0.0003
1	100	0	100	1.00	1	1	1.00	$\sqrt{2}$	7.98	7.04 ± 0.060	36.34	30.73 ± 0.351	0.0798	0.0692 ± 0.0007	0.0798	0.0704 ± 0.0007	0.0000	0.0000	0.0086 ± 0.0003
0	100	0.5	105	2.00	1	1	0.00	$\sqrt{2}$	6.65	6.68 ± 0.049	22.37	25.96 ± 0.363	0.0665	0.0611 ± 0.0005	0.0634	0.0667 ± 0.0005	0.0634	0.0593	0.0551 ± 0.0004
0.25	100	0.5	105	1.75	1	1	0.25	$\sqrt{2}$	7.20	7.08 ± 0.056	25.93	28.18 ± 0.271	0.0720	0.0656 ± 0.0006	0.0685	0.0705 ± 0.0006	0.0514	0.0479	0.0455 ± 0.0008
0.5	100	0.5	105	1.50	1	1	0.50	$\sqrt{2}$	7.88	7.60 ± 0.069	30.78	31.53 ± 0.644	0.0788	0.0717 ± 0.0007	0.0751	0.0758 ± 0.0004	0.0375	0.0347	0.0357 ± 0.0006
0.75	100	0.5	105	1.25	1	1	0.75	$\sqrt{2}$	8.80	8.16 ± 0.076	37.76	34.99 ± 0.330	0.0880	0.0784 ± 0.0008	0.0838	0.0818 ± 0.0006	0.0210	0.0192	0.0237 ± 0.0004
1	100	0.5	105	1.00	1	1	1.00	$\sqrt{2}$	10.09	8.88 ± 0.105	48.62	39.84 ± 0.887	0.1009	0.0872 ± 0.0011	0.0961	0.0888 ± 0.0011	0.0000	0.0000	0.0096 ± 0.0002

Table 5.4: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Unif(0, 1), Randomly Repeated, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	400	−0.5	390	2.00	1	1	0.00	$\sqrt{2}$	9.65	9.49 ± 0.082	58.68	62.91 ± 1.171	0.0241	0.0229 ± 0.0002	0.0247	0.0237 ± 0.0003	0.0247	0.0241	0.0220 ± 0.0003
0.25	400	−0.5	390	1.75	1	1	0.25	$\sqrt{2}$	10.21	9.97 ± 0.092	66.08	69.46 ± 1.444	0.0255	0.0242 ± 0.0002	0.0262	0.0249 ± 0.0003	0.0196	0.0191	0.0180 ± 0.0002
0.5	400	−0.5	390	1.50	1	1	0.50	$\sqrt{2}$	10.88	10.52 ± 0.046	75.69	76.20 ± 1.229	0.0272	0.0257 ± 0.0001	0.0279	0.0263 ± 0.0002	0.0140	0.0136	0.0135 ± 0.0001
0.75	400	−0.5	390	1.25	1	1	0.75	$\sqrt{2}$	11.73	11.15 ± 0.174	88.71	85.52 ± 2.168	0.0293	0.0275 ± 0.0005	0.0301	0.0279 ± 0.0004	0.0075	0.0073	0.0086 ± 0.0002
1	400	−0.5	390	1.00	1	1	1.00	$\sqrt{2}$	12.82	11.89 ± 0.121	107.39	97.19 ± 1.673	0.0321	0.0296 ± 0.0003	0.0329	0.0296 ± 0.0004	0.0000	0.0000	0.0029 ± 0.0000
0	400	0	400	2.00	1	1	0.00	$\sqrt{2}$	11.28	11.22 ± 0.094	72.68	78.61 ± 1.170	0.0282	0.0269 ± 0.0003	0.0282	0.0281 ± 0.0004	0.0282	0.0274	0.0257 ± 0.0002
0.25	400	0	400	1.75	1	1	0.25	$\sqrt{2}$	12.06	11.83 ± 0.126	83.06	86.28 ± 2.294	0.0302	0.0286 ± 0.0003	0.0302	0.0297 ± 0.0003	0.0226	0.0219	0.0210 ± 0.0003
0.5	400	0	400	1.50	1	1	0.50	$\sqrt{2}$	13.03	12.63 ± 0.208	96.90	97.07 ± 2.569	0.0326	0.0308 ± 0.0005	0.0326	0.0315 ± 0.0006	0.0163	0.0158	0.0160 ± 0.0002
0.75	400	0	400	1.25	1	1	0.75	$\sqrt{2}$	14.27	13.64 ± 0.187	116.28	111.47 ± 1.972	0.0357	0.0335 ± 0.0005	0.0357	0.0341 ± 0.0004	0.0089	0.0086	0.0102 ± 0.0002
1	400	0	400	1.00	1	1	1.00	$\sqrt{2}$	15.96	14.87 ± 0.229	145.35	131.18 ± 4.240	0.0399	0.0369 ± 0.0006	0.0399	0.0371 ± 0.0006	0.0000	0.0000	0.0033 ± 0.0001
0	400	0.5	410	2.00	1	1	0.00	$\sqrt{2}$	13.31	13.35 ± 0.084	89.50	97.45 ± 1.599	0.0333	0.0318 ± 0.0002	0.0325	0.0334 ± 0.0003	0.0325	0.0314	0.0302 ± 0.0001
0.25	400	0.5	410	1.75	1	1	0.25	$\sqrt{2}$	14.39	14.13 ± 0.084	103.72	107.09 ± 1.704	0.0360	0.0340 ± 0.0002	0.0351	0.0353 ± 0.0002	0.0263	0.0254	0.0248 ± 0.0002
0.5	400	0.5	410	1.50	1	1	0.50	$\sqrt{2}$	15.77	15.39 ± 0.158	123.12	124.27 ± 1.483	0.0394	0.0374 ± 0.0004	0.0385	0.0385 ± 0.0004	0.0192	0.0185	0.0190 ± 0.0002
0.75	400	0.5	410	1.25	1	1	0.75	$\sqrt{2}$	17.60	16.67 ± 0.286	151.04	142.90 ± 4.695	0.0440	0.0409 ± 0.0007	0.0429	0.0417 ± 0.0007	0.0107	0.0103	0.0120 ± 0.0002
1	400	0.5	410	1.00	1	1	1.00	$\sqrt{2}$	20.18	18.63 ± 0.287	194.47	170.08 ± 4.991	0.0505	0.0463 ± 0.0008	0.0492	0.0466 ± 0.0008	0.0000	0.0000	0.0037 ± 0.0001

5.2.2 Exponential Marginals

Table 5.5: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Exp(θ_b), Randomly Repeated, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	100	−0.5	95	8.00	4	4	0.00	$\sqrt{2}$	2.61	2.64 ± 0.018	4.08	5.88 ± 0.093	0.0261	0.0233 ± 0.0002	0.1097	0.0958 ± 0.0005	0.1097	0.0977	0.0709 ± 0.0006
0.25	100	−0.5	95	7.00	4	4	1.00	$\sqrt{2}$	2.77	2.74 ± 0.027	4.63	6.23 ± 0.116	0.0277	0.0247 ± 0.0003	0.1166	0.0997 ± 0.0008	0.0875	0.0773	0.0592 ± 0.0006
0.5	100	−0.5	95	6.00	4	4	2.00	$\sqrt{2}$	2.97	2.87 ± 0.011	5.35	6.66 ± 0.053	0.0297	0.0264 ± 0.0001	0.1252	0.1035 ± 0.0004	0.0626	0.0547	0.0466 ± 0.0004
0.75	100	−0.5	95	5.00	4	4	3.00	$\sqrt{2}$	3.23	3.01 ± 0.024	6.35	7.15 ± 0.101	0.0323	0.0283 ± 0.0003	0.1359	0.1088 ± 0.0010	0.0340	0.0294	0.0322 ± 0.0001
1	100	−0.5	95	4.00	4	4	4.00	$\sqrt{2}$	3.57	3.17 ± 0.031	7.81	7.65 ± 0.150	0.0357	0.0306 ± 0.0004	0.1502	0.1141 ± 0.0010	0.0000	0.0000	0.0165 ± 0.0002
0	100	0	100	8.00	4	4	0.00	$\sqrt{2}$	2.82	2.92 ± 0.012	4.54	6.58 ± 0.065	0.0282	0.0254 ± 0.0002	0.1128	0.1055 ± 0.0006	0.1128	0.1001	0.0766 ± 0.0005
0.25	100	0	100	7.00	4	4	1.00	$\sqrt{2}$	3.02	3.05 ± 0.026	5.19	7.04 ± 0.132	0.0302	0.0272 ± 0.0003	0.1206	0.1103 ± 0.0007	0.0905	0.0796	0.0638 ± 0.0006
0.5	100	0	100	6.00	4	4	2.00	$\sqrt{2}$	3.26	3.22 ± 0.018	6.06	7.60 ± 0.122	0.0326	0.0294 ± 0.0002	0.1303	0.1157 ± 0.0003	0.0651	0.0567	0.0503 ± 0.0003
0.75	100	0	100	5.00	4	4	3.00	$\sqrt{2}$	3.57	3.39 ± 0.021	7.27	8.16 ± 0.111	0.0357	0.0318 ± 0.0002	0.1427	0.1212 ± 0.0007	0.0357	0.0306	0.0348 ± 0.0003
1	100	0	100	4.00	4	4	4.00	$\sqrt{2}$	3.99	3.61 ± 0.008	9.08	8.91 ± 0.110	0.0399	0.0348 ± 0.0001	0.1596	0.1285 ± 0.0006	0.0000	0.0000	0.0175 ± 0.0002
0	100	0.5	105	8.00	4	4	0.00	$\sqrt{2}$	3.06	3.25 ± 0.010	5.05	7.47 ± 0.079	0.0306	0.0281 ± 0.0001	0.1166	0.1165 ± 0.0006	0.1166	0.1030	0.0826 ± 0.0006
0.25	100	0.5	105	7.00	4	4	1.00	$\sqrt{2}$	3.29	3.39 ± 0.015	5.81	7.99 ± 0.082	0.0329	0.0300 ± 0.0002	0.1254	0.1214 ± 0.0004	0.0940	0.0822	0.0688 ± 0.0008
0.5	100	0.5	105	6.00	4	4	2.00	$\sqrt{2}$	3.58	3.59 ± 0.020	6.84	8.57 ± 0.132	0.0358	0.0325 ± 0.0002	0.1364	0.1278 ± 0.0005	0.0682	0.0589	0.0542 ± 0.0002
0.75	100	0.5	105	5.00	4	4	3.00	$\sqrt{2}$	3.96	3.81 ± 0.020	8.30	9.33 ± 0.074	0.0396	0.0356 ± 0.0003	0.1507	0.1353 ± 0.0013	0.0377	0.0320	0.0375 ± 0.0003
1	100	0.5	105	4.00	4	4	4.00	$\sqrt{2}$	4.48	4.08 ± 0.021	10.54	10.20 ± 0.165	0.0448	0.0393 ± 0.0002	0.1706	0.1440 ± 0.0009	0.0000	0.0000	0.0182 ± 0.0003

