# Critically Loaded Time-Varying Multi-Server Queues: Computational Challenges and Approximations

Young Myoung Ko
Sabre Holdings, 3150 Sabre Drive, Southlake, Texas 76092, USA, YoungMyoung.Ko@sabre.com
Natarajan Gautam
Department of Industrial and Systems Engineering, Texas A&M University, 3131 TAMU, College Station, Texas 77843-3131, USA, gautam@tamu.edu

In this paper, we consider time-varying multi-server queues with abandonment and retrials. For their performance analysis, fluid and diffusion limits utilizing strong approximations have been widely used in the literature. Although those limits are asymptotically exact, they may not accurately approximate performance of multi-server queues even if the number of servers is large. To address that concern, this paper focuses on developing a methodology by taking fluid and diffusion limits in a non-traditional fashion. We show that our approximation is significantly more accurate and also asymptotically true. We illustrate the effectiveness of our methodology by performing several numerical experiments.

*Key words:* transient analysis; multi-server queues; nearly critically loaded; uniform acceleration; strong approximations

---

## 1. Introduction

In this paper, we are interested in developing accurate approximations for time-varying many-server queues with abandonment and retrials as depicted in Figure 1. Inspired by call centers, there have been extensive studies on multi-server queues, especially having a large number of servers. Most of the recent studies utilize asymptotic analysis as it makes the problem tractable and also provides good approximations under certain conditions. Asymptotic analysis, typically, utilizes convergence theorems to derive fluid and diffusion limits, that are well summarized in Billingsley (1999) and Whitt (2002). Methodologies to obtain fluid and diffusion limits, as described in Halfin and Whitt (1981), have been developed in the literature using two different ways in terms of the traffic intensity.

The first approach is to consider the convergence of a sequence of traffic intensities to a certain value. Depending on the value to which the sequence converges, there are three different operational regimes: efficiency driven (ED), quality and efficiency driven

(QED), and quality driven (QD). Roughly speaking, if the traffic intensity ($\rho$) of the limit process is strictly greater than 1, it is called ED regime. If $\rho = 1$, then that is QED, otherwise QD. Many research studies have been done under the ED and QED regimes for multi-server queues like call centers (Halfin and Whitt (1981), Puhalskii and Reiman (2000), Garnet et al. (2002), Whitt (2006a), Whitt (2006b), Pang and Whitt (2009)). Recently, the QED regime, also known as "Halfin-Whitt regime", has received a lot of attention; this is because it actually achieves both high utilization of servers and quality of service (Zeltyn and Mandelbaum (2005)), and is a favorable operational regime for call centers with strict performance constraints (Mandelbaum and Zeltyn (2009)).

The second way to obtain limit processes is to accelerate parameters keeping the traffic intensity fixed. An effective methodology called "uniform acceleration" coupled with the theory of strong approximations enables the analysis of time-dependent queues (Kurtz (1978), Mandelbaum and Pats (1995), Mandelbaum and Pats (1998), Mandelbaum et al. (1998), Hampshire et al. (2009)) and in fact provides the basis of this paper. The advantage of accelerating parameters as described in Kurtz (1978) is that it can be applied to a wide class of stochastic processes and can be nicely extended to time-dependent systems by combining with the results in Mandelbaum et al. (1998). However, it might not be applied to multi-server queues directly since the rate functions (e.g. net arrival rates and service rates) considered in Kurtz (1978) are assumed to be differentiable everywhere. But some rate functions in multi-server queues are not differentiable everywhere since they are of the forms, $\min(\cdot, \cdot)$ or $\max(\cdot, \cdot)$. To extend the theory to non-smooth rate functions, Mandelbaum et al. (1998) proves convergence to the limit processes by introducing a new derivative called *scalable Lipschitz derivative* and provides models for several queueing systems such as Jackson networks, multi-server queues with abandonment and retrials, multi-class preemptive priority queues, etc. In addition, several sets of ordinary differential equations are also provided to obtain the mean value and covariance matrix of the limit processes. It, however, turns out that the resulting sets of differential equations are computationally infeasible in general.

In a follow-on paper, Mandelbaum et al. (2002) provides numerical results for queue lengths and waiting times in multi-server queues with abandonment and retrials by adding a constraint to deal with computational intractability. Specifically, the authors restrict their attention to the cases where the time periods when the fluid limit is the same as the number of servers have measure zero. Note that when the fluid limit has the same value as the num-

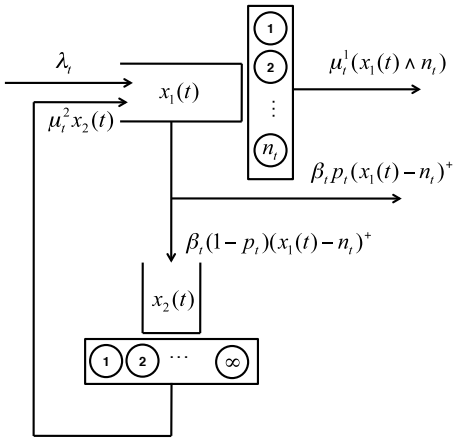
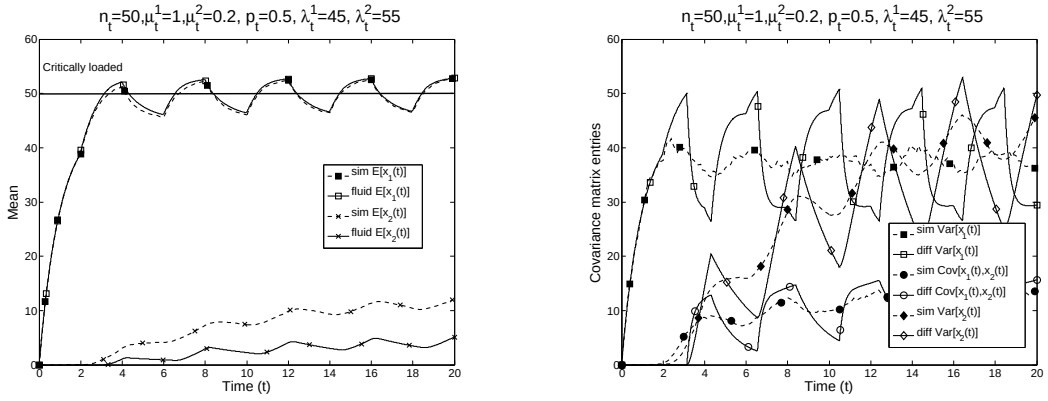



Figure 1: Multi-server queue with abandonment and retrials, Mandelbaum et al. (1998)

ber of servers, we say the queue is in the critically loaded phase. By doing that, they were able to apply the diffusion limit in Kurtz (1978) to the multi-server queues since in this case *scalable Lipschitz derivatives are essentially the same as the ordinary derivatives ignoring a set of non-differentiable points*. Adding this constraint seems restrictive in theory. However, in practice, it seems reasonable. For example, the number of servers is usually piecewise constant, and the fluid limit is a continuous function of time including non-linear terms. Therefore, the fluid limit possibly stays close to the number of servers but is unlikely to stay there for a positive-measured amount of time. Nevertheless, as pointed out in Mandelbaum et al. (2002), if the queues stay close to the critically loaded phase (called *lingering* in Mandelbaum et al. (2002)), their approach actually causes significantly inaccurate results despite the fact that it is asymptotically true. In this paper, we consider multi-server queues with the same restriction as that in Mandelbaum et al. (2002), yet we are specifically interested in a multi-server queue when the fluid limit stays close to the number of servers but crosses it at countable number of time points having measure zero; we would use the term *nearly critically loaded (phase)* to indicate the state of this kind of queues.

To explain this inaccuracy in detail, consider a multi-server queue with abandonment and retrials as shown in Figure 1 (the setting is identical to that in Mandelbaum et al. (2002)). As an example we select for all $t \geq 0$, numerical values of the number of servers ($n_t = 50$), service rate of each server ($\mu_t^1 = 1$), and sojourn rate in the retrial queue ($\mu_t^2 = 0.2$). We choose alternating arrival rates between $\lambda_t^1 = 45$ and $\lambda_t^2 = 55$ every two units of time (the parameters are defined in Section 2 and illustrated in Figure 1). In this numerical example, the fluid limit crosses the number of servers at finite number of time points on any

compact time interval, which implies the fluid and diffusion limits using Lipschitz derivatives are basically the same as those using Kurtz's method. Here, we consider two dimensional state space ($x_1(t)$ and $x_2(t)$ for the number of customers in the multi-server queue and the retrial queue respectively). We graph the estimated values of $E[x_1(t)]$ and $E[x_2(t)]$ in Figure 2 (a), and also $Var[x_1(t)]$, $Var[x_2(t)]$, and $Cov[x_1(t), x_2(t)]$ in Figure 2 (b). We use 5,000 independent simulation runs to benchmark those quantities since obtaining their exact values is not possible. Notice that the estimation of $E[x_1(t)]$ is reasonably accurate, while the others ($E[x_2(t)]$, $Var[x_1(t)]$, $Var[x_2(t)]$, and $Cov[x_1(t), x_2(t)]$) are not accurate at all. The reason for that is the system lingers around the non-differentiable points. We explain this in detail in Section 3.2. However, this does nourish the need for a methodology to accurately predict the system performance, and *Figure 3 illustrates how improved the estimation accuracy is with the methodology proposed in this paper.*

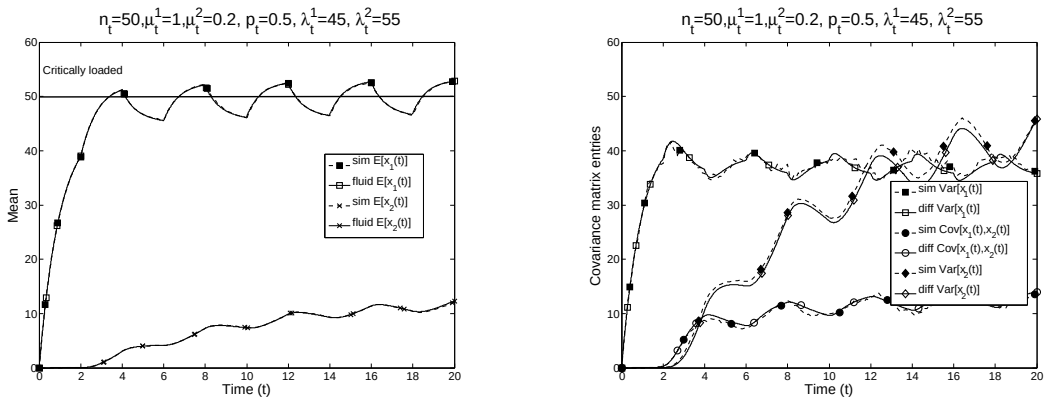



(a) Simulation vs Fluid limit

(b) Simulation vs Diffusion limit

Figure 2: Simulation vs Fluid and diffusion limits

Having motivated the need to develop a methodology for the critically loaded phase, we now describe its importance. According to Mandelbaum and Pats (1998) and Mandelbaum et al. (2002), time-dependent queues make transitions among three phases: underloaded, critically loaded, and overloaded. The phase of the system is determined by its fluid limit. The limit process in strong approximations does not require any regimes such as QD, QED, or ED. However, from Section 1.4 in Zeltyn and Mandelbaum (2005), we could find a rough correspondence between the operational regimes (QD, QED, and ED) and the phases in time-varying queues (underloaded, critically loaded, and overloaded). Explaining it briefly, Zeltyn and Mandelbaum (2005) models operational regimes from tracing data of real call




