MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

A Stochastic Analysis of Queues with Customer Choice and Delayed Information

Jamol Pender
School of Operations Research and Infomation Engineering, Cornell University, Ithaca, NY 14850 jjp274@cornell.edu,

Richard Rand
Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14850, rand@math.cornell.edu,

Elizabeth Wesson
Center for Applied Mathematics, Cornell University, Ithaca, NY 14850 enw27@cornell.edu,

Many service systems provide queue length information to customers thereby allowing customers to choose among many options of service. However, queue length information is often delayed and is often not provided in real time. Recent work by Dong et al. [9] explores the impact of these delays in an empirical study in U.S. hospitals. Work by Pender et al. [32] uses a two-dimensional fluid model to study the impact of delayed information and determine the exact threshold under which delayed information can cause oscillations in the dynamics of the queue length. In this work, we confirm that the fluid model analyzed by Pender et al. [32] can be rigorously obtained as a functional law of large numbers limit of a stochastic queueing process and we generalize their threshold analysis to arbitrary dimensions. Moreover, we prove a functional central limit theorem for the queue length process and show that the scaled queue length converges to a stochastic delay differential equation. Thus, our analysis sheds new insight on how delayed information can produce unexpected system dynamics.

Key words: queues, delayed information, delay differential equations, fluid model, diffusion limits, smartphone apps, oscillations

1. Introduction

Smartphone technology has changed the paradigm for communication between customers and service systems. One example of this communication is delay announcements, which have become important tools for managers to inform customers of their estimated waiting time. As a result, there is tremendous value in understanding the impact of providing waiting time or queue length information to customers. These announcements can affect the decisions of customers as well as the queue length dynamics of the system. Thus, the development of methods to support such announcements and interaction with customers has attracted the attention of the operations research community and is growing steadily.

Most of the current research that analyzes the impact of providing queue length or waiting time information to customers tends to focus on the impact of delay announcements. Delay annoucements are useful tools for managers of call centers and service systems to be able to interact and notify customers of their expected waiting time. For the most part, the literature only explores how customers respond to the delay announcements. Previous work by Armony and Maglaras [4], Guo and Zipkin [14], Hassin [18], Armony et al. [5], Guo and Zipkin [15], Jouini et al. [21, 22], Allon and Bassamboo [1], Allon et al. [2], Ibrahim et al. [19], Whitt [34] and references therein focus on