Table 5.6: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Exp(θ_b), Randomly Repeated, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	400	−0.5	390	8.00	4	4	0.00	$\sqrt{2}$	5.21	5.24 ± 0.006	16.33	19.98 ± 0.101	0.0130	0.0122 ± 0.0000	0.0534	0.0501 ± 0.0002	0.0534	0.0506	0.0428 ± 0.0002
0.25	400	−0.5	390	7.00	4	4	1.00	$\sqrt{2}$	5.54	5.50 ± 0.011	18.53	21.75 ± 0.125	0.0139	0.0130 ± 0.0000	0.0568	0.0526 ± 0.0002	0.0426	0.0402	0.0354 ± 0.0002
0.5	400	−0.5	390	6.00	4	4	2.00	$\sqrt{2}$	5.94	5.82 ± 0.038	21.42	23.97 ± 0.272	0.0149	0.0139 ± 0.0001	0.0610	0.0554 ± 0.0005	0.0305	0.0286	0.0273 ± 0.0003
0.75	400	−0.5	390	5.00	4	4	3.00	$\sqrt{2}$	6.46	6.17 ± 0.034	25.40	26.55 ± 0.232	0.0161	0.0150 ± 0.0001	0.0662	0.0585 ± 0.0004	0.0166	0.0155	0.0179 ± 0.0002
1	400	−0.5	390	4.00	4	4	4.00	$\sqrt{2}$	7.14	6.59 ± 0.018	31.25	29.81 ± 0.209	0.0178	0.0163 ± 0.0000	0.0732	0.0623 ± 0.0002	0.0000	0.0000	0.0075 ± 0.0001
0	400	0	400	8.00	4	4	0.00	$\sqrt{2}$	5.64	5.75 ± 0.019	18.17	22.40 ± 0.027	0.0141	0.0133 ± 0.0000	0.0564	0.0549 ± 0.0003	0.0564	0.0532	0.0463 ± 0.0002
0.25	400	0	400	7.00	4	4	1.00	$\sqrt{2}$	6.03	6.07 ± 0.024	20.76	24.62 ± 0.218	0.0151	0.0143 ± 0.0000	0.0603	0.0577 ± 0.0003	0.0452	0.0425	0.0385 ± 0.0003
0.5	400	0	400	6.00	4	4	2.00	$\sqrt{2}$	6.51	6.42 ± 0.045	24.23	26.99 ± 0.243	0.0163	0.0153 ± 0.0001	0.0651	0.0610 ± 0.0004	0.0326	0.0305	0.0295 ± 0.0003
0.75	400	0	400	5.00	4	4	3.00	$\sqrt{2}$	7.14	6.87 ± 0.015	29.07	30.24 ± 0.274	0.0178	0.0166 ± 0.0001	0.0714	0.0650 ± 0.0003	0.0178	0.0166	0.0195 ± 0.0002
1	400	0	400	4.00	4	4	4.00	$\sqrt{2}$	7.98	7.43 ± 0.057	36.34	34.36 ± 0.428	0.0199	0.0183 ± 0.0001	0.0798	0.0700 ± 0.0004	0.0000	0.0000	0.0079 ± 0.0001
0	400	0.5	410	8.00	4	4	0.00	$\sqrt{2}$	6.12	6.31 ± 0.029	20.19	25.19 ± 0.333	0.0153	0.0145 ± 0.0001	0.0597	0.0601 ± 0.0002	0.0597	0.0562	0.0502 ± 0.0003
0.25	400	0.5	410	7.00	4	4	1.00	$\sqrt{2}$	6.58	6.64 ± 0.028	23.24	27.40 ± 0.269	0.0165	0.0155 ± 0.0001	0.0642	0.0630 ± 0.0004	0.0482	0.0451	0.0414 ± 0.0003
0.5	400	0.5	410	6.00	4	4	2.00	$\sqrt{2}$	7.16	7.10 ± 0.039	27.35	30.32 ± 0.230	0.0179	0.0168 ± 0.0001	0.0698	0.0672 ± 0.0003	0.0349	0.0325	0.0319 ± 0.0002
0.75	400	0.5	410	5.00	4	4	3.00	$\sqrt{2}$	7.91	7.69 ± 0.055	33.20	34.56 ± 0.477	0.0198	0.0186 ± 0.0001	0.0772	0.0723 ± 0.0006	0.0193	0.0178	0.0212 ± 0.0002
1	400	0.5	410	4.00	4	4	4.00	$\sqrt{2}$	8.96	8.33 ± 0.028	42.14	39.53 ± 0.131	0.0224	0.0205 ± 0.0001	0.0874	0.0781 ± 0.0003	0.0000	0.0000	0.0083 ± 0.0001

5.2.3 Hyperexponential Marginals

Table 5.7: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Hyp($\alpha_b, \lambda_{b_1}, \lambda_{b_2}$), Randomly Repeated, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	100	−0.5	95	9.00	4.5	4.5	0.00	$\sqrt{2}$	2.47	2.50 ± 0.015	3.65	5.34 ± 0.057	0.0247	0.0219 ± 0.0002	0.1169	0.1015 ± 0.0007	0.1169	0.1032	0.0737 ± 0.0007
0.25	100	−0.5	95	7.88	4.5	4.5	1.12	$\sqrt{2}$	2.62	2.60 ± 0.015	4.14	5.65 ± 0.069	0.0262	0.0233 ± 0.0001	0.1243	0.1056 ± 0.0005	0.0932	0.0816	0.0619 ± 0.0005
0.5	100	−0.5	95	6.75	4.5	4.5	2.25	$\sqrt{2}$	2.82	2.72 ± 0.015	4.79	6.03 ± 0.060	0.0282	0.0248 ± 0.0001	0.1334	0.1099 ± 0.0006	0.0667	0.0578	0.0481 ± 0.0005
0.75	100	−0.5	95	5.62	4.5	4.5	3.38	$\sqrt{2}$	3.06	2.85 ± 0.006	5.69	6.47 ± 0.050	0.0306	0.0268 ± 0.0001	0.1450	0.1149 ± 0.0006	0.0362	0.0310	0.0339 ± 0.0003
1	100	−0.5	95	4.50	4.5	4.5	4.50	$\sqrt{2}$	3.38	3.02 ± 0.013	7.00	6.97 ± 0.068	0.0338	0.0290 ± 0.0002	0.1603	0.1208 ± 0.0006	0.0000	0.0000	0.0175 ± 0.0001
0	100	0	100	9.00	4.5	4.5	0.00	$\sqrt{2}$	2.66	2.78 ± 0.016	4.04	6.05 ± 0.066	0.0266	0.0242 ± 0.0002	0.1197	0.1122 ± 0.0006	0.1197	0.1054	0.0797 ± 0.0004
0.25	100	0	100	7.88	4.5	4.5	1.12	$\sqrt{2}$	2.84	2.90 ± 0.017	4.61	6.43 ± 0.048	0.0284	0.0258 ± 0.0002	0.1279	0.1162 ± 0.0004	0.0960	0.0837	0.0666 ± 0.0006
0.5	100	0	100	6.75	4.5	4.5	2.25	$\sqrt{2}$	3.07	3.04 ± 0.024	5.38	6.84 ± 0.091	0.0307	0.0276 ± 0.0003	0.1382	0.1218 ± 0.0007	0.0691	0.0595	0.0521 ± 0.0004
0.75	100	0	100	5.62	4.5	4.5	3.38	$\sqrt{2}$	3.36	3.20 ± 0.019	6.46	7.36 ± 0.097	0.0336	0.0298 ± 0.0002	0.1514	0.1277 ± 0.0007	0.0378	0.0321	0.0361 ± 0.0003
1	100	0	100	4.50	4.5	4.5	4.50	$\sqrt{2}$	3.76	3.40 ± 0.015	8.08	8.02 ± 0.055	0.0376	0.0327 ± 0.0001	0.1693	0.1348 ± 0.0007	0.0000	0.0000	0.0186 ± 0.0002
0	100	0.5	105	9.00	4.5	4.5	0.00	$\sqrt{2}$	2.87	3.04 ± 0.025	4.46	6.69 ± 0.092	0.0287	0.0262 ± 0.0002	0.1231	0.1220 ± 0.0005	0.1231	0.1079	0.0851 ± 0.0006
0.25	100	0.5	105	7.88	4.5	4.5	1.12	$\sqrt{2}$	3.09	3.19 ± 0.032	5.13	7.14 ± 0.130	0.0309	0.0281 ± 0.0003	0.1323	0.1277 ± 0.0013	0.0992	0.0861	0.0713 ± 0.0003
0.5	100	0.5	105	6.75	4.5	4.5	2.25	$\sqrt{2}$	3.36	3.37 ± 0.009	6.04	7.69 ± 0.088	0.0336	0.0304 ± 0.0001	0.1438	0.1335 ± 0.0004	0.0719	0.0616	0.0559 ± 0.0005
0.75	100	0.5	105	5.62	4.5	4.5	3.38	$\sqrt{2}$	3.71	3.57 ± 0.024	7.32	8.29 ± 0.092	0.0371	0.0332 ± 0.0003	0.1589	0.1416 ± 0.0009	0.0397	0.0334	0.0387 ± 0.0003
1	100	0.5	105	4.50	4.5	4.5	4.50	$\sqrt{2}$	4.19	3.83 ± 0.010	9.29	9.07 ± 0.076	0.0419	0.0367 ± 0.0001	0.1798	0.1507 ± 0.0008	0.0000	0.0000	0.0192 ± 0.0003

Table 5.8: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Hyp($\alpha_b, \lambda_{b_1}, \lambda_{b_2}$), Randomly Repeated, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	γ	$G_b'(0)$	$G_r'(0)$	$G_b(0,0)+G_r(0,0)$	σ	Q_{ROU}	Q_{sim}	V_{ROU}	V_{sim}	W_{ROU}	W_{sim}	B_{ROU}	B_{sim}	R_{ROU}	R_{ROU2}	R_{sim}
0	400	−0.5	390	9.00	4.5	4.5	0.00	$\sqrt{2}$	4.94	4.98 ± 0.028	14.61	18.16 ± 0.187	0.0123	0.0116 ± 0.0001	0.0569	0.0534 ± 0.0003	0.0569	0.0537	0.0451 ± 0.0003
0.25	400	−0.5	390	7.88	4.5	4.5	1.12	$\sqrt{2}$	5.25	5.19 ± 0.032	16.58	19.51 ± 0.118	0.0131	0.0122 ± 0.0001	0.0606	0.0556 ± 0.0002	0.0454	0.0427	0.0371 ± 0.0003
0.5	400	−0.5	390	6.75	4.5	4.5	2.25	$\sqrt{2}$	5.63	5.48 ± 0.023	19.17	21.40 ± 0.140	0.0141	0.0131 ± 0.0001	0.0650	0.0587 ± 0.0001	0.0325	0.0304	0.0285 ± 0.0002
0.75	400	−0.5	390	5.62	4.5	4.5	3.38	$\sqrt{2}$	6.12	5.86 ± 0.032	22.75	23.91 ± 0.270	0.0153	0.0142 ± 0.0001	0.0706	0.0623 ± 0.0005	0.0177	0.0164	0.0191 ± 0.0001
1	400	−0.5	390	4.50	4.5	4.5	4.50	$\sqrt{2}$	6.77	6.25 ± 0.048	28.02	26.76 ± 0.400	0.0169	0.0154 ± 0.0001	0.0781	0.0662 ± 0.0007	0.0000	0.0000	0.0081 ± 0.0001
0	400	0	400	9.00	4.5	4.5	0.00	$\sqrt{2}$	5.32	5.41 ± 0.020	16.15	20.05 ± 0.160	0.0133	0.0125 ± 0.0001	0.0598	0.0579 ± 0.0003	0.0598	0.0563	0.0485 ± 0.0002
0.25	400	0	400	7.88	4.5	4.5	1.12	$\sqrt{2}$	5.69	5.71 ± 0.011	18.46	22.05 ± 0.053	0.0142	0.0133 ± 0.0000	0.0640	0.0609 ± 0.0002	0.0480	0.0449	0.0403 ± 0.0001
0.5	400	0	400	6.75	4.5	4.5	2.25	$\sqrt{2}$	6.14	6.05 ± 0.027	21.53	24.22 ± 0.341	0.0154	0.0143 ± 0.0001	0.0691	0.0643 ± 0.0003	0.0345	0.0322	0.0310 ± 0.0002
0.75	400	0	400	5.62	4.5	4.5	3.38	$\sqrt{2}$	6.73	6.45 ± 0.047	25.84	26.87 ± 0.259	0.0168	0.0156 ± 0.0001	0.0757	0.0685 ± 0.0005	0.0189	0.0175	0.0205 ± 0.0002
1	400	0	400	4.50	4.5	4.5	4.50	$\sqrt{2}$	7.52	6.95 ± 0.048	32.30	30.42 ± 0.423	0.0188	0.0171 ± 0.0001	0.0846	0.0734 ± 0.0006	0.0000	0.0000	0.0084 ± 0.0001
0	400	0.5	410	9.00	4.5	4.5	0.00	$\sqrt{2}$	5.74	5.93 ± 0.021	17.84	22.39 ± 0.139	0.0144	0.0136 ± 0.0001	0.0630	0.0631 ± 0.0003	0.0630	0.0591	0.0524 ± 0.0002
0.25	400	0.5	410	7.88	4.5	4.5	1.12	$\sqrt{2}$	6.17	6.27 ± 0.029	20.52	24.55 ± 0.154	0.0154	0.0146 ± 0.0001	0.0678	0.0665 ± 0.0004	0.0508	0.0474	0.0433 ± 0.0001
0.5	400	0.5	410	6.75	4.5	4.5	2.25	$\sqrt{2}$	6.71	6.67 ± 0.027	24.15	27.25 ± 0.160	0.0168	0.0158 ± 0.0001	0.0737	0.0708 ± 0.0004	0.0368	0.0341	0.0335 ± 0.0003
0.75	400	0.5	410	5.62	4.5	4.5	3.38	$\sqrt{2}$	7.42	7.17 ± 0.026	29.29	30.53 ± 0.150	0.0185	0.0173 ± 0.0001	0.0814	0.0756 ± 0.0003	0.0204	0.0187	0.0220 ± 0.0002
1	400	0.5	410	4.50	4.5	4.5	4.50	$\sqrt{2}$	8.39	7.81 ± 0.013	37.15	35.11 ± 0.137	0.0210	0.0192 ± 0.0000	0.0921	0.0822 ± 0.0004	0.0000	0.0000	0.0089 ± 0.0001

One important thing to notice with the randomly repeated random variables is when the same variable is used with probability 1, then we would expect to have no customers renege from the queue. This is because if the customer did not balk, then they should not renege from the queue since it is the same random variables and the scaled queue length and workload processes are identical in the limit. However, we see some fraction of customers reneging from the queue. This difference between the limit theorem and reality is due to the fact that the workload and the scaled queue length are not identical when the arrival and service rates are finite. In fact the scaled queue length is actually larger than the workload since it contains customers that will eventually renege from the queue. But what is the significance of this difference? It is important to realize that this small difference can have a large impact on the mean and variance of the queue length and explains why some of the mean and variance values are not close to some of the simulated values.