(a) Simulation vs Adjusted fluid limit (b) Simulation vs Adjusted diffusion limit

Figure 3: Simulation vs Adjusted fluid and diffusion limits

centers. They say the call center is working in the ED regime when the occupancy is 100% with higher abandonment rates. Similarly, they associate the time slots with the other operational regimes such as QED and QD. The tracing table used for the explanation of the operational regimes, indeed, represents the dynamics of the time-varying multi-server queues. Therefore, the ED, QED and QD regimes could correspond to the overloaded, critically loaded and underloaded phases respectively. From the tracing data in Zeltyn and Mandelbaum (2005), we, again, recognize the importance of the critically loaded phase as nearly 100% utilization and low abandonment rates which most of companies want are achieved in that phase. Therefore, capturing the dynamics of multi-server queues near the critically loaded phase is also of significant importance. Nonetheless, from Figure 2, we found two major issues in the existing approach: 1) the fluid limit (where the non-differentiability issue is actually irrelevant) is itself inaccurate and 2) huge estimation inaccuracy against the simulation result is observed in the diffusion limit.

In this paper, we approach the above two issues from a different point of view and provide an effective solution to them. Considering those, the contributions of this paper can be summarized as follows:

1. We derive adjusted fluid and diffusion limits to address the above issues and show that they are asymptotically true.

2. We provide a reasonable approximation methodology to obtain the adjusted limits and verify their effectiveness by several numerical experiments.

Note that we sometimes use the term, *standard*, to indicate the fluid and diffusion limits in Kurtz (1978) and Mandelbaum et al. (1998) to distinguish them from the *adjusted* fluid and diffusion limits proposed in this paper. We now describe the organization of this paper. In Section 2, we state the problem considered in this paper. In Section 3, we summarize the fluid and diffusion limits in Kurtz (1978) and Mandelbaum et al. (1998), and describe the above issues in detail. In Section 4, we derive the adjusted fluid and diffusion limits and describe the relationship to the standard fluid and diffusion limits. In Section 5, we explain a Gaussian-based approximation to achieve computational tractability. Further intuition on the adjusted limits is provided in Section 6 to understand how the adjusted limits contribute to the estimation accuracy. In Section 7, we provide a number of numerical examples and compare against the standard approach as well as simulation. Finally, in Section 8, we make concluding remarks and explain directions for future work.

# 2. Problem description

Consider Figure 1 that illustrates a multi-server queue with abandonment and retrials as described in Mandelbaum et al. (1998) and Mandelbaum et al. (2002). There are $n_t$ number of servers in the service node at time $t$. Customers arrive to the service node according to a non-homogeneous Poisson process at rate $\lambda_t$. The service time of each customer follows a distribution having a memoryless property at rate $\mu_t^1$. Customers in the queue are served under the FCFS policy and the abandonment rate of customers is $\beta_t$ with exponentially distributed time to abandon. Abandoning customers leave the system with probability $p_t$ or go to a retrial queue with probability $1 - p_t$. The retrial queue is equivalent to an infinite-server-queue and hence each customer in the retrial queue waits there for a random amount of time with mean $1/\mu_t^2$ and returns to the service node.

Let $X(t) = \big(x_1(t), x_2(t)\big)$ be the system state where $x_1(t)$ is the number of customers in the service node and $x_2(t)$ is the number of customers in the retrial queue at time $t$. Then, $X(t)$ is the unique solution to the following integral equations:

$$\begin{aligned} x_1(t) &= x_1(0) + Y_1\Big(\int_0^t \lambda_s ds\Big) + Y_2\Big(\int_0^t x_2(s)\mu_s^2 ds\Big) - Y_3\Big(\int_0^t \big(x_1(s) \wedge n_s\big)\mu_s^1 ds\Big) \\ &\quad - Y_4\Big(\int_0^t \big(x_1(s) - n_s\big)^+ \beta_s(1-p_s) ds\Big) - Y_5\Big(\int_0^t \big(x_1(s) - n_s\big)^+ \beta_s p_s ds\Big), \qquad (1) \\ x_2(t) &= x_2(0) + Y_4\Big(\int_0^t \big(x_1(s) - n_s\big)^+ \beta_s(1-p_s) ds\Big) - Y_2\Big(\int_0^t x_2(s)\mu_s^2 ds\Big), \qquad (2) \end{aligned}$$

where $Y_i$'s are independent rate-1 Poisson processes.

The performance measures of interest are $E[X(t)]$ and $Cov[X(t), X(t)]$ (i.e. $Var[x_1(t)]$, $Var[x_2(t)]$, and $Cov[x_1(t), x_2(t)]$) for any given time $t \in [0, T]$, where $T < \infty$ is a constant. Especially, we have an interest in a multi-server queue that is *nearly critically loaded*, i.e., the system whose fluid limit stay close to, yet equals for a measure zero amount of time to the number of servers. Anyhow, as one may notice, the above two equations (1) and (2) cannot be solved directly. We would try to take advantage of an asymptotic methodology that is adequate for the analysis of time-varying systems with large number of servers. Nevertheless, as briefly mentioned in Section 1, we found that the existing methodologies are significantly inaccurate in the nearly critically loaded phase.

The objective of this paper is to develop a new approach to enhance the accuracy in estimating the mean value and covariance matrix for the multi-server queues with abandonment and retrials. To do so, we start by summarizing the fluid and diffusion limits from strong approximations and addressing the potential limitations in the following section.

# 3. Fluid and diffusion limits from strong approximations

In Section 3.1, we recapitulate the fluid and diffusion limits in Kurtz (1978) and Mandelbaum et al. (1998). In Section 3.2, we explain what produces estimation errors that we describe earlier in Section 1 and why existing methodologies do not fix them.

## 3.1. Fluid and diffusion limits

In this section, we obtain the fluid and diffusion limits by increasing the arrival rate and the number of servers according to the uniform acceleration technique. We consider the cases where the queues stay critically loaded only for the time periods having measure zero just like in Mandelbaum et al. (2002). In this case, accelerating the number of servers is basically the same as accelerating the service rate described in Kurtz (1978) if we adjust the definition of the system state suitably. Moreover, it is worthwhile to note that for $\eta \in \mathbf{N}$, the state of the queueing system $X^\eta(t)$ includes jumps but the limit process is continuous. Therefore, the weak convergence result that is presented is with respect to uniform topology in Space $D$ (Billingsley (1999) and Whitt (2002)).

Define $X^\eta(t)$ by accelerating the arrival rate and the number of servers by the factor of $\eta$ as follows:

$$\begin{aligned}
x_1^\eta(t) &= x_1^\eta(0) + Y_1\Big(\int_0^t \eta\lambda_s ds\Big) + Y_2\Big(\int_0^t x_2^\eta(s)\mu_s^2 ds\Big) - Y_3\Big(\int_0^t \big(x_1^\eta(s) \wedge \eta n_s\big)\mu_s^1 ds\Big) \\
&\quad -Y_4\Big(\int_0^t \big(x_1^\eta(s) - \eta n_s\big)^+ \beta_s(1-p_s) ds\Big) - Y_5\Big(\int_0^t \big(x_1^\eta(s) - \eta n_s\big)^+ \beta_s p_s ds\Big), \\
&= x_1^\eta(0) + Y_1\Big(\int_0^t \eta\lambda_s ds\Big) + Y_2\Big(\int_0^t \eta\Big(\frac{x_2^\eta(s)}{\eta}\mu_s^2\Big) ds\Big) - Y_3\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta} \wedge n_s\Big)\mu_s^1 ds\Big) \\
&\quad -Y_4\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta} - n_s\Big)^+ \beta_s(1-p_s) ds\Big) - Y_5\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta} - n_s\Big)^+ \beta_s p_s ds\Big), \quad (3) \\
x_2^\eta(t) &= x_2^\eta(0) + Y_4\Big(\int_0^t \big(x_1^\eta(s) - \eta n_s\big)^+ \beta_s(1-p_s) ds\Big) - Y_2\Big(\int_0^t x_2^\eta(s)\mu_s^2 ds\Big), \\
&= x_2^\eta(0) + Y_4\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta} - n_s\Big)^+ \beta_s(1-p_s) ds\Big) - Y_2\Big(\int_0^t \eta\Big(\frac{x_2^\eta(s)}{\eta}\mu_s^2\Big) ds\Big). \quad (4)
\end{aligned}$$

Note that equations (3) and (4) are essentially identical to equations (2.7) and (2.8) in Mandelbaum et al. (2002).

Next we move away from the specific $X^\eta(t)$ described above for the system in Figure 1 to a more generic $X^\eta(t)$ as described in Kurtz (1978) and Mandelbaum et al. (1998) to maintain the richness of the results. Let $X^\eta(t)$ be an arbitrary $d$-dimensional stochastic process which is the solution to the following integral equation:

$$X^\eta(t) = x_0^\eta + \sum_{i=1}^k l_i Y_i\Big(\int_0^t \eta f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big) ds\Big), \quad (5)$$

where $x_0^\eta = X^\eta(0)$ is a constant $d$-dimensional vector, $Y_i$'s are independent rate-1 Poisson processes, $l_i \in \mathbf{Z}^d$ for $i \in \{1, 2, \ldots, k\}$ are constant, and $f_i$'s are continuous functions such that $|f_i(t,x)| \le C_i(1+|x|)$ for some $C_i < \infty$, $t \le T$ and $T < \infty$.
Notice that equations (3) and (4) are a special case of equation (5) by defining $f_i(\cdot,\cdot)$'s as follows: For $x = (x_1, x_2)$ and $t \in [0, T]$,

$$\begin{aligned}
&f_1(t,x) = \lambda_t, f_2(t,x) = \mu_t^2 x_2, f_3(t,x) = \mu_t^1(x_1 \wedge n_t), \\
&f_4(t,x) = \beta_t(1-p_t)(x_1 - n_t)^+, f_5(t,x) = \beta_t p_t (x_1 - n_t)^+, \\
&\qquad l_1 = \begin{pmatrix}1\\0\end{pmatrix}, l_2 = \begin{pmatrix}1\\-1\end{pmatrix}, l_3 = \begin{pmatrix}-1\\0\end{pmatrix}, l_4 = \begin{pmatrix}-1\\1\end{pmatrix}, \text{ and } l_5 = \begin{pmatrix}-1\\0\end{pmatrix}.
\end{aligned}$$

Having said that, we proceed with the generic definition of $X^\eta(t)$ in equation (5).