5.3 Kibble's Bivariate Gamma

In this section, we simulate the balking and reneging random variables with Kibble's bivariate gamma distribution. Unlike in the previous cases, we cannot compare against our limit theorems see that our limit theorems since the sum of the partial derivatives of the joint distirbution is equal to zero. Thus, we are unable to compare against our heavy traffic limit theorems and new theory needs to be developed for this case.

Table 5.9: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = $\Gamma(\alpha_b, 1)$, $\Gamma(\alpha_r, 1)$, Kibble's Gamma, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), Service = Exp($\mu$)

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	100	−0.5	95	10.77 ± 0.238	85.09 ± 5.394	0.1057 ± 0.0025	0.0087 ± 0.0003	0.0082 ± 0.0004
0.25	100	−0.5	95	10.98 ± 0.305	87.98 ± 4.611	0.1079 ± 0.0031	0.0089 ± 0.0005	0.0084 ± 0.0003
0.5	100	−0.5	95	11.01 ± 0.293	88.02 ± 3.544	0.1082 ± 0.0030	0.0090 ± 0.0004	0.0081 ± 0.0003
0.75	100	−0.5	95	11.00 ± 0.091	89.18 ± 2.203	0.1082 ± 0.0010	0.0089 ± 0.0001	0.0075 ± 0.0002
1	100	−0.5	95	12.43 ± 0.477	116.23 ± 9.987	0.1238 ± 0.0050	0.0113 ± 0.0008	0.0018 ± 0.0001
0	100	0	100	16.67 ± 0.326	140.95 ± 5.785	0.1621 ± 0.0034	0.0171 ± 0.0006	0.0159 ± 0.0005
0.25	100	0	100	16.71 ± 0.330	141.48 ± 4.996	0.1624 ± 0.0032	0.0172 ± 0.0006	0.0158 ± 0.0005
0.5	100	0	100	16.85 ± 0.319	146.13 ± 4.934	0.1641 ± 0.0035	0.0173 ± 0.0006	0.0153 ± 0.0006
0.75	100	0	100	17.52 ± 0.115	155.36 ± 2.596	0.1712 ± 0.0012	0.0186 ± 0.0002	0.0140 ± 0.0002
1	100	0	100	20.20 ± 0.552	201.75 ± 4.221	0.2009 ± 0.0055	0.0242 ± 0.0006	0.0031 ± 0.0001
0	100	0.5	105	24.49 ± 0.601	190.82 ± 6.798	0.2348 ± 0.0058	0.0310 ± 0.0013	0.0276 ± 0.0012
0.25	100	0.5	105	24.71 ± 0.650	189.91 ± 5.164	0.2368 ± 0.0066	0.0313 ± 0.0012	0.0270 ± 0.0010
0.5	100	0.5	105	25.07 ± 0.480	197.99 ± 6.203	0.2408 ± 0.0047	0.0324 ± 0.0009	0.0261 ± 0.0009
0.75	100	0.5	105	26.37 ± 0.463	210.97 ± 6.544	0.2546 ± 0.0044	0.0351 ± 0.0011	0.0236 ± 0.0005
1	100	0.5	105	32.54 ± 0.575	280.16 ± 6.272	0.3225 ± 0.0062	0.0497 ± 0.0014	0.0050 ± 0.0002

Table 5.10: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = $\Gamma(\alpha_b, 1)$, $\Gamma(\alpha_r, 1)$, Kibble's Gamma, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = $\text{Exp}(\mu + \sqrt{\mu} \cdot \beta)$, Service = $\text{Exp}(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	400	−0.5	390	25.52 ± 0.940	461.02 ± 31.962	0.0634 ± 0.0024	0.0031 ± 0.0002	0.0031 ± 0.0003
0.25	400	−0.5	390	25.35 ± 0.168	464.43 ± 9.898	0.0629 ± 0.0004	0.0032 ± 0.0001	0.0031 ± 0.0001
0.5	400	−0.5	390	25.53 ± 0.433	467.24 ± 22.750	0.0634 ± 0.0011	0.0032 ± 0.0001	0.0031 ± 0.0001
0.75	400	−0.5	390	25.69 ± 0.649	478.81 ± 14.044	0.0638 ± 0.0016	0.0032 ± 0.0001	0.0030 ± 0.0001
1	400	−0.5	390	27.89 ± 0.382	591.99 ± 21.773	0.0697 ± 0.0010	0.0039 ± 0.0001	0.0004 ± 0.0000
0	400	0	400	40.90 ± 0.623	809.78 ± 24.267	0.1010 ± 0.0015	0.0069 ± 0.0002	0.0067 ± 0.0002
0.25	400	0	400	41.47 ± 1.197	832.80 ± 28.919	0.1024 ± 0.0030	0.0070 ± 0.0003	0.0067 ± 0.0003
0.5	400	0	400	41.87 ± 0.785	837.89 ± 27.115	0.1035 ± 0.0019	0.0072 ± 0.0003	0.0067 ± 0.0002
0.75	400	0	400	41.81 ± 1.247	856.04 ± 45.522	0.1034 ± 0.0031	0.0071 ± 0.0004	0.0062 ± 0.0003
1	400	0	400	51.76 ± 0.801	1251.72 ± 51.752	0.1292 ± 0.0021	0.0105 ± 0.0003	0.0009 ± 0.0000
0	400	0.5	410	65.45 ± 1.254	1116.06 ± 35.183	0.1605 ± 0.0031	0.0145 ± 0.0005	0.0137 ± 0.0005
0.25	400	0.5	410	66.25 ± 1.633	1121.40 ± 41.443	0.1625 ± 0.0041	0.0148 ± 0.0006	0.0137 ± 0.0005
0.5	400	0.5	410	65.75 ± 1.904	1143.58 ± 48.394	0.1613 ± 0.0047	0.0147 ± 0.0007	0.0131 ± 0.0007
0.75	400	0.5	410	66.85 ± 0.997	1208.90 ± 18.474	0.1643 ± 0.0025	0.0152 ± 0.0003	0.0121 ± 0.0002
1	400	0.5	410	87.71 ± 2.162	1727.66 ± 94.629	0.2188 ± 0.0056	0.0241 ± 0.0011	0.0017 ± 0.0001

5.4 NORTA

In this section, we simulate the balking and reneging random variables using the NORTA method with several different distributions. We use uniform, exponential, and lognormal distributions to simulate the queueing dynamics. Unlike in the previous cases, we cannot compare against our limit theorems see that our limit theorems since the sum of the partial derivatives of the joint distirbution is unknown and cannot be computed in closed form. Thus, we are unable to compare against our heavy traffic limit theorems and new theory needs to be developed for this case as well.

5.4.1 Uniform Marginals

Table 5.11: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = $\text{Unif}(0, 1)$, NORTA, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = $\text{Exp}(\mu + \sqrt{\mu} \cdot \beta)$, Service = $\text{Exp}(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	100	−0.5	95	4.70 ± 0.091	17.03 ± 0.559	0.0439 ± 0.0009	0.0468 ± 0.0010	0.0403 ± 0.0007
0.25	100	−0.5	95	4.75 ± 0.043	17.37 ± 0.206	0.0446 ± 0.0005	0.0476 ± 0.0006	0.0370 ± 0.0006
0.5	100	−0.5	95	4.92 ± 0.024	18.40 ± 0.260	0.0466 ± 0.0002	0.0492 ± 0.0004	0.0325 ± 0.0005
0.75	100	−0.5	95	5.15 ± 0.053	20.09 ± 0.290	0.0493 ± 0.0005	0.0515 ± 0.0008	0.0250 ± 0.0004
1	100	−0.5	95	5.72 ± 0.049	23.93 ± 0.493	0.0564 ± 0.0005	0.0573 ± 0.0004	0.0076 ± 0.0001
0	100	0	100	5.61 ± 0.091	21.20 ± 0.578	0.0519 ± 0.0009	0.0560 ± 0.0010	0.0472 ± 0.0007
0.25	100	0	100	5.71 ± 0.061	21.82 ± 0.292	0.0532 ± 0.0007	0.0572 ± 0.0006	0.0432 ± 0.0006
0.5	100	0	100	5.96 ± 0.041	23.33 ± 0.326	0.0559 ± 0.0004	0.0596 ± 0.0004	0.0378 ± 0.0005
0.75	100	0	100	6.29 ± 0.078	25.76 ± 0.399	0.0599 ± 0.0008	0.0630 ± 0.0008	0.0291 ± 0.0005
1	100	0	100	7.15 ± 0.048	31.43 ± 0.584	0.0704 ± 0.0005	0.0717 ± 0.0003	0.0086 ± 0.0001
0	100	0.5	105	6.68 ± 0.100	25.96 ± 0.641	0.0611 ± 0.0010	0.0666 ± 0.0011	0.0549 ± 0.0008
0.25	100	0.5	105	6.83 ± 0.078	26.81 ± 0.296	0.0629 ± 0.0009	0.0683 ± 0.0009	0.0502 ± 0.0008
0.5	100	0.5	105	7.17 ± 0.053	28.89 ± 0.499	0.0668 ± 0.0006	0.0718 ± 0.0005	0.0439 ± 0.0006
0.75	100	0.5	105	7.68 ± 0.089	32.38 ± 0.349	0.0726 ± 0.0009	0.0769 ± 0.0009	0.0339 ± 0.0007
1	100	0.5	105	8.93 ± 0.051	40.18 ± 0.617	0.0878 ± 0.0006	0.0895 ± 0.0004	0.0097 ± 0.0002

Table 5.12: Simulated Results vs. Heavy Traffic Approximations (Balking, Reneging) = Unif$(0,1)$, NORTA, $\frac{Q}{\mu}$ for Balking, W for Reneging, Arrivals = Exp$(\mu + \sqrt{\mu} \cdot \beta)$, Service = Exp$(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	400	−0.5	390	9.46 ± 0.065	62.65 ± 1.427	0.0228 ± 0.0002	0.0236 ± 0.0003	0.0219 ± 0.0001
0.25	400	−0.5	390	9.65 ± 0.077	65.06 ± 1.145	0.0233 ± 0.0002	0.0240 ± 0.0003	0.0208 ± 0.0002
0.5	400	−0.5	390	9.88 ± 0.060	68.47 ± 0.988	0.0240 ± 0.0002	0.0247 ± 0.0001	0.0185 ± 0.0002
0.75	400	−0.5	390	10.42 ± 0.127	75.72 ± 2.152	0.0254 ± 0.0003	0.0260 ± 0.0004	0.0145 ± 0.0002
1	400	−0.5	390	11.93 ± 0.125	98.25 ± 2.630	0.0297 ± 0.0003	0.0298 ± 0.0003	0.0029 ± 0.0001
0	400	0	400	11.19 ± 0.074	77.86 ± 1.823	0.0268 ± 0.0002	0.0279 ± 0.0003	0.0257 ± 0.0002
0.25	400	0	400	11.44 ± 0.077	81.14 ± 1.245	0.0275 ± 0.0002	0.0285 ± 0.0003	0.0242 ± 0.0002
0.5	400	0	400	11.81 ± 0.056	86.51 ± 0.914	0.0285 ± 0.0001	0.0295 ± 0.0001	0.0216 ± 0.0001
0.75	400	0	400	12.57 ± 0.127	96.86 ± 1.944	0.0306 ± 0.0003	0.0314 ± 0.0004	0.0170 ± 0.0001
1	400	0	400	14.88 ± 0.154	131.38 ± 4.295	0.0370 ± 0.0004	0.0372 ± 0.0003	0.0033 ± 0.0000
0	400	0.5	410	13.27 ± 0.087	95.60 ± 2.152	0.0317 ± 0.0002	0.0330 ± 0.0003	0.0300 ± 0.0002
0.25	400	0.5	410	13.63 ± 0.077	100.42 ± 1.063	0.0326 ± 0.0002	0.0339 ± 0.0003	0.0284 ± 0.0002
0.5	400	0.5	410	14.16 ± 0.069	107.65 ± 1.549	0.0340 ± 0.0002	0.0354 ± 0.0001	0.0253 ± 0.0002
0.75	400	0.5	410	15.25 ± 0.116	122.35 ± 1.780	0.0370 ± 0.0003	0.0381 ± 0.0005	0.0200 ± 0.0001
1	400	0.5	410	18.69 ± 0.159	172.20 ± 4.655	0.0465 ± 0.0004	0.0467 ± 0.0004	0.0037 ± 0.0001

5.4.2 Exponential Marginals

Table 5.13: Simulated Results vs. Heavy Traffic Approximations (Balking, Reneging) = Exp(θ_b), NORTA, $\frac{Q}{\mu}$ for Balking, W for Reneging, Arrivals = Exp$(\mu + \sqrt{\mu} \cdot \beta)$, Service = Exp$(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	100	−0.5	95	2.47 ± 0.018	5.15 ± 0.052	0.0219 ± 0.0002	0.1105 ± 0.0007	0.0650 ± 0.0004
0.25	100	−0.5	95	2.54 ± 0.017	5.41 ± 0.046	0.0229 ± 0.0002	0.1137 ± 0.0006	0.0566 ± 0.0002
0.5	100	−0.5	95	2.63 ± 0.022	5.66 ± 0.080	0.0241 ± 0.0002	0.1173 ± 0.0005	0.0464 ± 0.0005
0.75	100	−0.5	95	2.76 ± 0.021	6.08 ± 0.087	0.0260 ± 0.0002	0.1222 ± 0.0007	0.0329 ± 0.0004
1	100	−0.5	95	2.95 ± 0.010	6.68 ± 0.065	0.0289 ± 0.0001	0.1304 ± 0.0012	0.0092 ± 0.0001
0	100	0	100	2.73 ± 0.014	5.78 ± 0.052	0.0240 ± 0.0001	0.1210 ± 0.0007	0.0701 ± 0.0002
0.25	100	0	100	2.81 ± 0.018	6.06 ± 0.061	0.0252 ± 0.0002	0.1250 ± 0.0006	0.0608 ± 0.0002
0.5	100	0	100	2.92 ± 0.022	6.40 ± 0.077	0.0267 ± 0.0002	0.1294 ± 0.0005	0.0498 ± 0.0005
0.75	100	0	100	3.08 ± 0.021	6.91 ± 0.098	0.0289 ± 0.0002	0.1355 ± 0.0008	0.0352 ± 0.0003
1	100	0	100	3.33 ± 0.017	7.68 ± 0.071	0.0326 ± 0.0002	0.1458 ± 0.0012	0.0095 ± 0.0001
0	100	0.5	105	3.00 ± 0.016	6.43 ± 0.056	0.0262 ± 0.0002	0.1323 ± 0.0007	0.0751 ± 0.0004
0.25	100	0.5	105	3.10 ± 0.021	6.78 ± 0.068	0.0276 ± 0.0002	0.1369 ± 0.0005	0.0651 ± 0.0002
0.5	100	0.5	105	3.24 ± 0.025	7.18 ± 0.080	0.0294 ± 0.0003	0.1422 ± 0.0007	0.0531 ± 0.0005
0.75	100	0.5	105	3.43 ± 0.026	7.78 ± 0.103	0.0320 ± 0.0003	0.1496 ± 0.0011	0.0374 ± 0.0005
1	100	0.5	105	3.74 ± 0.014	8.70 ± 0.069	0.0366 ± 0.0002	0.1620 ± 0.0011	0.0096 ± 0.0002

Table 5.14: Simulated Results vs. Heavy Traffic Approximations (Balking, Reneging) = Exp(θ_b), NORTA, $\frac{Q}{\mu}$ for Balking, W for Reneging, Arrivals = Exp$(\mu + \sqrt{\mu} \cdot \beta)$, Service = Exp$(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	400	−0.5	390	4.94 ± 0.025	17.76 ± 0.137	0.0116 ± 0.0001	0.0588 ± 0.0003	0.0400 ± 0.0004
0.25	400	−0.5	390	5.06 ± 0.030	18.52 ± 0.143	0.0119 ± 0.0001	0.0599 ± 0.0004	0.0360 ± 0.0002
0.5	400	−0.5	390	5.25 ± 0.036	19.81 ± 0.254	0.0125 ± 0.0001	0.0621 ± 0.0003	0.0303 ± 0.0002
0.75	400	−0.5	390	5.51 ± 0.024	21.46 ± 0.134	0.0133 ± 0.0001	0.0648 ± 0.0003	0.0215 ± 0.0002
1	400	−0.5	390	6.15 ± 0.027	25.95 ± 0.256	0.0153 ± 0.0001	0.0720 ± 0.0003	0.0032 ± 0.0000
0	400	0	400	5.39 ± 0.030	19.79 ± 0.180	0.0125 ± 0.0001	0.0639 ± 0.0003	0.0431 ± 0.0004
0.25	400	0	400	5.53 ± 0.032	20.71 ± 0.184	0.0130 ± 0.0001	0.0652 ± 0.0004	0.0387 ± 0.0002
0.5	400	0	400	5.76 ± 0.043	22.23 ± 0.315	0.0137 ± 0.0001	0.0679 ± 0.0003	0.0325 ± 0.0002
0.75	400	0	400	6.08 ± 0.027	24.28 ± 0.175	0.0146 ± 0.0001	0.0713 ± 0.0003	0.0231 ± 0.0002
1	400	0	400	6.86 ± 0.024	29.69 ± 0.234	0.0170 ± 0.0001	0.0800 ± 0.0002	0.0032 ± 0.0000
0	400	0.5	410	5.88 ± 0.033	21.96 ± 0.182	0.0136 ± 0.0001	0.0694 ± 0.0004	0.0464 ± 0.0004
0.25	400	0.5	410	6.05 ± 0.035	23.05 ± 0.175	0.0141 ± 0.0001	0.0711 ± 0.0005	0.0416 ± 0.0002
0.5	400	0.5	410	6.32 ± 0.044	24.85 ± 0.354	0.0149 ± 0.0001	0.0742 ± 0.0003	0.0349 ± 0.0003
0.75	400	0.5	410	6.71 ± 0.031	27.27 ± 0.203	0.0161 ± 0.0001	0.0783 ± 0.0004	0.0248 ± 0.0002
1	400	0.5	410	7.66 ± 0.029	33.80 ± 0.241	0.0190 ± 0.0001	0.0889 ± 0.0003	0.0033 ± 0.0000

5.4.3 Lognormal Marginals

Table 5.15: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Logn(μ_b, σ_b), NORTA, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp$(\mu + \sqrt{\mu} \cdot \beta)$, Service = Exp$(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	100	−0.5	95	3.08 ± 0.021	6.82 ± 0.072	0.0276 ± 0.0002	0.0750 ± 0.0007	0.0556 ± 0.0007
0.25	100	−0.5	95	3.15 ± 0.020	7.05 ± 0.051	0.0286 ± 0.0002	0.0769 ± 0.0008	0.0486 ± 0.0006
0.5	100	−0.5	95	3.25 ± 0.012	7.37 ± 0.045	0.0299 ± 0.0001	0.0806 ± 0.0001	0.0415 ± 0.0005
0.75	100	−0.5	95	3.34 ± 0.021	7.73 ± 0.068	0.0313 ± 0.0002	0.0844 ± 0.0008	0.0326 ± 0.0001
1	100	−0.5	95	3.50 ± 0.014	8.21 ± 0.015	0.0335 ± 0.0002	0.0899 ± 0.0004	0.0202 ± 0.0001
0	100	0	100	3.42 ± 0.022	7.57 ± 0.076	0.0304 ± 0.0002	0.0864 ± 0.0006	0.0620 ± 0.0007
0.25	100	0	100	3.51 ± 0.021	7.88 ± 0.034	0.0316 ± 0.0002	0.0894 ± 0.0008	0.0540 ± 0.0006
0.5	100	0	100	3.64 ± 0.016	8.25 ± 0.044	0.0332 ± 0.0002	0.0939 ± 0.0001	0.0458 ± 0.0004
0.75	100	0	100	3.76 ± 0.024	8.70 ± 0.069	0.0350 ± 0.0002	0.0986 ± 0.0011	0.0359 ± 0.0003
1	100	0	100	3.96 ± 0.014	9.28 ± 0.033	0.0378 ± 0.0002	0.1062 ± 0.0004	0.0216 ± 0.0001
0	100	0.5	105	3.79 ± 0.020	8.32 ± 0.065	0.0332 ± 0.0002	0.0989 ± 0.0007	0.0685 ± 0.0005
0.25	100	0.5	105	3.89 ± 0.023	8.71 ± 0.046	0.0348 ± 0.0003	0.1028 ± 0.0008	0.0595 ± 0.0006
0.5	100	0.5	105	4.05 ± 0.014	9.14 ± 0.039	0.0368 ± 0.0001	0.1086 ± 0.0004	0.0503 ± 0.0004
0.75	100	0.5	105	4.21 ± 0.024	9.67 ± 0.082	0.0390 ± 0.0003	0.1144 ± 0.0012	0.0391 ± 0.0003
1	100	0.5	105	4.45 ± 0.018	10.35 ± 0.036	0.0424 ± 0.0002	0.1237 ± 0.0006	0.0229 ± 0.0003

Table 5.16: Simulated Results vs. Heavy Traffic Approximations
(Balking, Reneging) = Logn(μ_b, σ_b), NORTA, $\frac{Q}{\mu}$ for Balking, W for Reneging,
Arrivals = Exp$(\mu + \sqrt{\mu} \cdot \beta)$, Service = Exp$(\mu)$