Typically the process $X^\eta(t)$ (usually called a scaled process) is obtained by considering $\eta$ times faster arrival rate and larger number of servers. This type of setting is used in the

literature and is denoted as "uniform acceleration" in Massey and Whitt (1998), Mandelbaum et al. (1998), and Mandelbaum et al. (2002). Then, the following theorem provides the fluid limit to which $\{X^\eta(t)\}_{\eta\geq 1}$ converges almost surely as $\eta \to \infty$. For that, we first define

$$F(t,x) = \sum_{i=1}^{k} l_i f_i(t,x). \tag{6}$$

**Theorem 1** (Fluid limit, Kurtz (1978), Mandelbaum et al. (1998))**.** *If there is a constant $M < \infty$ such that $|F(t,x) - F(t,y)| \leq M|x-y|$ for all $t \in [0,T]$ and $T < \infty$. Then, $\lim_{\eta\to\infty} \frac{X^\eta(t)}{\eta} = \bar{X}(t)$ a.s. where $\bar{X}(t)$ is the solution to the following integral equation:*

$$\bar{X}(t) = x_0 + \sum_{i=1}^{k} l_i \int_0^t f_i\big(s, \bar{X}(s)\big) ds.$$

Note that $\bar{X}(t)$ is a deterministic time-varying quantity. We will connect $\bar{X}(t)$ and $X(t)$ defined in equation (5) via equation (7), but before that we provide the following result. Once we have the fluid limit, we can obtain the diffusion limit from the scaled centered process $(D^\eta(t))$. Define $D^\eta(t)$ to be $\sqrt{\eta}\Big(\frac{X^\eta(t)}{\eta} - \bar{X}(t)\Big)$. Then, the limit process of $D^\eta(t)$ is provided by the following theorem.

**Theorem 2** (Diffusion limit, Kurtz (1978), Mandelbaum et al. (1998))**.** *Suppose $F$ is differentiable almost everywhere with respect to $x$ and the set $\{t : \partial F(t, \bar{X}(t))$ does not exist.$\}$ has measure zero. For some $M < \infty$ and $i \in \{1, \ldots, k\}$, if $f_i$'s and $F$ satisfy.*

$$|f_i(t,x) - f_i(t,y)| \leq M|x-y| \quad \text{and} \quad \left|\frac{\partial}{\partial x_i} F(t,x)\right| \leq M, \qquad \text{for almost all } t \in [0,T],$$

*then $\lim_{\eta\to\infty} D^\eta(t) = D(t)$ where $D(t)$ is the solution to*

$$D(t) = \sum_{i=1}^{k} l_i \int_0^t \sqrt{f_i\big(s, \bar{X}(s)\big)} dW_i(s) + \int_0^t \partial F\big(s, \bar{X}(s)\big) D(s) ds,$$

*$W_i(\cdot)$'s are independent standard Brownian motions, and $\partial F(t,x)$ is the gradient matrix of $F(t,x)$ with respect to $x$.*

**Remark 1.** *According to Ethier and Kurtz (1986), if $D(0)$ is a constant or a Gaussian random vector, then $D(t)$ is a Gaussian process.*

Now, we have the fluid and diffusion limits for $X^\eta(t)$. Therefore, for a large $\eta$, $X^\eta(t)$ is approximated by

$$X^\eta(t) \approx \eta \bar{X}(t) + \sqrt{\eta} D(t). \tag{7}$$

If we follow this approximation, we can also approximate the mean and covariance matrix of $X^\eta(t)$ denoted by $E\big[X^\eta(t)\big]$ and $Cov\big[X^\eta(t), X^\eta(t)\big]$ respectively as

$$E\big[X^\eta(t)\big] \approx \eta\bar{X}(t) + \sqrt{n}E\big[D(t)\big], \tag{8}$$

$$Cov\big[X^\eta(t), X^\eta(t)\big] \approx \eta Cov\big[D(t), D(t)\big]. \tag{9}$$

In equations (8) and (9), only $\bar{X}(t)$ is known. Therefore, in order to get approximated values of $E\big[X^\eta(t)\big]$ and $Cov\big[X^\eta(t), X^\eta(t)\big]$, we need to obtain $E\big[D(t)\big]$ and $Cov\big[D(t), D(t)\big]$. The following theorem provides a methodology to obtain $E\big[D(t)\big]$ and $Cov\big[D(t), D(t)\big]$.

**Theorem 3** (Mean and covariance matrix of linear stochastic systems, Arnold (1992))**.** *Let $Y(t)$ be the solution to the following linear stochastic differential equation.*

$$dY(t) = A(t)Y(t)dt + B(t)dW(t), \quad Y(0) = 0,$$

*where $A(t)$ is a $d \times d$ matrix, $B(t)$ is a $d \times k$ matrix, and $W(t)$ is a $k$-dimensional standard Brownian motion. Suppose $A(t)$ and $B(t)$ are measurable and bounded on a compact time interval $[0, T]$. Let $M(t) = E\big[Y(t)\big]$ and $\Sigma(t) = Cov\big[Y(t), Y(t)\big]$. Then, $M(t)$ and $\Sigma(t)$ can be obtained as the unique solution to the following ordinary differential equations:*

$$\begin{aligned} \frac{d}{dt}M(t) &= A(t)M(t) \\ \frac{d}{dt}\Sigma(t) &= A(t)\Sigma(t) + \Sigma(t)A(t)' + B(t)B(t)'. \end{aligned} \tag{10}$$

**Corollary 1.** *If $M(0) = 0$, then $M(t) = 0$ for $t \geq 0$.*

By Corollary 1, if $D(0) = 0$, then $E\big[D(t)\big] = 0$ for $t \geq 0$. Therefore, if $\bar{X}(0) = X(0) = x_0$, then we can rewrite approximate equation (8) to be

$$E\big[X^\eta(t)\big] \approx \eta\bar{X}(t).$$

**Remark 2.** *Mandelbaum et al. (2002) provides a set of differential equations (see equations (2.15)-(2.19)) to obtain the mean and covariance matrix of the diffusion limit. Those differential equations are indeed the ones in Theorem 3.*

Until now, we explain the fluid and diffusion limits in Kurtz (1978) and Mandelbaum et al. (1998). In theory, they are asymptotically true and seems promising when approximating multi-server queues with a large number of servers. Unfortunately, they do not always provide good approximation results even if the number of servers is very large. In the next section, we describe the possibility of inaccuracy of them in detail.

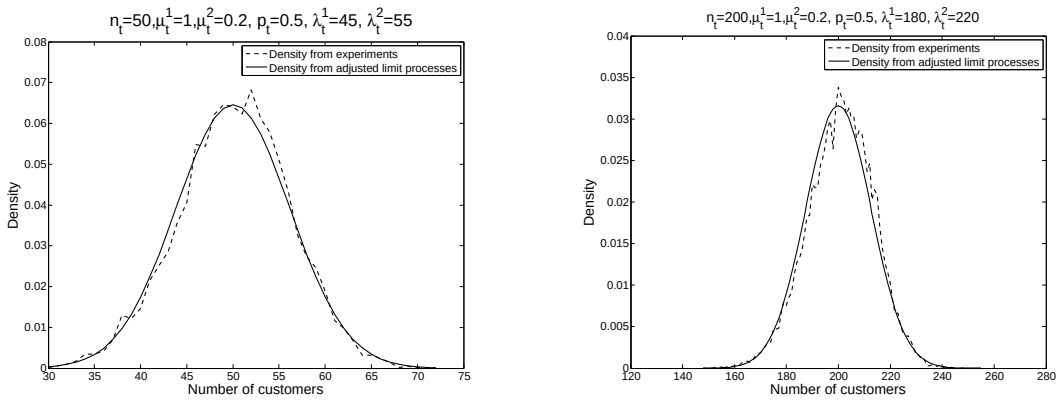



Figure 4: Empirical density vs Gaussian density

## 3.2. Inaccuracy of the fluid and diffusion limits as approximations

In this section, we would like to explain the possibility for both fluid and diffusion limits to be inaccurate. Consider the following equation to get the exact value of $E\big[X^\eta(t)\big]$ by the following theorem.

**Theorem 4** (Expected value of $X^\eta(t)$)**.** *Consider $X^\eta(t)$ defined in equation (5). Then, for $t \in [0,T]$, $E\big[X^\eta(t)\big]$ is the solution to the following equation.*

$$E\big[X^\eta(t)\big] = x_0^\eta + \sum_{i=1}^{k} l_i \int_0^t \eta E\bigg[f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big)\bigg] ds \tag{11}$$

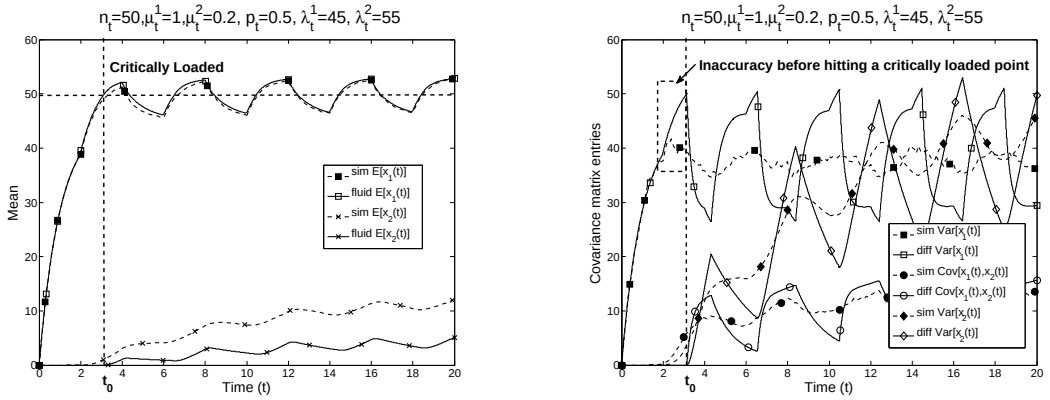
*Proof.* Take expectation on both sides of equation (5). Then,

$$\begin{aligned} E\big[X^\eta(t)\big] &= x_0^\eta + \sum_{i=1}^{k} l_i E\bigg[Y_i\Big(\int_0^t \eta f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big) ds\Big)\bigg] \\ &= x_0^\eta + \sum_{i=1}^{k} l_i E\bigg[\int_0^t \eta f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big) ds\bigg] \\ &= x_0^\eta + \sum_{i=1}^{k} l_i \int_0^t \eta E\bigg[f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big)\bigg] ds \text{ by Fubini theorem in Folland (1999).} \end{aligned}$$

□

Comparing Theorems 1 and 4, notice that we cannot conclude that $\eta\bar{X}(t)$ in Theorem 1 is sufficient to approximate $E\big[X^\eta(t)\big]$ since $E\Big[f_i\big(t, \frac{X^\eta(t)}{\eta}\big)\Big] \neq f_i\Big(t, E\big[\frac{X^\eta(t)}{\eta}\big]\Big)$. Yes, we know that they eventually become identical when $\eta$ goes to infinity. However, the problem lies in

the fact that we have no choice but to use the same $\bar{X}(t)$ (with scaling) as approximations no matter which $\eta$ values (number of servers) the real systems (e.g. call centers) have, i.e. the same $\bar{X}(t)$ is always used for $\eta = 50$, $\eta = 500$, and $\eta = 5,000$. Therefore, if one can derive a new fluid limit specifically designed for each fixed $\eta$ value, we expect more improved results, and that is the adjusted fluid limit that we will explain in Section 4.