ρ	μ	β	λ	Q_{sim}	V_{sim}	W_{sim}	B_{sim}	R_{sim}
0	400	−0.5	390	7.41 ± 0.039	32.26 ± 0.409	0.0176 ± 0.0001	0.0294 ± 0.0003	0.0257 ± 0.0003
0.25	400	−0.5	390	7.56 ± 0.034	33.46 ± 0.287	0.0181 ± 0.0001	0.0305 ± 0.0003	0.0234 ± 0.0002
0.5	400	−0.5	390	7.70 ± 0.050	34.87 ± 0.358	0.0185 ± 0.0001	0.0317 ± 0.0003	0.0200 ± 0.0003
0.75	400	−0.5	390	8.01 ± 0.067	37.31 ± 0.444	0.0194 ± 0.0002	0.0340 ± 0.0004	0.0157 ± 0.0003
1	400	−0.5	390	8.52 ± 0.065	41.55 ± 0.530	0.0210 ± 0.0002	0.0381 ± 0.0006	0.0076 ± 0.0001
0	400	0	400	8.24 ± 0.045	35.80 ± 0.440	0.0195 ± 0.0001	0.0350 ± 0.0004	0.0299 ± 0.0003
0.25	400	0	400	8.43 ± 0.035	37.30 ± 0.271	0.0200 ± 0.0001	0.0365 ± 0.0003	0.0271 ± 0.0002
0.5	400	0	400	8.63 ± 0.051	39.05 ± 0.369	0.0206 ± 0.0001	0.0381 ± 0.0003	0.0232 ± 0.0003
0.75	400	0	400	9.01 ± 0.073	41.96 ± 0.458	0.0218 ± 0.0002	0.0411 ± 0.0004	0.0181 ± 0.0003
1	400	0	400	9.68 ± 0.069	47.15 ± 0.512	0.0238 ± 0.0002	0.0466 ± 0.0007	0.0085 ± 0.0002
0	400	0.5	410	9.13 ± 0.049	39.30 ± 0.475	0.0215 ± 0.0001	0.0413 ± 0.0004	0.0344 ± 0.0003
0.25	400	0.5	410	9.37 ± 0.035	41.03 ± 0.275	0.0222 ± 0.0001	0.0433 ± 0.0003	0.0312 ± 0.0002
0.5	400	0.5	410	9.63 ± 0.056	43.13 ± 0.389	0.0229 ± 0.0001	0.0454 ± 0.0004	0.0267 ± 0.0003
0.75	400	0.5	410	10.11 ± 0.082	46.54 ± 0.458	0.0243 ± 0.0002	0.0493 ± 0.0006	0.0208 ± 0.0003
1	400	0.5	410	10.95 ± 0.085	52.64 ± 0.466	0.0269 ± 0.0002	0.0565 ± 0.0008	0.0094 ± 0.0002

6 Conclusions and Extensions

In this paper, we analyze the dynamics of a critically loaded single server unobservable queue where customers can either balk because the line is too long or abandon from the queue if their wait is too excessive. In a ticket queue, a reneging customer is only detected when their hypothetical service time would have begun. One important feature is that the distribution of balking and reneging random variables are dependent. This is realistic since when a customer joins a long line, they would naturally be more likely to wait longer, which creates a positive dependence that needs to be studied.

In order to develop approximations for the unobservable queue with dependent balking and reneging random variables, we prove a heavy traffic limit theorem for the diffusion scaled queue length and workload processes. Our main result illustrates that the unobservable queue with dependent balking and reneging random variables can be approximated with a reflected Ornstein-Uhlenbeck process, which has a steady state distribution that is given by a truncated normal distribution. What makes the dependent case different from the independent case is that the joint distribution near the origin shifts the restorative drift of the regulated OU process causing less customers to renege from the queue and increasing the number of customers in the queue.

To assess the accruacy of our approximations, we appeal to simulation. By performing extensive simulation studies, we find that in a broad range of parameter and distributional settings, our approximations based

on the steady state distribution are very accurate. One extension of our work is to prove an interchange of limits for the steady state queue length. This would provide even further evidence and support for our steady state approximations. Moreover, it is also worth considering an even more realistic situation where customers wander and return (possibly late) for service. We will consider these extensions in future work.

A Appendix

A.1 Preliminary and Standard Results

Lemma A.1. *Let $\{X_i, i \geq 1\}$ be an i.i.d sequence of random variables with finite variance σ^2. Then for each $T \geq 0$,*

$$\lim_{n\to\infty} E\left[\sup_{i=1,\ldots,nT} \frac{|X_i|}{\sqrt{n}}\right] \to 0 \tag{A.1}$$

Proof. See Jennings and Reed [18].

□

Lemma A.2. *(i) For any $t \geq 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(A^n(t) > 2\lambda nt\right) = 0.$$

(ii) For any $\varepsilon, K, t \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{i\leq Knt} v^n(i) > \frac{\varepsilon}{\sqrt{n}}\right) = 0.$$

(iii) For any ε and $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\leq t} \left|\bar{A}^n(s) - \mu s\right| > \varepsilon\right) = 0.$$

(iv) For any ε and $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\leq t} \left|\sum_{i=1}^{A^n(s)} v^n(i) - s\right| > \varepsilon\right) = 0.$$

(v) For any ε and $t > 0$, there exists a $\delta > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{u<v\leq t, v-u<\delta} \left|\sum_{i=A^n(u)+1}^{A^n(v)} v^n(i) - (v-u)\right| > \frac{\varepsilon}{\sqrt{n}}\right) = 0.$$

(vi) For any δ, η, t and $K > 0$ we have that

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{s\in[0,t]} A^n(s + K/\sqrt{n}) - A(s) > \lambda(K+\delta)\sqrt{n}\right) < \eta.$$

Proof. See Appendix of [17]. □

Lemma A.3. *For any $\varepsilon, \eta > 0$ we have that*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \sum_{i=A^n((s-\delta)^+)}^{A^n(s)} 1(r_i \leq K/\sqrt{n}) > \sqrt{n}\varepsilon\right) \quad < \quad \eta$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \sum_{i=A^n((s-\delta)^+)}^{A^n(s)} 1(b_i \leq K/\sqrt{n}) \cdot 1(r_i \leq K/\sqrt{n}) > \sqrt{n}\varepsilon\right) \quad < \quad \eta$$

Let B_a and B_s be independent, standard Brownian motions; i.e., $B_a(0) = B_s(0) = 0$ and the processes have zero drift and unitary infinitesimal variance. The diffusion scaled arrival processes and service completion processes converge to scaled versions of these Brownian motions. These are standard results in the queueing literature see, for example, [17, 18]

Proposition A.4. *Under the assumptions of Theorem 4.1,*

$$\tilde{A}^n \Rightarrow \mu\sigma_a B_a \quad \text{and} \quad \tilde{S}^n_\alpha \Rightarrow \mu\sigma_s B_s$$

as $n \to \infty$.

The following lemmas, which besides Lemma A.6, are provided without proof, use the fact that the derivatives of the balking and abandonment distributions exist at zero; see (2.2). The first lemma is used throughout this section and follows from a straightforward application of Taylor's Expansion.

Lemma A.5. *For any $K > 0$,*

$$\frac{G_b(K/\sqrt{n})}{K/\sqrt{n}} + \frac{G_r(K/\sqrt{n})}{K/\sqrt{n}} - \frac{G(K/\sqrt{n}, K/\sqrt{n})}{K/\sqrt{n}} < 2\gamma$$

for sufficiently large n.

The second lemma is similar.

Lemma A.6. *For any $\delta, K > 0$,*

$$\sup_{s\in[0,K]} \left(\frac{G_b\left(\frac{s+\delta}{\sqrt{n}}\right) - G_b\left(\frac{s}{\sqrt{n}}\right) + G_r\left(\frac{s+\delta}{\sqrt{n}}\right) - G_r\left(\frac{s}{\sqrt{n}}\right) + G\left(\frac{s+\delta}{\sqrt{n}}, \frac{s+\delta}{\sqrt{n}}\right) - G\left(\frac{s}{\sqrt{n}}, \frac{s}{\sqrt{n}}\right)}{\delta/\sqrt{n}} \right) < 2\gamma$$

for sufficiently large n.

A.2 Asymptotic Boundedness

We argue first that the scaled queue length processes and the workload processes are asymptotically bounded.

Lemma A.7. *Under (4.30), we have that for any $t, \eta > 0$ there exists a $K = K(\eta) > 0$ such that,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{Q}^n(s) > K \right) < \eta \tag{A.2}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}^n(s) > K \right) < \eta. \tag{A.3}$$

Proof. The proof is identical to that in [17]. □

Lemma A.7 emphasizes the orders of magnitude of the queueing and workload processes. Namely, the queue lengths weighted by the processing time $1/\mu^n$ and workload processes convergeto zero at rate $1/\sqrt{n}$. This lemma will be used frequently in conjunction with the balking and abandonment distributions to place bounds on abandonment and balking frequencies.

So far our propositions have been able to replace the queueing and workload processes with upper bounds early in the proofs. For the following result, where we show that the centered and scaled balking and approximate abandonment processes converge to zero, such substitutions cannot be made immediately.

Proposition A.8. *Centered balking and reneging processes are negligible. Under the assumptions of Theorem 4.1, and any $\varepsilon, \eta, t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{M}^n_b(s)\right| > \varepsilon \right) < \eta \tag{A.4}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{M}^n_r(s)\right| > \varepsilon \right) < \eta. \tag{A.5}$$

Proof. In order to prove this result, one needs to use Kolmogorov's inequality and Burkholder's inequality like in Proposition 5.6 in [17]. We omit the details since they are roughly identical to that proof. □

The implications here are that the balking and reneging random variables can be replaced with their respective distribution functions.

A.3 State Space Collapse

The purpose of this section to show the relationship between the queue length and the workload process. The asymptotic boundedness shown in Lemma A.7 shows that the queue length is of the order $\sqrt{n}$. In queueing a model where no customers can balk or abandon, this fact, via the weak law of large numbers, is clearly sufficient to draw a linear relationship between the queue length and the workload of the form $Q/\mu \approx W$. We will also show this relationship even though we have balking, abandonment, and clearing times in our unobservable queue. The main ingredient in showing this linear relationship is that the number of jobs in queue who do not contribute to the workload is negligible with respect to $\sqrt{n}$. Thus, we have the following result showing that the queue length and workload processes converge to the same limit, albeit scalar multiples of each other.

Proposition A.9. *State space collapse. Under (4.30), we have that for any $t, \varepsilon, \eta > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{Q}^n(s) - \mu\tilde{W}^n(s)\right| > \varepsilon\right) < \eta.$$

Proof. Proposition A.9 In order to prove state space collapse for the queue length and workload process we will use an argument that exploits the fact that the number of balking and reneging customers is negligible in the limit. First we fix $t > 0$ and for each $s \geq 0$, let $\hat{Q}^n(s)$ denote the difference between the index of the last arriving job and the index of the job currently in service. Note that, for each $s \geq 0$, the inflated queue can be constructed such that $\hat{Q}^n(s) \geq Q^n(s)$. The process $\hat{Q}^n$ ignores the balking and abandonment that has taken place since the arrival of the job currently in service. One can think of the process as progressing in a manner similar to a ticket queue for which, in addition to the abandoned tickets, balking is not accounted for until service would have begun for the departing job. First, we start with the triangle inequality where our state space collapse result can be bounded by

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{Q}^n(s) - \mu\tilde{W}^n(s)\right| > \varepsilon\right) \\
&= \mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{Q}^n(s) + \hat{Q}^n(s) - \hat{Q}^n(s) + \hat{W}^n(s) - \hat{W}^n(s) - \mu\tilde{W}^n(s)\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(\sup_{s\in[0,t]} \left|Q^n(s) - \hat{Q}^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right) + \mathbb{P}\left(\mu^n \sup_{s\in[0,t]} \left|W^n(s) - \hat{W}^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right) \\
&\quad + \mathbb{P}\left(\sup_{s\in[0,t]} \left|\hat{Q}^n(s) - \mu^n\hat{W}^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right) \\
&< \eta
\end{aligned}$$

for sufficiently large n. Then, we note by the functional weak law of large numbers that,

$$\mathbb{P}\left(\sup_{s\in[0,t]} \left|\hat{Q}^n(s) - \mu^n\hat{W}^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right) < \frac{\eta}{3} \tag{A.6}$$

for sufficiently large n. Now it remains to bound the difference between the queue length and workload processes with their inflated versions. We begin with the queue length and its inflated version. The rest of the proof follows from a similar argument of state space collapse from [17]. □

Our next result is to show that the service idleness process I^n converges to zero. However, we need a lemma that measures for a sufficiently small time interval the amount of potential workload contribution associated with balking or abandoning jobs arriving during the interval is smaller than order $1/\sqrt{n}$. This result is a key element in the proof of tightness of the scaled workload processes; see Proposition A.13.

Lemma A.10. *For any $\eta, t > 0$ and $K > 0$, there exists a δ such that*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\le t} \sum_{i=A^n(s)+1}^{A^n(s+\delta)} v_i^n \cdot \left(1(b_i \le K/\sqrt{n}) + 1(r_i \le K/\sqrt{n})\right) > \frac{\varepsilon}{\sqrt{n}}\right) < \eta. \tag{A.7}$$

Proposition A.11. *Convergence of the server busy process. Under the assumptions of Theorem 4.1, and any $\varepsilon, \eta, t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} I^n(s) > \varepsilon\right) < \eta.$$

Proof. Fix ε, η, and $t > 0$. The server cannot work faster than rate one. It follows that for each $s \le t$,

$$T^n(s) \le s. \tag{A.8}$$

Moreover, we also have the following equality for the idleness and service allocation processes

$$T^n(s) + I^n(s) = s. \tag{A.9}$$

Using (4.6) we can provide a lower bound on the service allocation process for any $s \ge 0$,

$$T^n(s) \ge -W^n(s) + \sum_{i=1}^{A^n(s)} v_i^n \cdot 1(b_i > Q^n(t_i-)/\mu^n) \cdot 1(r_i > W^n(t_i-)). \tag{A.10}$$

It follows by rearranging A.9, and (A.10) that

$$\begin{aligned}
\mathbb{P}\left(\sup_{s\in[0,t]} |T^n(s) - s| > \varepsilon\right) &= \mathbb{P}\left(\sup_{s\in[0,t]} T^n(s) < s - \varepsilon\right) \qquad (A.11)\\
&\le \mathbb{P}\left(\sup_{s\in[0,t]} W^n(s) > \frac{\varepsilon}{4}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} \left|\sum_{i=1}^{A^n(s)} v_i^n - s\right| > \frac{\varepsilon}{4}\right)\\
&\quad + \mathbb{P}\left(\sum_{i=1}^{A^n(t)} v_i^n \cdot (1(b_i < Q^n(t_i-)/\mu^n) + 1(r_i < W^n(t_i-))) > \frac{\varepsilon}{4}\right)\\
&\quad + \mathbb{P}\left(\sum_{i=1}^{A^n(t)} v_i^n \cdot 1(b_i < Q^n(t_i-)/\mu^n) \cdot 1(r_i < W^n(t_i-)) > \frac{\varepsilon}{4}\right)\\
&\le \mathbb{P}\left(\sup_{s\in[0,t]} W^n(s) > \frac{\varepsilon}{4}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} \left|\sum_{i=1}^{A^n(s)} v_i^n - s\right| > \frac{\varepsilon}{4}\right)\\
&\quad + 2 \cdot \mathbb{P}\left(\sum_{i=1}^{A^n(t)} v_i^n \cdot (1(b_i < Q^n(t_i-)/\mu^n) + 1(r_i < W^n(t_i-))) > \frac{\varepsilon}{4}\right)
\end{aligned}$$

For sufficiently large n, and from Lemma A.2 and Lemma A.7, we can bound the first two terms on the right hand side,

$$\mathbb{P}\left(\sup_{s\in[0,t]} W^n(s) > \frac{\varepsilon}{4}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} \left|\sum_{i=1}^{A^n(s)} v_i^n - s\right| > \frac{\varepsilon}{4}\right) < \frac{\eta}{2}. \tag{A.12}$$

For the third term, by Lemma A.2 and Proposition A.10,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{A^n(t)} v_i^n \cdot (1(b_i < Q^n(t_i-)/\mu^n) + 1(r_i < W^n(t_i-))) > \frac{\varepsilon}{4}\right) \qquad (A.13)\\
&\leq \mathbb{P}\left(A^n(t) > 2\mu n t\right) + \mathbb{P}\left(\sum_{i=1}^{2\mu n t} v_i^n \cdot (1(b_i < Q^n(t_i-)/\mu^n) + 1(r_i < W^n(t_i-))) > \frac{\varepsilon}{4}\right)\\
&\leq \mathbb{P}\left(\sum_{i=1}^{2\mu n t} \left(v_i^n - \frac{1}{\mu n}\right) \cdot (1(b_i < Q^n(t_i-)/\mu^n) + 1(r_i < W^n(t_i-))) > \frac{\varepsilon}{12}\right)\\
&+ \mathbb{P}\left(\sum_{i=1}^{2\mu n t} \frac{1}{\mu n} \cdot (1(b_i < Q^n(t_i-)/\mu^n) - G_b(Q^n(t_i-)/\mu^n) + 1(r_i < W^n(t_i-)) - G_r(W^n(t_i-))) > \frac{\varepsilon}{12}\right)\\
&+ \mathbb{P}\left(\sum_{i=1}^{2\mu n t} \frac{1}{\mu n} \cdot (G_b(Q^n(t_i-)/\mu^n) + G_r(W^n(t_i-))) > \frac{\varepsilon}{12}\right)\\
&< \frac{\eta}{4}.
\end{aligned}$$

The above result follows from (A.10) - (A.13) and a modification of the proofs of Propositions A.3 - A.8. This completes the proof.

□

A.4 Reneging is Asymptotically Equivalent to Balking

Recall that a customer can only renege from the queue if it decides to join the queue. This creates a complication that makes for involved expressions for the reneging processes in (4.4). Furthermore, notice that the expressions include both the queue length process as well as the workload process. The process $\tilde{\varepsilon}_q^n$ allows one to replace the workload process with the queue length process, to ignore whether or not the jobs have balked when considering whether they will abandon, and to ignore the timing of when reneging is detected by the system. The following proposition justifies this approximation.

For each $n \geq 1$, we define the processes $R^{1,n} = \{R^{1,n}(t), t \geq 0\}$ and $R^{2,n} = \{R^{2,n}(t), t \geq 0\}$, where

$$R^{1,n}(t) = \sum_{i=1}^{A^n(t)} 1(r_i \leq Q^n(t_i^n-)/\mu^n) - 1(b_i \leq Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \leq Q^n(t_i^n-)/\mu^n) \qquad (A.14)$$

and

$$R^{2,n}(t) = \sum_{i=1}^{A^n(t)} 1(r_i \leq W^n(t_i^n-)) - 1(b_i \leq Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \leq W^n(t_i^n-)). \qquad (A.15)$$

One should note that notice that $R^{2,n}$ removes the indicator function that makes the customer only leave when they would have entered service; this is one of the essential parts of the ticket queue as the customer is only detected when they would have entered service. Next, $R^{1,n}$ is the fictitious reneging process whereby reneging happens upon arrival and the reneging process also depends on the scaled queue length instead of the workload. The diffusion scaled analogs have the form $\tilde{R}^{k,n} = \{\tilde{R}^{k,n}(t), t \geq 0\}$, where $\tilde{R}^{k,n}(t) = (1/\sqrt{n})R^{k,n}(t)$ for each $k = 1, 2$. Ultimately, we would like to replace R^n with $R^{1,n}$; see Proposition A.12. Thus, the following proposition allows us to make this conclusion rigorously.

Proposition A.12. *Reneging effectively ignores balking and takes place upon arrival. Under the assumptions of Theorem 4.1,*

$$\tilde{\varepsilon}_q^n \to 0$$

in probability as $n \to \infty$.

Proof. Notice that for any $t \geq 0$ and α that

$$\tilde{\varepsilon}_q^n(t) = \left(\tilde{R}^n(t) - \tilde{R}^{2,n}(t)\right) + \left(\tilde{R}^{2,n}(t) - \tilde{R}^{1,n}(t)\right).$$

Thus, it suffices to show the following, for any ε, η and $t > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{R}^{1,n}(s) - \tilde{R}^{2,n}(s)\right| > \varepsilon\right) < \eta,$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{R}^{2,n}(s) - \tilde{R}^{n}(s)\right| > \varepsilon\right) < \eta.$$

Fix ε, η and $t > 0$. First notice that for any $n \geq 0$, and $s \geq 0$, we have that $R^{2,n}(s) - R^n(s) \geq 0$. By expanding the difference we have that

$$\begin{aligned}
R^{2,n}(s) - R^n(s) &= \sum_{i=1}^{A^n(s)} 1(r_i \leq W^n(t_i^n)) - \sum_{i=1}^{A^n(s)} 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) \leq s - t_i^n) \\
&- \sum_{i=1}^{A^n(s)} 1(b_i \leq Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \leq W^n(t_i^n)) \\
&+ \sum_{i=1}^{A^n(s)} 1(b_i \leq Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) \leq s - t_i^n) \\
&= \sum_{i=1}^{A^n(s)} 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) > s - t_i^n) \\
&- \sum_{i=1}^{A^n(s)} 1(b_i \leq Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) > s - t_i^n) \\
&\leq \sum_{i=1}^{A^n(s)} 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) > s - t_i^n).
\end{aligned}$$

Note that we have eliminated the indicator associated with balking. Moreover, for an abandoning customer the workload upon arrival must exceed the patience quantity. Hence we can replace the patience quantity in the last of the indicators with the workload upon arrival.

Both the standard and the ticket queue have the same bounds:

$$\mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{R}^{2,n}(s) - \tilde{R}^{n}(s)\right| > \varepsilon\right) \tag{A.16}$$

$$\leq \mathbb{P}\left(\sup_{s\in[0,t]} \sum_{i=1}^{A^n(s)} 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) > s - t_i^n) > \sqrt{n}\varepsilon/2\right) \tag{A.17}$$

$$+ \mathbb{P}\left(\sup_{s\in[0,t]} \sum_{i=1}^{A^n(s)} 1(b_i \leq Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \leq W^n(t_i^n)) \cdot 1(W^n(t_i^n) > s - t_i^n) > \sqrt{n}\varepsilon/2\right). \tag{A.18}$$

Next we replace the workload quantities with an upper bound. This will restrict how far in the past the

summation is taken over for the arrivals.

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}^{2,n}(s)-\tilde{R}^{n}(s)\right|>\varepsilon\right)\\
&\leq \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=1}^{A^n(s)}1(r_i\leq K/\sqrt{n})\cdot 1(K/\sqrt{n}>s-t_i^n)>\sqrt{n}\varepsilon/2\right)\\
&+ \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=1}^{A^n(s)}1(b_i\leq K/\sqrt{n})\cdot 1(r_i\leq K/\sqrt{n})\cdot 1(K/\sqrt{n}>s-t_i^n)>\sqrt{n}\varepsilon/2\right)\\
&\quad +\mathbb{P}\left(\sup_{s\in[0,t]}W^n(s)>K/\sqrt{n}\right)+\mathbb{P}\left(\sup_{s\in[0,t]}Q^n(s)>K\sqrt{n}\right)\\
&\leq \frac{\eta}{3}+\mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=1}^{A^n(s)}1(r_i\leq K/\sqrt{n})\cdot 1(K/\sqrt{n}>s-t_i^n)>\sqrt{n}\varepsilon/2\right)\\
&+ \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=1}^{A^n(s)}1(b_i\leq K/\sqrt{n})\cdot 1(r_i\leq K/\sqrt{n})\cdot 1(K/\sqrt{n}>s-t_i^n)>\sqrt{n}\varepsilon/2\right)
\end{aligned}
$$

Notice that in the first term on the righthand side above, the only jobs that contribute positively to the summation are those jobs i whose arrival time is after $s-\tilde{K}/\sqrt{n}$; that $t_i^n>s-\tilde{K}/\sqrt{n}$. By Lemma A.7,

$$
\begin{aligned}
\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}^{2,n}(s)-\tilde{R}^{n}(s)\right|>\varepsilon\right) &\leq \frac{\eta}{3}+\mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\tilde{K}/\sqrt{n})^+)}^{A^n(s)}1(r_i\leq K/\sqrt{n})>\sqrt{n}\varepsilon/2\right)\\
&+ \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\tilde{K}/\sqrt{n})^+)}^{A^n(s)}1(b_i\leq K/\sqrt{n})\cdot 1(r_i\leq K/\sqrt{n})>\sqrt{n}\varepsilon/2\right)
\end{aligned}
$$

for sufficiently large n. For any arbitrarily chosen $\delta>0$ it is true that $\delta>\tilde{K}/\sqrt{n}$ for sufficiently large n. It follows then by Lemma A.3 that we can choose a $\delta>0$ so that

$$
\begin{aligned}
\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}^{2,n}(s)-\tilde{R}^{n}(s)\right|>\varepsilon\right) &\leq \frac{\eta}{3}+\mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\tilde{K}/\sqrt{n})^+)}^{A^n(s)}1(r_i\leq K/\sqrt{n})>\sqrt{n}\varepsilon/2\right)\\
&+ \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\tilde{K}/\sqrt{n})^+)}^{A^n(s)}1(b_i\leq K/\sqrt{n})\cdot 1(r_i\leq K/\sqrt{n})>\sqrt{n}\varepsilon/2\right)\\
&\leq \frac{\eta}{3}+\mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\delta)^+)}^{A^n(s)}1(r_i\leq K/\sqrt{n})>\sqrt{n}\varepsilon/2\right)\\
&+ \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\delta)^+)}^{A^n(s)}1(b_i\leq K/\sqrt{n})\cdot 1(r_i\leq K/\sqrt{n})>\sqrt{n}\varepsilon/2\right)\\
&< \eta
\end{aligned}
$$

for sufficiently large n. This completes the proof for the first term. Once again we fix ε,η and $t>0$. First notice that for any $n\geq 0$, and $s\geq 0$, we have that $R^{1,n}(s)-R^{2,n}(s)\geq 0$. By expanding the difference we have that

$$
\begin{aligned}
R^{1,n}(s) - R^{2,n}(s) &= \sum_{i=1}^{A^n(s)} 1(r_i \le W^n(t_i^n)) - \sum_{i=1}^{A^n(s)} 1(r_i \le Q^n(t_i^n-)/\mu^n) \\
&- \sum_{i=1}^{A^n(s)} 1(b_i \le Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \le W^n(t_i^n)) \\
&+ \sum_{i=1}^{A^n(s)} 1(b_i \le Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \le Q^n(t_i^n-)/\mu^n).
\end{aligned}
$$