(a) Simulation vs Fluid limit

(b) Simulation vs Diffusion limit

Figure 5: Simulation vs Fluid and diffusion limits

Figures 5 (a) and (b) show the estimation of the mean value and covariance matrix of the multi-server queues respectively. We use the annotated version of Figure 2 (via Figure 5) here for a clearer explanation. Since the number of servers is 50, as shown in Figure 5 (a), the mean value of $x_1(t)$ is fluctuating close to the number of servers. From the figure, we also confirm that the fluid limit is quite inaccurate for the estimation of the mean value of $x_2(t)$. For the covariance matrix, as shown in Figure 5 (b), the diffusion limit brings about immense estimation inaccuracy (sharp spikes) in the nearly critically loaded phase. Recall that under the parameters in Figure 5 *the fluid and diffusion limits using the scalable Lipschitz derivatives in Mandelbaum et al. (1998) are virtually the same as those in Kurtz (1978) and Mandelbaum et al. (2002).* Therefore, the methodology in Mandelbaum et al. (1998) also results in the sharp spikes. In fact, it turns out that the sharp spikes in the diffusion limits arise from the sudden changes in the drift matrix of the diffusion limits at the non-differentiable points of rate functions. We will revisit and explain it in Section 6.

In the next section, we describe our approach to the above issues in both fluid and diffusion limits. Instead of accelerating parameters, we keep $\eta$ fixed and construct a new sequence $\{Z^{\eta,\nu}(t)\}_{\nu \geq 1}$ which converges to $E[X^{\eta}(t)]$ almost surely. We derive fluid and diffusion limits

for the new sequence and show that they are asymptotically identical to the standard fluid and diffusion limits in Kurtz (1978) and Mandelbaum et al. (1998).

# 4. Adjusted fluid and diffusion limits

The basic idea of our approach is to derive new fluid and diffusion limits *for a fixed $\eta$ (fixed number of servers)*. In this approach, we want to approximate multi-server queues having *a finite number of servers*. To do so, for a fixed $\eta$, we first define new rate functions $g_i^\eta(t,x)$'s from the existing $f_i(t,x)$'s as follows:

$$g_i^\eta(t,x) \;=\; \eta E\Big[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big)\Big].$$

With new rate functions, we construct a new sequence of stochastic processes, $\{Z^{\eta,\nu}(t)\}_{\nu\geq 1}$ such that $Z^{\eta,\nu}(t)$ is the solution to the following integral equations:

$$Z^{\eta,\nu}(t) = \nu x_0^\eta + \sum_{i=1}^k l_i Y_i\Big(\int_0^t \nu g_i^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big). \tag{12}$$

Notice that once we show that $g_i^\eta(t,x)$'s are Lipschitz functions on $[0,T]$, we can apply Theorems 1 and 2, and are able to obtain new fluid and diffusion limits for $Z^{\eta,\nu}(t)$. From the following lemmas, we prove that the functions $g_i^{\eta,\nu}(t,\cdot)$'s are actually Lipschitz functions.

**Lemma 1.** *For a fixed $\eta$ and $i \in \{1,2,\ldots,k\}$, if $|f_i(t,x)| \leq C_i(1+|x|)$ on $[0,T]$, then $g_i^\eta(t,x)$'s satisfy*

$$|g_i^\eta(t,x)| \;\leq\; D_i(1+|x|) \quad \textit{for some } D_i < \infty.$$

*Proof.*

$$\begin{aligned}
|g_i^\eta(t,x)| \;&=\; \Big|\eta E\Big[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big)\Big]\Big| \\
&\leq\; \Big|\eta E\Big[C_i\Big(1 + \Big|\frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big|\Big)\Big| \\
&\leq\; \Big|\eta C_i\Big(1 + E\Big[\Big|\frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta}\Big|\Big] + \Big|\frac{x}{\eta}\Big|\Big)\Big| \\
&\leq\; D_i(1+|x|),
\end{aligned}$$

where

$$D_i = \eta C_i \sup_{t\leq T}\left(1 + E\Big[\Big|\frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta}\Big|\Big]\right).$$

□

For the next lemma, we would like to define

$$G^\eta(t,x) = \sum_{i=1}^{k} l_i g_i^\eta(t,x). \tag{13}$$

**Lemma 2.** *For a fixed $\eta$ and $i \in \{1,2,\ldots,k\}$, if $|f_i(t,x) - f_i(t,y)| \le M|x-y|$ on $[0,T]$, then $g_i^\eta(t,x)$'s satisfy*

$$|g_i^\eta(t,x) - g_i^\eta(t,y)| \le M|x-y|,$$

*and if $|F(t,x) - F(t,y)| \le M|x-y|$, then $G^\eta(t,x)$ satisfies*

$$|G^\eta(t,x) - G^\eta(t,y)| \le M|x-y|.$$

*Proof.* For any $t \in [0,T]$,

$$\begin{aligned}
|g_i^\eta(t,x) - g_i^\eta(t,y)| &= \eta\left|E\left[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big)\right] - E\left[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{y}{\eta}\Big)\right]\right| \\
&= \eta\left|E\left[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big) - f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{y}{\eta}\Big)\right]\right| \\
&\le M|x-y|.
\end{aligned}$$

Since $M$ does not depend on $t$, $|f_i(t,x) - f_i(t,y)| \le M|x-y|$ on $[0,T]$. □

Hence, with the results in Lemmas 1, and 2, we now derive the adjusted fluid limit.

**Theorem 5** (Adjusted fluid limit)**.** *Under the same assumptions in Theorem 1, i.e., for all $t \in [0,T]$*

$$\big|f_i(t,x)\big| \le C_i\big(1+|x|\big) \quad \textit{for } i \in \{1,\ldots,k\}, \tag{14}$$

$$\big|F(t,x) - F(t,y)\big| \le M|x-y|, \tag{15}$$

*for a fixed $\eta$,*

$$\lim_{\nu\to\infty} \frac{Z^{\eta,\nu}(t)}{\nu} = \bar{Z}^\eta(t) \quad a.s., \tag{16}$$

*where $\bar{Z}^\eta(t)$ is the solution to the following integral equation:*

$$\bar{Z}^\eta(t) = x_0^\eta + \sum_{i=1}^{k} l_i \int_0^t g_i^\eta\big(s, \bar{Z}^\eta(s)\big)ds, \tag{17}$$

*and furthermore*

$$\bar{Z}^\eta(t) = E\big[X^\eta(t)\big] = x_0^\eta + \sum_{i=1}^{k} l_i \int_0^t \eta E\left[f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big)\right]ds. \tag{18}$$

*Proof.* From Lemmas 1 and 2, (14) and (15) imply

$$|g_i^\eta(t,x)| \le D_i(1+|x|) \quad \text{and} \quad |G^\eta(t,x) - G^\eta(t,y)| \le M|x-y|.$$

Therefore, by Theorem 1, we have equation (17), and by the definition of $g_i^\eta(t,x)$'s, we have equation (18). □

Comparing equation (18) with equation (11) in Theorem 4, we notice that Theorem 5 via equation (18) could provide the exact estimation of $E\big[X^\eta(t)\big]$. Once we have the adjusted fluid limit, we can derive the adjusted diffusion limit from it. The following theorem explains the adjusted diffusion limit.

**Theorem 6** (Adjusted diffusion limit)**.** *Under the same settings in Theorem 2, for a fixed $\eta$, suppose the Lebesgue measure of the set $\{t : \partial G^\eta(t, \bar{Z}^\eta(t))$ does not exist.$\}$ is zero. Define a sequence of scaled centered processes $\{V^{\eta,\nu}(t)\}$ on a time interval $[0,T]$ to be*

$$V^{\eta,\nu}(t) = \sqrt{\nu}\Big(\frac{Z^{\eta,\nu}(t)}{\nu} - \bar{Z}^\eta(t)\Big),$$

*where $Z^{\eta,\nu}(t)$ and $\bar{Z}^\eta(t)$ are solutions to equations (12) and (17) respectively. If $f_i(t,x)$'s and $F(t,x)$ satisfy equations (14) and (15) respectively, then $\lim_{\nu\to\infty} V^{\eta,\nu}(t) = V^\eta(t)$, where*

$$V^\eta(t) = \sum_{i=1}^k l_i \int_0^t \sqrt{g_i^\eta\big(s, \bar{Z}^\eta(s)\big)}\, dW_i(s) + \int_0^t \partial G^\eta\big(s, \bar{Z}^\eta(s)\big) V^\eta(s) ds,$$

*$W_i(\cdot)$'s are independent standard Brownian motions, and $\partial G^\eta\big(t, \bar{Z}^\eta(t)\big)$ is the gradient matrix of $G^\eta\big(t, \bar{Z}^\eta(t)\big)$ with respect to $\bar{Z}^\eta(t)$. Furthermore, $V^\eta(t)$ is a Gaussian process.*

*Proof.* From definition of $G^\eta(t,x)$ in (13), we can easily verify that $G^\eta(t,x)$ is differentiable almost everywhere, and hence $|G^\eta(t,x) - G^\eta(t,y)| \le M|x-y|$ implies

$$\left|\frac{\partial}{\partial x_i} G^\eta(t,x)\right| \le M_i \quad \text{for some } M_i < \infty, \text{ almost all } t \le T, \text{ and } i \in \{1,\dots,d\}.$$

Therefore, by Theorem 2, we prove this theorem. □

From Theorems 5 and 6, we obtain the adjusted fluid and diffusion limits from $Z^{\eta,\nu}(t)$. Recall that we call the fluid and diffusion limits in Section 3 as *standard* and the ones derived in this section as *adjusted* limits. Now one may ask a natural question. What is the relationship between the standard and adjusted limits? The following theorem suggests that the adjusted limits are asymptotically identical to the standard fluid and diffusion limits.