This implies that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]} \left| R^{1,n}(s) - R^{2,n}(s) \right| > \varepsilon\sqrt{n}/2 \right) \\
&\le \mathbb{P}\left(\sup_{s\in[0,t]} |Q^n(s)/\mu^n - W^n(s)| > \frac{\delta}{\sqrt{n}} \right) \\
&+ \mathbb{P}\left(\sup_{s\in[0,t]} \left| \sum_{i=1}^{A^n(s)} 1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \right| > \varepsilon\sqrt{n}/2 \right) \\
&+ \mathbb{P}\left(\sup_{s\in[0,t]} \left| \sum_{i=1}^{A^n(s)} 1(b_i \le Q^n(t_i^n-)/\mu^n) \cdot 1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \right| > \varepsilon\sqrt{n}/2 \right) \\
&\le \mathbb{P}\left(\sup_{s\in[0,t]} |Q^n(s)/\mu^n - W^n(s)| > \frac{\delta}{\sqrt{n}} \right) + \mathbb{P}\left(A^n(s) > 2n\lambda s \right) \\
&+ 2\cdot\mathbb{P}\left(\sup_{s\in[0,t]} \left| \sum_{i=1}^{\lfloor 2n\lambda s\rfloor} 1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \right| > \varepsilon\sqrt{n}/2 \right) \\
&\le \frac{\eta}{2} + 2\cdot\mathbb{P}\left(\sup_{s\in[0,t]} \left| \sum_{i=1}^{\lfloor 2n\lambda s\rfloor} 1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \right| > \varepsilon\sqrt{n}/2 \right).
\end{aligned}
$$

Now it remains to show that

$$
\mathbb{P}\left(\sup_{s\in[0,t]} \left| \sum_{i=1}^{\lfloor 2n\lambda s\rfloor} 1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \right| > \varepsilon\sqrt{n}/2 \right) < \eta/4.
$$

This can be shown by part (iii) of Lemma A.2,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{2n\lambda t} 1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) > \sqrt{n}\varepsilon/2 \right) \\
&\le \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \mathbb{E}\left[1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \middle| \mathcal{F}_{i-1}^n \right] > \frac{\sqrt{n}\varepsilon}{4} \right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \Big(1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \right. \\
&\qquad \left. - \mathbb{E}\left[1(r_i \in [Q^n(t_i^n-)/\mu^n - \delta/\sqrt{n}, Q^n(t_i^n-)/\mu^n + \delta/\sqrt{n}]) \middle| \mathcal{F}_{i-1}^n \right] \Big) > \frac{\sqrt{n}\varepsilon}{4} \right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})-G_r(Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n})\right)>\frac{\sqrt{n}\varepsilon}{4}\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(r_i\in[Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n},Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n}])\right.\right.\\
&\qquad\left.\left.-\left(G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})-G_r(Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n})\right)\right)>\frac{\sqrt{n}\varepsilon}{4}\right)
\end{aligned}
\tag{A.19}
$$

As for the second term on the far right hand side of (A.19), notice that each summand contributes an amount equal to the increase of the abandonment distribution function over an interval of length 2δ. By Equation (A.19) and Lemma A.6,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})-G_r(Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n}])\right)>\frac{\sqrt{n}\varepsilon}{4}\right)\\
&\leq \mathbb{P}\left(\sup_{s\in[0,t]}\tilde{W}^n(s)>K\right)+1\left(2\mu nt\sup_{s\in[0,K]}\left(G_r((s+2\delta)/\sqrt{n})-G_r(s/\sqrt{n})\right)>\frac{\sqrt{n}\varepsilon}{4}\right)\\
&< \frac{\eta}{8}
\end{aligned}
\tag{A.20}
$$

for sufficiently large n. Now consider the third term on the right hand side of (A.19). The following steps are similar to those in the proof of Proposition A.8. By Lemma A.5,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(r_i\in[Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n},Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n}])\right.\right.\\
&\qquad\left.\left.-\left(G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})-G_r(Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n})\right)\right)>\frac{\sqrt{n}\varepsilon}{4}\right)\\
&\leq \frac{16}{n\varepsilon^2}\mathbb{E}\left[\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(r_i\in[Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n},Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n}])\right.\right.\right.\\
&\qquad\left.\left.\left.-\left(G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})-G_r(Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n})\right)\right)\right)^2\right]\\
&\leq \frac{16}{n\varepsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(r_i\in[Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n},Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n}])\right.\right.\\
&\qquad\left.\left.-\left(G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})-G_r(Q^n(t_i^n-)/\mu^n-\delta/\sqrt{n})\right)\right)^2\right]\\
&\leq \frac{32}{n\varepsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor 2\mu nt\rfloor}G_r(Q^n(t_i^n-)/\mu^n+\delta/\sqrt{n})\right]\\
&\leq \frac{64\mu t}{\varepsilon^2}\left(\mathbb{P}\left(\sup_{s\in[0,t]}W^n(s)>K/\sqrt{n}\right)+G_r((K+\delta)/\sqrt{n})\right)\\
&< \frac{\eta}{4},
\end{aligned}
\tag{A.21}
$$

for sufficiently large n greater than $\left(\frac{4096\mu t\theta(K+\delta)}{\varepsilon^2\eta}\right)^2$.

□

A.5 Tightness

Next we argue that the scaled processes are tight. This result is a key step in proving Proposition A.14. which in turn is used to extract the restorative drift of the limiting diffusion process.

Proposition A.13. *Tightness of the scaled processes. The processes* $\{\tilde{Q}^n, n \geq 1\}$ *and* $\{\tilde{W}^n, n \geq 1\}$ *are tight.*

Proof. By exploiting Theorem 13.2 of [3], tightness follows from the asymptotic boundedness of the queue length and workload processes in Lemma A.7 and the fact that for any $\varepsilon > 0$, we have that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{Q}^n(v) - \tilde{Q}^n(u) \right| > \varepsilon \right) = 0 \tag{A.22}$$

and

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{W}^n(v) - \tilde{W}^n(u) \right| > \varepsilon \right) = 0. \tag{A.23}$$

We will show that (A.23) holds. Then (A.22) follows from the fact that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{Q}^n(v) - \tilde{Q}^n(u) \right| > \varepsilon \right) \tag{A.24}$$

$$\leq \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{Q}^n(v) - \tilde{Q}^n(u) + \tilde{W}^n(v) - \tilde{W}^n(v) + \tilde{W}^n(u) - \tilde{W}^n(u) \right| > \varepsilon \right) \tag{A.25}$$

$$\leq \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{Q}^n(v) - \tilde{W}^n(v) \right| + \left| \tilde{W}^n(u) - \tilde{Q}^n(u) \right| + \left| \tilde{W}^n(v) - \tilde{W}^n(u) \right| > \varepsilon \right) \tag{A.26}$$

$$\leq \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{Q}^n(v) - \tilde{W}^n(v) \right| > \varepsilon/3 \right) \tag{A.27}$$

$$+ \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{W}^n(u) - \tilde{Q}^n(u) \right| > \varepsilon/3 \right) \tag{A.28}$$

$$+ \limsup_{n \to \infty} \mathbb{P}\left(\sup_{u,v \in [0,t], v-u<\delta} \left| \tilde{W}^n(v) - \tilde{W}^n(u) \right| > \varepsilon/3 \right). \tag{A.29}$$

Thus, by Theorem A.9 and (A.23) we have the tightness of the queue length process. Now it remains to prove the tightness of the workload process and then our proof is complete. Since the workload only changes when new arrivals occur and decreases at rate one due to service, the remainder of proof can be shown by a similar argument of tightness in [17].

□

A.6 Convergence of the State Dependent Drift

The process $\tilde{\delta}_q^n$ swaps the sum of the abandonment and balking distributions, evaluated at the scaled queue length at the times of arrivals, with a smooth function involving the derivatives of the distribution functions evaluated at zero and then multiplied by the scaled queue lengths. The following proposition justifies this step and is proven in the appendix.

Proposition A.14. *Under the assumptions of Theorem 4.1,*

$$\tilde{\delta}_q^n \to 0$$

in probability as $n \to \infty$.

Proof. Proposition A.14 We can break δ_q^n into four seperate parts that will serve to streamline the proof. For every $t \geq 0$

$$\begin{aligned}
\delta_q^n(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} G_b(Q^n(t_i^n-)/\mu^n) - G_b(Q^n(t_i^n-)/(\mu n)) + G_r(Q^n(t_i^n-)/\mu^n) - G_r(Q^n(t_i^n-)/(\mu n)) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} G(Q^n(t_i^n-)/\mu^n, Q^n(t_i^n-)/\mu^n) - G(Q^n(t_i^n-)/\mu n, Q^n(t_i^n-)/\mu n) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \left(G_b(Q^n(t_i^n-)/(\mu n)) + G_r(Q^n(t_i^n-)/(\mu n)) - G(Q^n(t_i^n-)/\mu n, Q^n(t_i^n-)/\mu n) - \gamma \frac{Q^n(t_i^n-)}{\mu n} \right) \\
&\quad + \gamma \left(\int_0^t \tilde{Q}^n(s-) d\left(\frac{\bar{A}^n(s)}{\mu} \right) - \int_0^t \tilde{Q}^n(s) ds \right).
\end{aligned}$$

In order to prove our result, we need to show that each of these separate parts converges to zero in probability as $n \to \infty$.