**Theorem 7** (Relationship between standard and adjusted limits)**.** *For $t \in [0,T]$, if $\eta f_i(t, x/\eta) = f_i(t,x)$ for $i \in \{1,2,\ldots,k\}$, then,*

$$\lim_{\eta\to\infty} \frac{\bar{Z}^\eta(t)}{\eta} = \bar{X}(t), \quad \textit{and} \tag{19}$$

$$\lim_{\eta\to\infty} \frac{V^\eta(t)}{\sqrt{\eta}} = D(t). \tag{20}$$

*Proof.* It is enough to show that $\lim_{\eta\to\infty} g^\eta(t,\eta x)/\eta = f_i(t,x)$ for $t \in [0,T]$, which results in the integral equations for LHS of equations (19) and (20) identical to those for RHS. If $\eta f_i(t, x/\eta) = f_i(t,x)$ for $i \in \{1,2,\ldots,k\}$,

$$\begin{aligned}
\lim_{\eta\to\infty} \frac{g^\eta(t,\eta x)}{\eta} &= \lim_{\eta\to\infty} E\left[ f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + x\Big)\right] \\
&= E\left[ \lim_{\eta\to\infty} f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + x\Big)\right] \quad \text{by uniform integrability} \\
&= f_i(t,x).
\end{aligned}$$

□

Now, we turn our attention to the solution of the adjusted fluid and diffusion limits for a fixed $\eta$. *Theoretically, Theorem 5 guarantee the exact estimation to $E\big[X^\eta(t)\big]$.* However, the functions $g_i^\eta$'s, in fact, cannot be identified unless we know the distribution of $X^\eta(t)$, which forces us to develop a methodology to approximate $g_i^\eta$'s for the sake of computational feasibility. Nonetheless, when applying the adjusted fluid limit to the multi-server queues with abandonment and retrials, we have a good candidate distribution to obtain $g_i^\eta$'s. So, the following section will describe a computational methodology to get approximated adjusted limits.

# 5. Approximation of adjusted limits with Gaussian density

In general, there is no clear way to find the distribution of $X^\eta(t)$. Without knowledge of it, it is not possible to obtain $g_i^\eta(t,\cdot)$. However, we could approximate its distribution based on the asymptotic distribution. As mentioned in Section 3.1, equation (7) implies that the distribution of $X^\eta(t)$ becomes closer to Gaussian distribution as $\eta$ (or the number of servers) increases, and it was experimentally shown that empirical density is actually quite close to the Gaussian density even if the number of servers are not very large: see left graph in Figure 4.

Similar empirical results are found in Mandelbaum and Pats (1998) and Mandelbaum et al. (2002). Therefore, using Gaussian distribution to approximate the distribution of $X^\eta(t)$ is reasonable especially when the number of servers is large. Note that the empirical density in Figure 4 is obtained when the queue is in the critically loaded phase.

Once we decide to use the Gaussian density, it provides the following two additional benefits:

1. Gaussian distribution can be completely characterized by the mean and covariance matrix which can be obtained from the fluid and diffusion limits.

2. By using Gaussian density, $g_i^\eta(t,\cdot)$'s will be smooth even if $f_i(t,\cdot)$'s are not, which enables us to apply Theorem 6 without measure-zero assumption.

The second benefit is not obvious and hence we provide a proof of that.

**Lemma 3.** *For any fixed $t > 0$, assume that $X^\eta(t)$ follows a multivariate normal distribution with the mean $\mu = (\mu_1, \ldots, \mu_d)'$ and covariance matrix $\Sigma$. Then, $g_i^\eta(t,x)$'s are differentiable everywhere with respect to $x$.*

*Proof.* WLOG, we prove this lemma for $\eta = 1$. Define

$$\phi(x,y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left( -\frac{(y-\mu+x)'\Sigma^{-1}(y-\mu+x)}{2} \right).$$

Using Gaussian density,

$$g_i^\eta(t,x) = \int_{\mathbf{R}^d} f_i(t,y)\phi(x,y)dy.$$

For $j \in \{1,\ldots,d\}$, since $\phi(x,y)$ is differentiable with respect to $x_j$ and $|f_i(t,y)\frac{d}{dx_j}\phi(x,y)|$ is integrable,

$$\begin{aligned} \frac{d}{dx_j} g_i^\eta(t,x) &= \frac{d}{dx_j} \int_{\mathbf{R}^d} f_i(t,y)\phi(x,y)dy \\ &= \int_{\mathbf{R}^d} f_i(t,y) \frac{d}{dx_j}\phi(x,y)dy \quad \text{by Theorem 2.27 in Folland (1999),} \end{aligned} \tag{21}$$

where $x_j$ is $j^{\text{th}}$ component of $x$. Therefore, $g_i^\eta(t,x)$ is differentiable with respect to $x_j$. □

**Remark 3.** *For $t > 0$, the covariance matrix, $\Sigma(t)$ is invertible unless at least one of the components is deterministic quantity. If so, we obtain invertible $\Sigma(t)$ by excluding those deterministic components from the state.*

Now, we have $g_i^\eta(t,\cdot)$'s which are differentiable. Then, we can apply Theorem 6 to obtain the diffusion limit for $Z^{\eta,\nu}(t)$.

Finally, we approximate the adjusted fluid and diffusion limits by utilizing Gaussian density. Therefore, we compare the adjusted limits with the empirical mean and covariance matrix. Note when we explain Theorem 5, we do not consider $\Sigma^\eta(t)$, the covariance matrix of $X^\eta(t)$. However, in order to obtain $g_i^\eta(t,\cdot)$'s from Gaussian density, we should consider $\Sigma^\eta(t)$. In order to reflect that, we rewrite $g_i^\eta(t,x)$'s as follows:

$$g_i^\eta\big(t,x\big) \;\to\; g_i^\eta\big(t,x,u\big) \quad \text{for } i\in\{1,\ldots,k\} \text{ and} \tag{22}$$

$$G^\eta\big(t,x\big) \;\to\; G^\eta\big(t,x,u\big). \tag{23}$$

Note that the $u$ term in equations (22) and (23) represents the covariance matrix of $X^\eta(t)$.

**Proposition 1** (Mean and covariance matrix using adjusted limits )**.** *The quantities $\bar{Z}^\eta(t)$ and $\Sigma^\eta(t)$ are obtained by solving the following simultaneous ordinary differential equations with initial values given by $\bar{Z}^\eta(0)=x_0^\eta$ and $\Sigma^\eta(0)=0$:*

$$\frac{d}{dt}\bar{Z}^\eta(t) \;=\; \sum_{i=1}^{k} l_i g_i^\eta\big(t,\bar{Z}^\eta(t),\Sigma^\eta(t)\big), \tag{24}$$

$$\frac{d}{dt}\Sigma^\eta(t) \;=\; A^\eta(t)\Sigma^\eta(t)+\Sigma^\eta(t)A^\eta(t)'+B^\eta(t)B^\eta(t)', \tag{25}$$

*where $A(t)$ is the gradient matrix of $G^\eta\big(t,\bar{Z}^\eta(t),\Sigma^\eta(t)\big)$ with respect to $\bar{Z}^\eta(t)$, and $B(t)$ is the $d\times k$ matrix such that its $i^{th}$ column is $l_i\sqrt{g_i^\eta\big(t,\bar{Z}^\eta(t),\Sigma^\eta(t)\big)}$.*

*Proof.* By rewriting (17) in Theorem 5 as a differential equation form, we have (24), and by Theorem 3, we have (25). Note that since both $\bar{Z}^\eta(t)$ and $\Sigma^\eta(t)$ are variables, we should solve (24) and (25) simultaneously. □

Now, it is the time to return to the multi-server queues as given in Section 2. We explain the procedure to obtain the mean and covariance matrix of those queues using adjusted limits.

1. Setting up variables: we have a two-dimensional state space. Define $\bar{Z}^\eta(t)=(\bar{z}_1^\eta(t),\bar{z}_2^\eta(t))'$ to be the adjusted fluid limit and $\Sigma^\eta(t)=\begin{pmatrix}\sigma_1^\eta(t)^2 & \mathrm{cov}^\eta(t)\\ \mathrm{cov}^\eta(t) & \sigma_2^\eta(t)^2\end{pmatrix}$ to be the covariance matrix of the adjusted diffusion limit.

2. Deriving $g$ functions: using Gaussian density, we can obtain new rate functions, $g_i^\eta(t,\cdot,\cdot)$'s, which correspond to $f_i(t,\cdot)$'s as follows:
For $x=(x_1,x_2)'$ and $u=\begin{pmatrix} u_1 & u_2 \\ u_2 & u_3 \end{pmatrix}$,

$$
\begin{aligned}
g_1^\eta(t,x,u) &= \eta\lambda_t,\\
g_2^\eta(t,x,u) &= \mu_t^2 x_2,\\
g_3^\eta(t,x,u) &= \mu_t^1\big(\eta n_t + (x_1-\eta n_t)\Phi(\eta n_t, x_1, \sqrt{u_1}) - u_1\phi(\eta n_t, x_1, \sqrt{u_1})\big),\\
g_4^\eta(t,x,u) &= \beta_t(1-p_t)\Big((x_1-\eta n_t)\big(1-\Phi(\eta n_t, x_1, \sqrt{u_1})\big) + u_1\phi(\eta n_t, x_1, \sqrt{u_1})\Big), \quad \text{and}\\
g_5^\eta(t,x,u) &= \beta_t p_t\Big((x_1-\eta n_t)\big(1-\Phi(\eta n_t, x_1, \sqrt{u_1})\big) + u_1\phi(\eta n_t, x_1, \sqrt{u_1})\Big),
\end{aligned}
$$

where $\Phi(a,b,c)$ and $\phi(a,b,c)$ are function values at point $a$ of the Gaussian CDF and PDF respectively with mean $b$ and standard deviation $c$.
The details of deriving $g_i^\eta$'s are provided in the Online Supplement (Section 1). Interestingly, typical queueing systems include rate functions that are constant (w.r.t. state variables) or of the form "$\min(\cdot,\cdot)$". Therefore, $g_i^\eta$'s derived here are reusable for other queues. Since $f_1(t,x)$ and $f_2(t,x)$ are constant and linear with respect to $x$ respectively, $g_1^\eta(t,x,u)=\eta f_1(t,x/\eta)$ and $g_2^\eta(t,x,u)=\eta f_2(t,x/\eta)$.

3. Constructing $A^\eta(t)$ and $B^\eta(t)$ matrices for diffusion limits: define

$$
\begin{aligned}
\Phi^\eta(t) &= \Phi(\eta n_t, \bar{z}_1^\eta(t), \sigma_1^\eta(t)),\\
\phi^\eta(t) &= \phi(\eta n_t, \bar{z}_1^\eta(t), \sigma_1^\eta(t)),\\
\alpha_1^\eta(t) &= \eta n_t + (\bar{z}_1^\eta(t)-\eta n_t)\Phi^\eta(t) - \sigma_1^\eta(t)^2\phi^\eta(t), \text{ and}\\
\alpha_2^\eta(t) &= (\bar{z}_1^\eta(t)-\eta n_t)(1-\Phi^\eta(t)) + \sigma_1^\eta(t)^2\phi^\eta(t).
\end{aligned}
$$

Recall that $A^\eta(t)$ and $B^\eta(t)$ satisfy

$$
\begin{aligned}
A^\eta(t) &= \partial G^\eta\big(t, \bar{Z}^\eta(t), \Sigma^\eta(t)\big) \text{ and}\\
B^\eta(t) &= \sum_{i=1}^k l_i \sqrt{g_i^\eta\big(t, \bar{Z}^\eta(t), \Sigma^\eta(t)\big)}.
\end{aligned}
$$

Note that $\partial G^\eta\big(t, \bar{Z}^\eta(t), \Sigma^\eta(t)\big)$ is the gradient matrix of $G^\eta\big(t, \bar{Z}^\eta(t), \Sigma^\eta(t)\big)$ with respect to $\bar{Z}^\eta(t)$ (Theorem 6). For the multi-server queues considered here, $A^\eta(t)$ and $B^\eta(t)$

are as follows:

$$A^\eta(t) = \begin{pmatrix} -\mu_t^1\Phi^\eta(t) - \beta_t(1-\Phi^\eta(t)) & \mu_t^2 \\ \beta_t(1-p_t)(1-\Phi^\eta(t)) & -\mu_t^2 \end{pmatrix}, \text{ and}$$

$$B^\eta(t) = \begin{pmatrix} \sqrt{\eta\lambda_t} & 0 \\ \sqrt{\mu_t^2\bar{z}_2^\eta(t)} & -\sqrt{\mu_t^2\bar{z}_2^\eta(t)} \\ -\sqrt{\mu_t^1\alpha_1^\eta(t)} & 0 \\ -\sqrt{\beta_t(1-p_t)\alpha_2^\eta(t)} & \sqrt{\beta_t(1-p_t)\alpha_2^\eta(t)} \\ -\sqrt{\beta_t p_t\alpha_2^\eta(t)} & 0 \end{pmatrix}'.$$

4. Solving ordinary differential equations (ODEs) to obtain $\bar{Z}^\eta(t)$ and $\Sigma^\eta(t)$: we can numerically solve the following system of ODEs using a mathematical software such as MATLAB (which only takes seconds to solve).

$$\begin{aligned} \frac{d}{dt}\bar{Z}^\eta(t) &= \sum_{i=1}^{k} l_i g_i^\eta\big(t, \bar{Z}^\eta(t), \Sigma^\eta(t)\big), \\ \frac{d}{dt}\Sigma^\eta(t) &= A^\eta(t)\Sigma^\eta(t) + \Sigma^\eta(t)A^\eta(t)' + B^\eta(t)B^\eta(t)'. \end{aligned}$$

Although we derive new rate functions for our adjusted limits, we need some intuition regarding how they contribute to increasing accuracy especially in the critically loaded phases. Thus, in the next section, we revisit the inaccuracy in the previous approaches and explain how our adjusted limits treat this.

# 6. Intuition behind the function $g_i^\eta$

In this section, we explain some intuition regarding the functions $g_i^\eta(t,\cdot)$'s. For the sake of clarity, we consider a simple $M_t/M_t/n_t$ queue which is a special case of the multi-server queues with abandonment and retrials ($\beta_t = 0$, and $\mu_t^1 = \mu_t$). We use $\eta = 1$ for the sake of illustration and remove the superscript $\eta$, i.e., we use $g_i(t,\cdot)$ instead of $g_i^\eta(t,\cdot)$. Let $X(t)$ denote the number of customers in the system at time $t$. Then, $X(t)$ is the solution to the following integral equation:

$$X(t) = X(0) + Y_1\Big(\int_0^t \lambda_s ds\Big) - Y_2\Big(\int_0^t \big(X(s)\wedge n_s\big)\mu_s ds\Big).$$

Here, for convenience, define $f_1(t,x) = \lambda_t$, $f_2(t,x) = \big(x\wedge n_t\big)\mu_t$, and $F(t,x) = \lambda_t - \big(x\wedge n_t\big)\mu_t$.

Applying theorems in Section 3.1, we have the fluid and diffusion limits, $\bar{X}(t)$ and $U(t)$ respectively, from the following integral equations:

$$\begin{aligned}\bar{X}(t) &= X(0) + \int_0^t \lambda_s - \big(\bar{X}(s) \wedge n_s\big)\mu_s ds, \text{ and} \\ U(t) &= U(0) + \int_0^t \Big(\sqrt{\lambda_s}, \sqrt{\big(\bar{X}(s) \wedge n_s\big)\mu_s}\Big)\begin{pmatrix} dW_1(s) \\ dW_2(s)\end{pmatrix} + \int_0^t \partial F(s, \bar{X}(s))U(s)ds,\end{aligned}$$

where

$$\partial F(t, \bar{X}(t)) = \begin{cases} -\mu_t & \text{if } \bar{X}(t) \le n_t, \\ 0 & \text{otherwise.}\end{cases}$$

Notice that the drift part $\partial F(t, \bar{X}(t))$ of the diffusion limit is completely determined by the fluid limit, and here we notice a possibility that the diffusion limit could produce the sharp spikes described in Section 3.2. When the $\bar{X}(t)$ is much smaller than the number of server $n_t$ (underloaded phase), then $Pr[X(t) \ge n_t]$ is likely to be very small, i.e. if we run several independent realizations of the process, only small fraction of them are in overloaded or critically loaded phases. In this case, the drift part of the diffusion limit is $-\mu_t$. Now, suppose that $\bar{X}(t)$ is smaller than but fairly close to $n_t$ (still underloaded phase). Then, $Pr[X(t) \ge n_t]$ would be relatively large. However, the drift part is still the same, $-\mu_t$. The drift part which significantly affects the covariance matrix structure does not reflect $Pr[X(t) \ge n_t]$ at all, while $Pr[X(t) \ge n_t]$ is closely related to the covariance matrix. Furthermore, consider the case where $\bar{X}(t)$ becomes slightly larger than $n_t$. Then, the drift part suddenly changes to zero. As a result, if $\bar{X}(t)$ is fluctuating close to $n_t$, then the drift part of the diffusion limit show repeated jumps between the values $-\mu_t$ and 0 although the state of the queue itself does not changes much. Undoubtedly, it produces sharp spikes in the covariance matrix as shown in Figure 5 and make the quality of the approximation worse.

Now, we turn our attention to the functions $g_i(t, \cdot)$'s. Let us follow the procedure to obtain $g_2(t, x)$. Note that $g_1(t, x) = f_1(t, x)$. Define $G(t, x) = g_1(t, x) - g_2(t, x) = \lambda_t - g_2(t, x)$. For a fixed $t_0$, let $X = X(t_0)$, $\mu = \mu_{t_0}$ and $n = n_{t_0}$ and $x = E[X]$. Then,

$$g_2(t_0, x) = E\big[\mu(X \wedge n)\big] = \mu\Big\{E[X\mathbb{I}_{X \le n}] + nPr[X > n]\Big\}. \tag{26}$$

From equation (26), notice the following characteristics of the function $g_2(t, x)$:

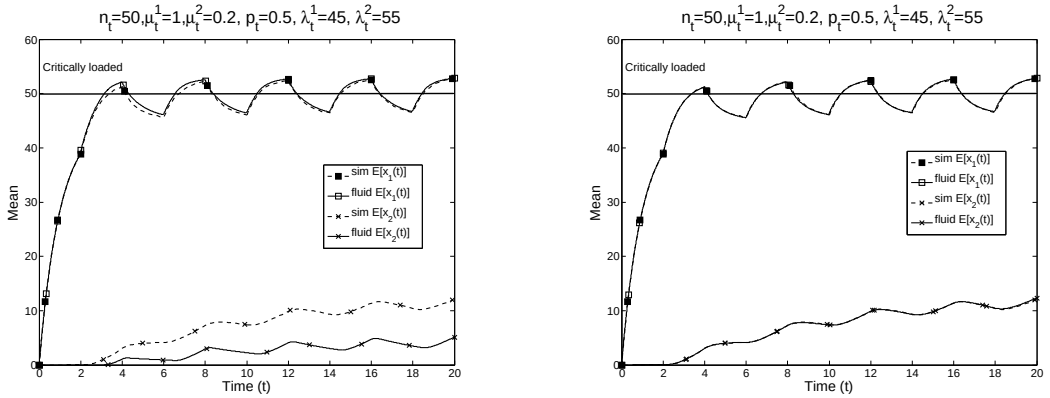
1. If $Pr[X > n]$ becomes closer to 1, $\partial g_2(t_0, x)/\partial x$ approaches 0;

2. If $Pr[X > n]$ gets closer to 0, $\partial g_2(t_0, x)/\partial x$ gets to be closer to $\mu$.

Notice that the drift part $\partial G(t,\cdot)$ determines its value between $-\mu_t$ and 0 according to $Pr[X(t) > n_t]$. Therefore, even if the queue makes phase transitions frequently, the drift part of the adjusted diffusion limit does not make sudden changes. In the following section, we can actually verify this intuition from many experimental results and see the effectiveness of the adjusted limits.

# 7. Numerical results

In this section, we show several numerical results to justify how effectively the adjusted limits approximate the multi-server queues with abandonment and retrials. We compare our adjusted limits against the standard ones. Under the similar settings in Mandelbaum et al. (2002), we use 5,000 independent simulation runs and use them as a reference model. We use constant rates for all parameters except the arrival rate. The arrival rate alternates between 45 and 55 every two time units. Figures 6 and 7 show the estimation of mean values from one experiment. The number of servers ($n_t$) is 50 and the service rate of each server is 1 for all $t \geq 0$.



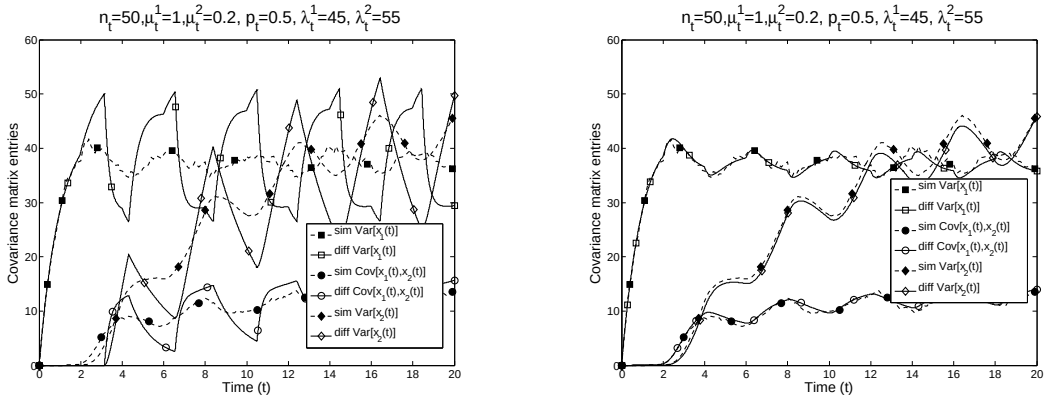

(a) Mean numbers by the fluid limit

(b) Mean numbers by the adjusted fluid limit

Figure 6: Comparison of mean values, $E\big[X(t)\big]$

As seen in Figure 6, the multi-server queue is nearly critically loaded, i.e. $\bar{x}_1(t) \approx n_t$. As Mandelbaum et al. (2002) points out, the standard fluid limit shows significant inaccuracy for $E\big[x_2(t)\big]$. On the other hand, the adjusted fluid limit provides an excellent approximation result. Especially, one can recognize remarkable improvement in the estimation of $E\big[x_2(t)\big]$.