Fix $t > 0$ and select arbitrary constants $\varepsilon, \eta > 0$. We first show that both

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} |G_b(Q^n(t_i^n-)/(\mu^n)) - G_b(Q^n(t_i^n-)/(\mu n))| > \frac{\varepsilon}{4} \right) < \frac{\eta}{8}, \tag{A.30}$$

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} |G_r(Q^n(t_i^n-)/(\mu^n)) - G_r(Q^n(t_i^n-)/(\mu n))| > \frac{\varepsilon}{4} \right) < \frac{\eta}{8}, \tag{A.31}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} |G(Q^n(t_i^n-)/(\mu^n), Q^n(t_i^n-)/(\mu^n)) - G(Q^n(t_i^n-)/(\mu n), Q^n(t_i^n-)/(\mu n))| > \frac{\varepsilon}{4} \right) < \frac{\eta}{4}, \tag{A.32}$$

We will prove (A.30) and the proof of (A.31) follows trivially. The right hand side of (A.30) can be expanded:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} |G_b(Q^n(t_i^n-)/(\mu^n)) - G_b(Q^n(t_i^n-)/(\mu n))| > \frac{\varepsilon}{4} \right) \qquad \text{(A.33)} \\
&\leq \mathbb{P}\left(2\mu n t \cdot \left(\sup_{s\in[0,t]} |G_b(Q^n(s)/\mu^n) + G_b(Q^n(s)/(\mu n))| \right) > \frac{\varepsilon\sqrt{n}}{4} \right) + \mathbb{P}\left(A^n(t) > 2\mu n t\right) \\
&\leq \mathbb{P}\left(\sup_{s\in[0,t]} Q^n(s) > K\sqrt{n} \right) + \mathbb{P}\left(A^n(t) > 2\mu n t\right) + \mathbb{P}\left(\sup_{x\leq K} \left| G_b(x\sqrt{n}/(\mu n)) + G_b(x\sqrt{n}/(\mu n)) \right| > \frac{\varepsilon}{8\mu t\sqrt{n}} \right).
\end{aligned}$$

Fix $K > 0$ so that by (A.2) of Lemma A.7,

$$\mathbb{P}\left(\sup_{s\in[0,t]} Q^n(s) > K\sqrt{n} \right) < \frac{\eta}{16}. \tag{A.34}$$

Moreover, we have that for any $\delta > 0$

$$\sup_{x\leq K} \left| \frac{x\sqrt{n}}{\mu^n} - \frac{x\sqrt{n}}{\mu n} \right| \leq \left| \frac{K\sqrt{n}}{\mu^n} - \frac{K\sqrt{n}}{\mu n} \right| < \frac{\delta}{\sqrt{n}}$$

for sufficiently large n. Set $\delta = \varepsilon/(16\mu t\gamma)$. The first result, (A.30), follows from (A.33), (A.34), and part (iii) of Lemma A.2 and A.6. We can also show that (A.31) follows from a similar argument and (A.32) also

follows from a similar agrument and the following Fréchet bound for the joint distribution of the balking and reneging random variables.

$$G(x,y) \leq \min\left(G_b(x), G_r(y)\right). \tag{A.35}$$

Next we show that

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{A^n(t)}\left|G_b(Q^n(t_i^n-)/(\mu n)) + G_r(Q^n(t_i^n-)/(\mu n)) - G(Q^n(t_i^n-)/(\mu n), Q^n(t_i^n-)/(\mu n)) - \gamma\frac{Q^n(t_i^n-)}{\mu n}\right| > \frac{\varepsilon}{4}\right) < \frac{\eta}{4} \tag{A.36}$$

as $n \to \infty$. We explore the derivative of the abandonment and balking distributions:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{A^n(t)}\left|G_b(Q^n(t_i^n-)/(\mu n)) + G_r(Q^n(t_i^n-)/(\mu n)) - \gamma\frac{Q^n(t_i^n-)}{\mu n}\right| > \frac{\varepsilon}{4}\right) \qquad (A.37)\\
&\leq \mathbb{P}\left(A^n(t)\sup_{s\in[0,t]}\left|G_b(Q^n(s)/(\mu n)) + G_r(Q^n(s)/(\mu n)) - \gamma\frac{Q^n(s)}{\mu n}\right| > \frac{\varepsilon\sqrt{n}}{4}\right)\\
&\leq \mathbb{P}\left(\sup_{s\in[0,t]} Q^n(s) > K\sqrt{n}\right) + \mathbb{P}\left(A^n(t) > 2\mu n t\right)\\
&\quad +1\left(\sup_{s\in[0,K/\mu]}\left|G_b(s/\sqrt{n}) + G_r(s/\sqrt{n}) - G(s/\sqrt{n}, s/\sqrt{n}) - \frac{\gamma s}{\sqrt{n}}\right| > \frac{\varepsilon}{8\mu t\sqrt{n}}\right)
\end{aligned}$$

Recall that the derivatives at zero of both G_b and G_r exist and sum to γ. Hence, for a any given $\delta > 0$ there exists an h_0 such that

$$\sup_{s\leq h}\left|\frac{G_b(s)}{s} + \frac{G_r(s)}{s} - \frac{G(s,s)}{s} - \gamma\right| < \delta$$

for all $h \leq h_0$. Let $\delta = \frac{\varepsilon}{8Kt}$ and let h_0 be a constant so that the above inequality holds. Now choose an n_0 such that $K/(\mu\sqrt{n}) \leq h_0$ for all $n \geq n_0$. It follows that for each $n \geq n_0$,

$$\begin{aligned}
&\sup_{s\in[0,K/(\mu\sqrt{n})]}\left|G_b(s) + G_r(s) - G(s,s) - \gamma s\right| \qquad (A.38)\\
&= \sup_{s\in[0,K/\mu]}\left|G_b(s/\sqrt{n}) + G_r(s/\sqrt{n}) - G(s/\sqrt{n}, s/\sqrt{n}) - \frac{\gamma s}{\sqrt{n}}\right|\\
&< \delta\frac{K}{\mu\sqrt{n}} \qquad (A.39)\\
&= \frac{\varepsilon}{8\mu t\sqrt{n}}.
\end{aligned}$$

The result (A.36) follows from (A.34), (A.37), (A.38) and part (iii) of Lemma A.2. Third, we show that

$$\limsup_{n\to\infty}\mathbb{P}\left(\left|\int_0^t \tilde{Q}^n(s-)d\left(\frac{\bar{A}^n(s)}{\mu}\right) - \int_0^t \tilde{Q}^n(s)ds\right| > \varepsilon\right) < \eta. \tag{A.40}$$

By Proposition A.13, the process $\{\tilde{Q}^n(s), s \leq t\}$ is tight. Consider a subsequence $\{n'\}$ over which the process $\tilde{Q}^{n'}$ has a limit, say $\tilde{Q}$. By the Skorohod Representation Theorem, there exists an alternative probability space on which are defined a sequence $\{(\hat{\tilde{Q}}^{n'}, \hat{\bar{A}}^{n'}), n \geq 1\}$ and, by part (ii) of Lemma A.2, a limit process $(\hat{\tilde{Q}}, \hat{\bar{A}})$ such that

$$(\hat{\tilde{Q}}^{n'}, \hat{\bar{A}}^{n'}) \stackrel{D}{=} (\tilde{Q}^{n'}, \bar{A}^{n'})$$

for each n' and such that $(\hat{\tilde{Q}}^{n'}, \hat{\bar{A}}^{n'}) \to (\hat{\tilde{Q}}, \hat{\bar{A}})$ almost surely as $n' \to \infty$. It is also true that

$$\int_0^u \hat{\tilde{Q}}^{n'}(s-)d\left(\frac{\hat{\bar{A}}^{n'}(s)}{\mu}\right) \stackrel{D}{=} \int_0^u \tilde{Q}^{n'}(s-)d\left(\frac{\bar{A}^{n'}(s)}{\mu}\right),$$

and

$$\int_0^t \hat{\tilde{Q}}^{n'}(s)ds \stackrel{D}{=} \int_0^t \tilde{Q}^{n'}(s)ds$$

for each n'. Applying Lemma 8.3 from [10] twice we have

$$\sup_{u\in[0,t]} \left| \int_0^u \hat{\tilde{Q}}^{n'}(s-)d\left(\frac{\hat{\tilde{A}}^{n'}(s)}{\mu}\right) - \int_0^u \hat{\tilde{Q}}(s)ds \right| \to 0$$

and

$$\sup_{u\in[0,t]} \left| \int_0^t \hat{\tilde{Q}}^{n'}(s)ds - \int_0^t \hat{\tilde{Q}}(s)ds \right| \to 0$$

almost surely as $n' \to \infty$, so that

$$\sup_{u\in[0,t]} \left| \int_0^u \hat{\tilde{Q}}^{n'}(s-)d\left(\frac{\hat{\tilde{A}}^{n'}(s)}{\mu}\right) - \int_0^t \hat{\tilde{Q}}^{n'}(s)ds \right| \to 0$$

almost surely as $n' \to \infty$. It follows that in our original probability space

$$\sup_{u\in[0,t]} \left| \int_0^u \tilde{Q}^{n'}(s-)d\left(\frac{\hat{\tilde{A}}^{n'}(s)}{\mu}\right) - \int_0^t \tilde{Q}^{n'}(s)ds \right| \to 0$$

as $n' \to \infty$. This limit holds on the arbitrarily chosen subsequence $\{n'\}$. Hence (A.40) holds.

Finally, our result follows from (A.30), (A.30), (A.31), (A.36), (A.40), and part (ii) of Lemma A.2. □

References

[1] I. J. Adan and V. G. Kulkarni. Single-server queue with markov-dependent inter-arrival and service times. *Queueing Systems*, 45(2):113–134, 2003.

[2] N. Balakrishnan and C.-D. Lai. *Continuous bivariate distributions*. Springer Science & Business Media, 2009.

[3] P. Billingsley. *Convergence of probability measures*, volume 493. John Wiley & Sons, 1999.

[4] S. C. Borst, O. J. Boxma, and M. Combé. Collection of customers: a correlated m/g/1 queue. In *ACM SIGMETRICS Performance Evaluation Review*, volume 20, pages 47–59. ACM, 1992.

[5] M. C. Cario and B. L. Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Citeseer, 1997.

[6] X. Chao. Monotone effect of dependency between interarrival and service times in a simple queueing system. *Operations research letters*, 17(1):47–51, 1995.

[7] I. Cidon, R. Guérin, A. Khamisy, and M. Sidi. Analysis of a correlated queue in a communication system. *Information Theory, IEEE Transactions on*, 39(2):456–465, 1993.

[8] B. Conolly and Q. Choo. The waiting time process for a generalized correlated queue with exponential demand and service. *SIAM Journal on Applied Mathematics*, 37(2):263–275, 1979.

[9] B. Conolly and N. Hadidi. A correlated queue. *Journal of Applied Probability*, pages 122–136, 1969.

[10] J. Dai and W. Dai. A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Systems*, 32(1-3):5–40, 1999.

[11] P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 8(1):329–384, 2003.

[12] K. W. Fendick, V. R. Saksena, and W. Whitt. Dependence in packet queues. *Communications, IEEE Transactions on*, 37(11):1173–1183, 1989.

[13] K. W. Fendick, V. R. Saksena, and W. Whitt. Investigating dependence in packet queues with the index of dispersion for work. *Communications, IEEE Transactions on*, 39(8):1231–1244, 1991.

[14] N. Hadidi. Queues with partial correlation. *SIAM Journal on Applied Mathematics*, 40(3):467–475, 1981.

[15] N. Hadidi. Further results on queues with partial correlation. *Operations research*, 33(1):203–209, 1985.

[16] S. K. Iyer and D. Manjunath. Queues with dependency between interarrival and service times using mixtures of bivariates. *Stochastic models*, 22(1):3–20, 2006.

[17] O. B. Jennings and J. Pender. Comparisons of standard and ticket queues in heavy traffic. *Submitted for publication to Queueing Systems*, 2015.

[18] O. B. Jennings and J. E. Reed. An overloaded multiclass fifo queue with abandonments. *Operations research*, 60(5):1282–1295, 2012.

[19] C. Langaris. Busy-period analysis of a correlated queue with exponential demand and service. *Journal of applied probability*, pages 476–485, 1987.

[20] X. S. Y. D. Z. H. Li, X. Optimal staffing of ticket queues. *Working Paper*, 2014.

[21] A. W. Marshall and I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 62(317):30–44, 1967.

[22] A. Müller. On the waiting times in queues with dependency between interarrival and service times. *Operations Research Letters*, 26(1):43–47, 2000.

[23] A. Müller. On the interplay between variability and negative dependence for bivariate distributions. *Operations Research Letters*, 31(2):90–94, 2003.

[24] G. Pang and W. Whitt. The impact of dependent service times on large-scale service systems. *Manufacturing & Service Operations Management*, 14(2):262–278, 2012.

[25] G. Pang and W. Whitt. Infinite-server queues with batch arrivals and dependent service times. *Probability in the Engineering and Informational Sciences*, 26(02):197–220, 2012.

[26] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*, 73(2):119–146, 2013.

[27] J. Pender. The Truncated Normal Distribution: Applications to Queues with Impatient Customers. *Operations Research Letters*, 43(1):40–45, 2015.

[28] J. Reed and A. R. Ward. Approximating the gi/gi/1+ gi queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33(3):606–644, 2008.

[29] A. R. Ward and P. W. Glynn. A diffusion approximation for a markovian queue with reneging. *Queueing Systems*, 43(1-2):103–128, 2003.

[30] A. R. Ward and P. W. Glynn. A diffusion approximation for a gi/gi/1 queue with balking or reneging. *Queueing Systems*, 50(4):371–400, 2005.

[31] S. H. Xu, L. Gao, and J. Ou. Service performance analysis and improvement for a ticket queue with balking customers. *Management science*, 53(6):971–990, 2007.