For the mean value of $x_1(t)$, the adjusted fluid limit provides a lot better approximation result.




(a) Covariance matrix by the diffusion limit

(b) Covariance matrix by the adjusted diffusion limit

Figure 7: Comparison of covariance matrix entries, $Cov\big[X(t), X(t)\big]$

When we see the estimation of the covariance matrix, we also notice the adjusted diffusion limit shows dramatic improvement. As seen in Figure 7, the standard diffusion limit causes "spikes" as also pointed out in Section 3.2. The adjusted diffusion limit, however, provides excellent accuracy without spikes. Besides this specific example, in order to verify the effectiveness of our methodology, we conduct several experiments with different parameter combinations.

Table 1: Experiment setting

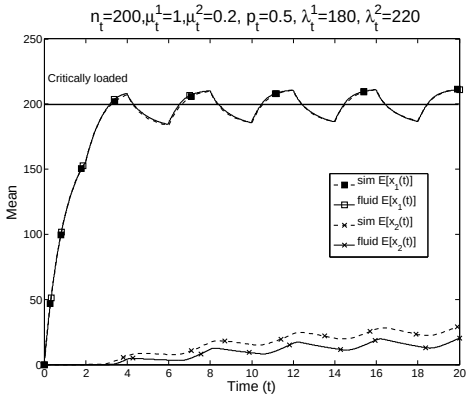
| exp | svrs | $\lambda_1$ | $\lambda_2$ | $\mu_1$ | $\mu_2$ | $\beta$ | prob | alter | time |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 45 | 55 | 1 | 0.2 | 2.0 | 0.5 | 2 | 20 |
| 2 | 200 | 180 | 220 | 1 | 0.2 | 2.0 | 0.5 | 2 | 20 |
| 3 | 300 | 270 | 330 | 1 | 0.2 | 2.0 | 0.5 | 2 | 20 |
| 4 | 400 | 360 | 440 | 1 | 0.2 | 2.0 | 0.5 | 2 | 20 |
| 5 | 400 | 390 | 410 | 1 | 0.2 | 2.0 | 0.5 | 2 | 20 |
| 6 | 50 | 45 | 55 | 1 | 2.0 | 2.0 | 0.5 | 2 | 20 |
| 7 | 50 | 45 | 55 | 1 | 0.2 | 5.0 | 0.5 | 2 | 20 |
| 8 | 50 | 45 | 55 | 1 | 0.2 | 2.0 | 0.9 | 2 | 20 |

Table 1 describes the setting of each experiment. In Table 1, "svrs" is the number of servers ($n_t$), "alter" is the time length for which each arrival rate lasts, and "time" is

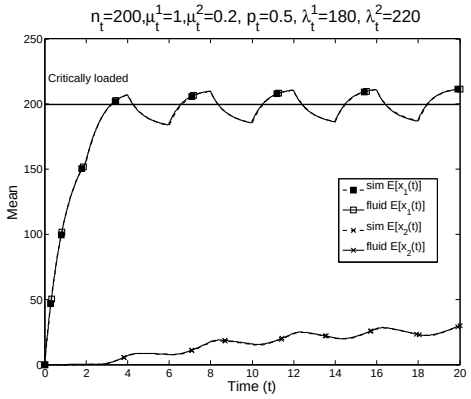
the end time of our analysis. We already recognize that the standard fluid and diffusion limits work well when it *does not linger* too long close to the critically loaded phase. For comparison, therefore, our experiments contain several cases where the system *does linger* relatively longer. Experiments 1-4 are intended to see how the estimation accuracy would be improved as we increase the number of servers along with the arrival rates. Experiment 5 is set in order to observe the effect of *lingering* in the nearly critically loaded phase even if we the number of servers is fairly large. We change parameters other than number of servers and arrival rates in experiments 6-8 to see the effects of them. In fact, from a large number of experiments not listed in Table 1, we observe that they do not affect estimation accuracy significantly.

Here we explain the overall results. As seen in Figures 8 and 9, the standard fluid and diffusion limits show improvement in estimation accuracy when we scale the number of servers along with the arrival rates since the standard limits are asymptotically true. In fact, the improvement when using the adjusted limits does not seem obvious. However, it is because they already provide excellent estimation results even when the number of servers is relatively small, and the adjusted limits always outperform the standard fluid and diffusion limits. We also see the effect of lingering near the critically loaded phase in Figure 10. Although the number of servers is fairly large, it does debase the quality of approximations significantly when we use the standard fluid and diffusion limits. On the other hand, the adjusted ones provide excellent accuracy for both mean and covariance matrix in all cases.

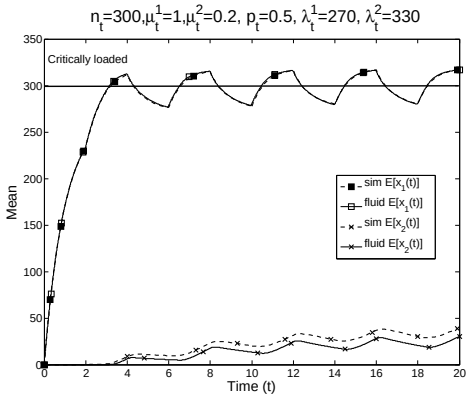
Figure 11 illustrates the average percentile difference of both approaches against simulation. Figure 11 (a) is obtained by averaging all difference across time. From Figure 11 (a), we notice that the adjusted limits show promise relative to the standard ones. In Figure 11 (b), we graph the differences especially at the time points when the queues are critically loaded. It turns out that the adjusted limits are still accurate, while the standard ones show even worse accuracy as expected. Note that, in Figure 11 (b), huge estimation difference, more than 300%, is observed when estimating $Cov[x_1(t), x_2(t)]$ using the standard fluid and diffusion limits. However, the graph is cropped at the 70% level for the illustration purpose. We know that those results are from our limited experiments and hence do not make an absolute conclusion about two methodologies. Nonetheless, we could not deny the fact that the adjusted limits provide accurate estimation results consistently, while the standard fluid and diffusion limits result in inconsistent accuracy.



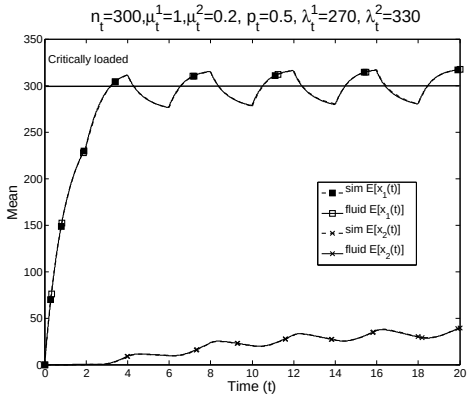

(a) Standard fluid limit of exp. 2



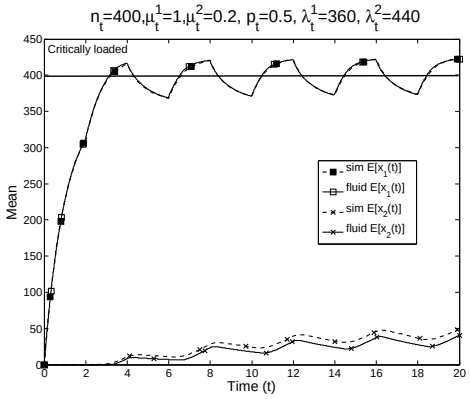

(b) Adjusted fluid limit of exp. 2



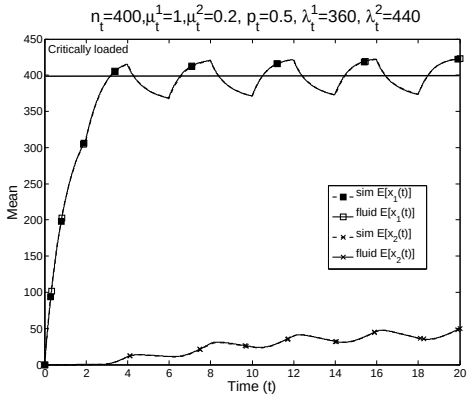

(c) Standard fluid limit of exp. 3




(d) Adjusted fluid limit of exp. 3

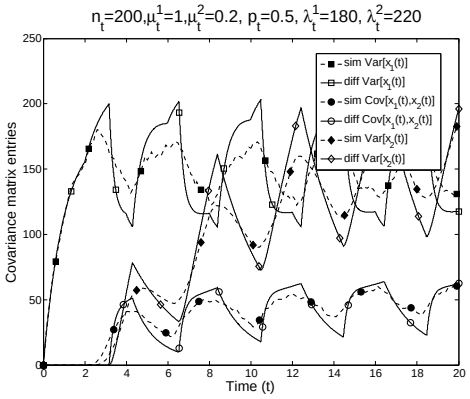



(e) Standard fluid limit of exp. 4

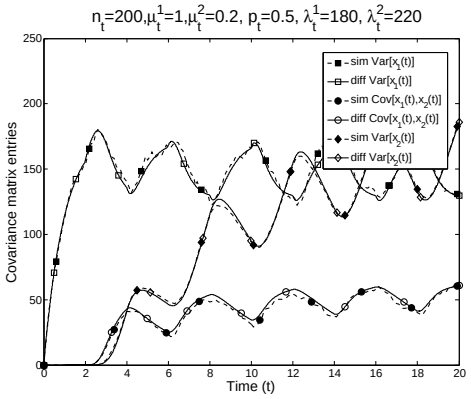



(f) Adjusted fluid limit of exp. 4

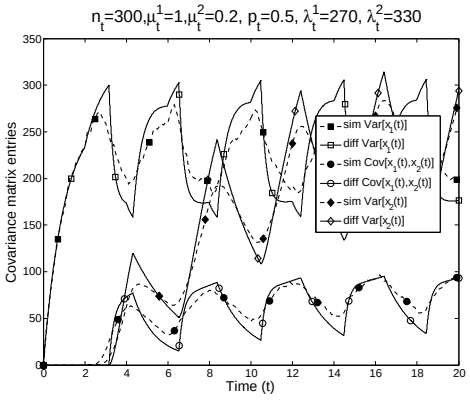
Figure 8: Comparison of mean values, $E\big[X(t)\big]$



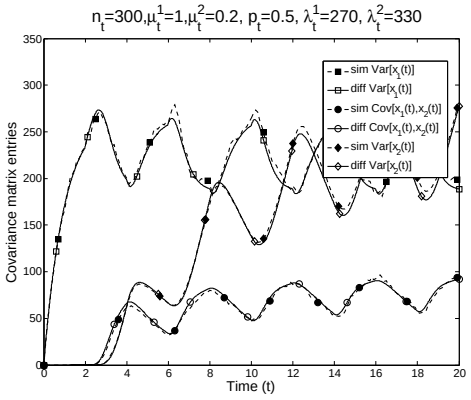

(a) Standard diffusion limit of exp. 2



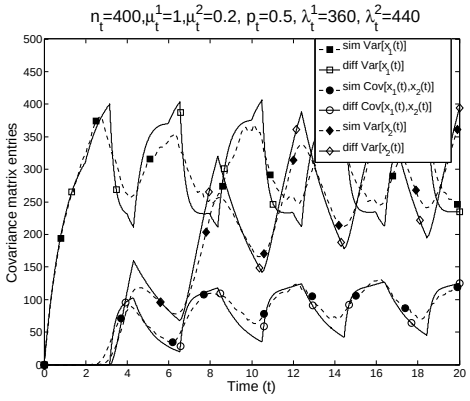

(b) Adjusted diffusion limit of exp. 2



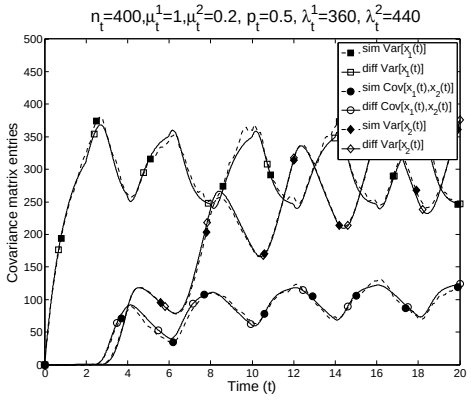

(c) Standard diffusion limit of exp. 3




(d) Adjusted diffusion limit of exp. 3

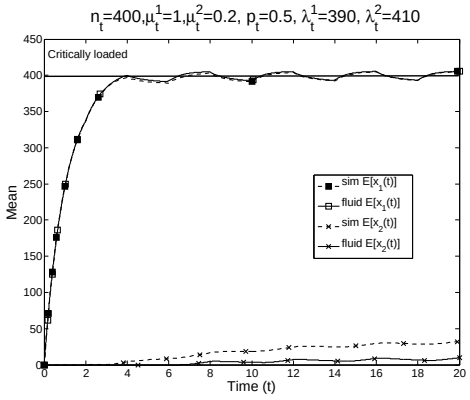



(e) Standard diffusion limit of exp. 4

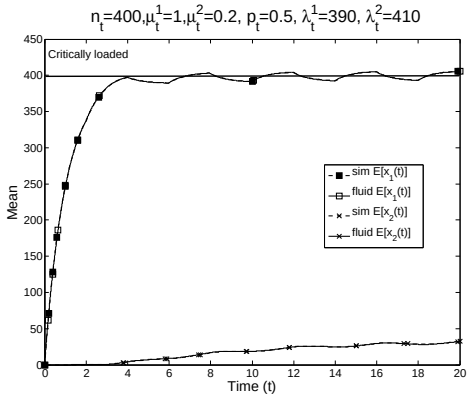



(f) Adjusted diffusion limit of exp. 4

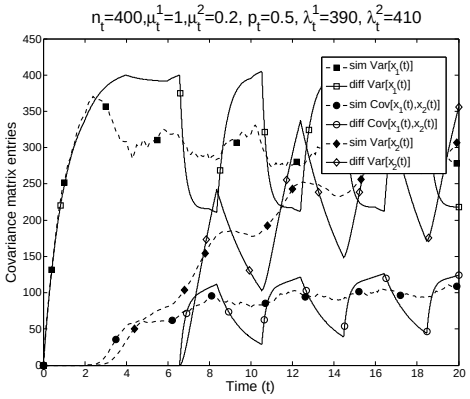
Figure 9: Comparison of covariance matrices, $Cov\big[X(t), X(t)\big]$



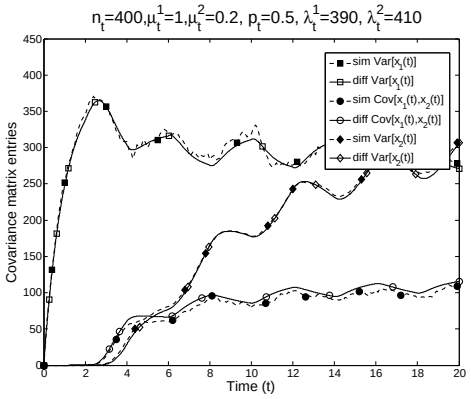

(a) Fluid limit of exp. 5




(b) Adjusted fluid limit of exp. 5

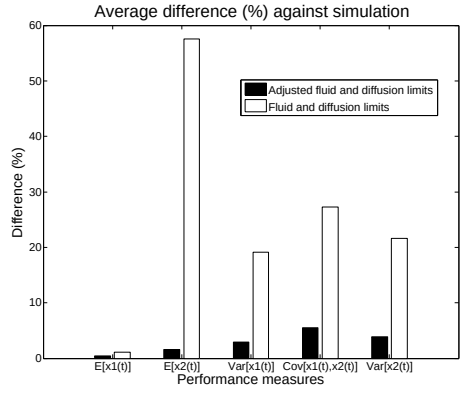



(c) Diffusion limit of exp. 5

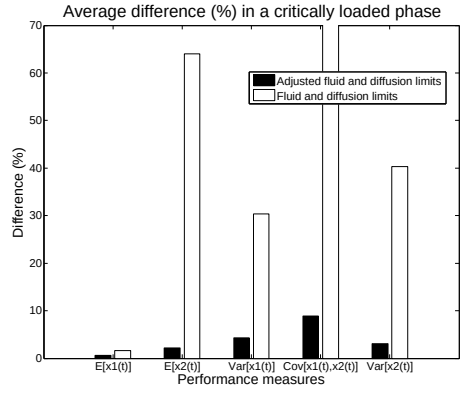



(d) Adjusted diffusion limit of exp. 5

Figure 10: Comparison in the nearly critically loaded phase




(a) Average difference for all experiments




(b) Average difference at a critically loaded point

Figure 11: Average difference against simulation

# 8. Conclusion

In this paper, we explain the fluid and diffusion limits used in the analysis of time-varying multi-server queues with abandonment and retrials and show potential problems that one faces in obtaining accuracy in the nearly critically loaded phase. To address those problems, we proposed adjusted fluid and diffusion limits *specifically designed for the approximation of the multi-server queues with finite number of servers*. Since the existing fluid and diffusion limits are *asymptotically* true, we show that our adjusted limits are also asymptotically true. It turns out that the adjusted limits achieve great improvement in approximation accuracy of performance measures, which was verified by a number of numerical experiments. Due to space restriction, we have not shown all examples where our method (i.e., the adjusted limits) works well.

As standard approach holds for a general class of queues *in theory*, our adjusted limits are applicable for their analysis as well since standard and adjusted limits are asymptotically identical (Theorem 7). We, however, observe that in some other types of queues other than multi-server queues considered here, e.g. multi-class queues, our approach may not provide accurate results. It is because empirical density for those queues may not be close to Gaussian density, and the convergence to Gaussian density is slow. Nevertheless, standard limits show even worse results *as approximations*, and adjusted limits outperform them. For the reference, we provide some numerical results for multi-class preemptive queues in the Online Supplement (Section 2).

For better approximations to those types of queues, one can investigate the properties of specific rate functions that affect the shape of empirical density and can devise a new methodology to find the functions $g_i^\eta(\cdot,\cdot)$'s from other density functions in the future. In addition, one can consider reflected diffusions (as opposed to the ones in this paper) especially in cases where non-negativity is an issue, such as a small number of servers or low loads.

# Acknowledgments

The authors are grateful to Dr. William A. Massey and Dr. Martin I. Reiman for their inputs and valuable discussions. The authors also thank two anonymous reviewers, associate editor and area editor for their insightful comments and suggestions that significantly improved the content and presentation of this paper. This research was partially supported by NSF grant

CMMI-0946935.

# References

Arnold, Ludwig. 1992. *Stochastic Differential Equations: Theory and Applications*. Krieger Publishing Company.

Billingsley, Patrick. 1999. *Convergence of Probability Measures*. A John Wiley & Sons, Inc., Publication.

Ethier, Stewart N., Thomas G. Kurtz. 1986. *Markov Processes: Characterization and Convergence*. 1st ed. A John Wiley & Sons, Inc., Publication.

Folland, Gerald B. 1999. *Real Analysis : Modern Techniques and Their Applications*. 2nd ed. A John Wiley & Sons, Inc., Publication.

Garnet, O., Avi Mandelbaum, Martin I. Reiman. 2002. Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management* **4** 208–227.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research* **29** 567–588.

Hampshire, Robert C., Otis B. Jennings, William A. Massey. 2009. A Time-Varying Call Center Design via Lagrangian Mechanics. *Probability in the Engineering and Informational Sciences* **23** 231–259.

Kurtz, Thomas G. 1978. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications* **6** 223–240.

Mandelbaum, Avi, William A. Massey, Martin I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* **30** 149–201.

Mandelbaum, Avi, William A. Massey, Martin I. Reiman, Alexander Stolyar, Brian Rider. 2002. Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials. *Telecommunication Systems* **21** 149–171.

Mandelbaum, Avi, Gennady Pats. 1995. State-dependent queues: approximations and applications. *Institute for Mathematics and Its Applications* **71** 239–282.

Mandelbaum, Avi, Gennady Pats. 1998. State-Dependent Stochastic Networks. Part I: Approximations and Applications with Continuous Diffusion Limits. *The Annals of Applied Probability* **8** 569–646.

Mandelbaum, Avishai, Sergey Zeltyn. 2009. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. *Operations Research* **57** 1189–1205.

Massey, William A, Ward Whitt. 1998. Uniform acceleration expansions for Markov chains with time-varying rates. *The Annals of Applied Probability* **8** 1130–1155.

Pang, Guodong, Ward Whitt. 2009. Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems* **61** 167–202.

Puhalskii, Anatolii A., Martin I. Reiman. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability* **32** 564–595.

Whitt, Ward. 2002. *Stochastic Process Limits*. 1st ed. Springer.

Whitt, Ward. 2006a. Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments. *Management Science* **50** 1449–1461.

Whitt, Ward. 2006b. Fluid Models for Multiserver Queues with Abandonments. *Operations Research* **54** 37–54.

Zeltyn, Sergey, Avi Mandelbaum. 2005. Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/n+ G Queue. *Queueing Systems* **51** 361–402.