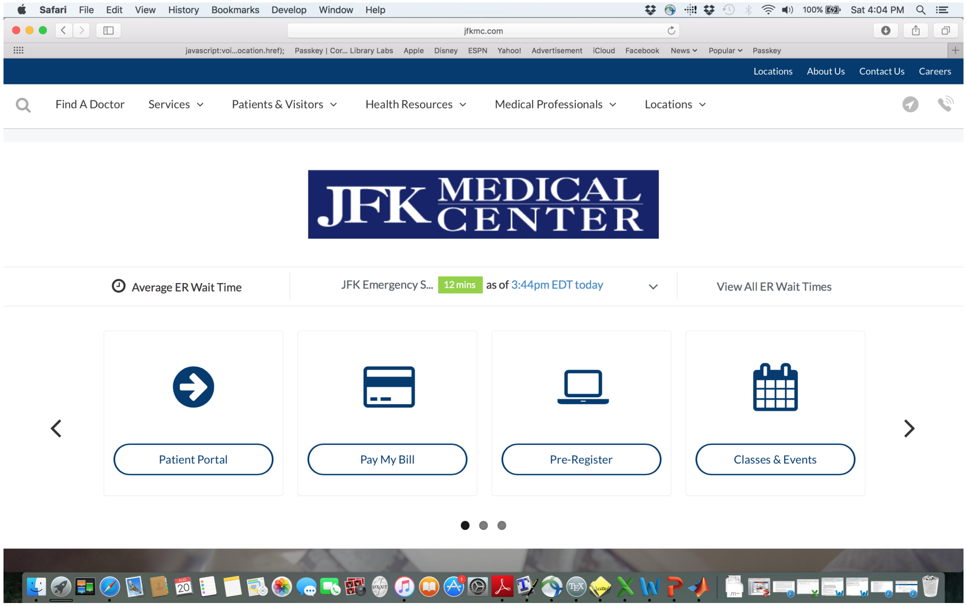


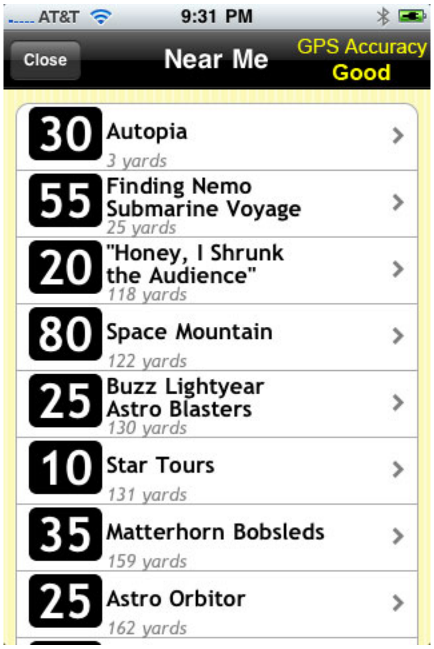
FIGURE 1. JFK Medical Center Online Reporting.

this aspect of the annoucements. Thus, previous work does not focus on the situation where the information given to customers in the form of an announcement is delayed and how this delay in information can affect the dynamics of the underlying service.

The analysis of this paper is similar to the main thrust of the delay announcement literature in that it is concerned with the impact of the information on the dynamics of the queueing process. However, it differs from the mainstream literature since we focus on when the information itself is delayed and is not given to customers in real-time. This is an important distinction from the current literature, which focuses on delay announcements given in real-time. Moreover, we should mention that this work also applies to systems where the delay information could be given in real-time, but the customer needs time to travel to receive their service. This is common for services that use app technology. Smartphone apps allow customers to join the system before arriving at the service and in this context the travel time is the delay of information. One example of a system is the Citibike bikesharing network in New York City. Customers can look and see where there are bikes on an app. However, in the time it takes for them to leave their home and get to a station, all of the bikes could have disappeared. Thus, the information that they used was in real-time, however, their travel time makes it delayed and somewhat unreliable.

Recently, there also is work that considers how the loss of information can impact queueing systems. Work by Jennings and Pender [20], Pender [28, 29] compares ticket queues with standard queues. In a ticket queue, the manager is unaware of when a customer abandons and is only notified of the abandonment when the customer would have entered service. This artificially inflates the queue length process and the work of Jennings and Pender [20], Pender [28, 29] compares the difference in queue length between the standard and ticket queue. However, this work does not consider the aspect of customer choice and delays in providing the information to customers, which is the case in many healthcare settings.

One important application of our work is in healthcare systems and networks. Recently, many healthcare providers have started to post their waiting times and queue lengths online, highway billboards, and even through apps. One example of this type of posting is given in Figure 1, which is an online snapshot of the waiting time at JFK Medical Center in Boynton Beach, Florida. In Figure 1, the average wait time is reported to be 12 minutes. However, in the top right of the figure we see that the time of the snapshot was 4:04pm while the time of a 12 minute wait is as of



30	Autopia 3 yards
55	Finding Nemo Submarine Voyage 25 yards
20	"Honey, I Shrunk the Audience" 118 yards
80	Space Mountain 122 yards
25	Buzz Lightyear Astro Blasters 130 yards
10	Star Tours 131 yards
35	Matterhorn Bobsleds 159 yards
25	Astro Orbitor 162 yards

Figure 2. Disneyland Park Wait Times App.

3:44pm. Thus, there is a delay of 20 minutes in the reporting of the wait times in the emergency room and this can have an important impact on the system dynamics as we will show in the rest of the paper.

Another relevant application of our work is for amusement parks like Disneyland or Six Flags. In Figure 6, we show a snapshot of the Disneyland app. The Disneyland app lists waiting times and the rider's current distance from each ride in the themepark. Customers obviously have the opportunity to choose which ride that they would want to go on, however, this choice depends on the information that they are given through the app. However, the wait times on the app might not be posted in real-time or customers might need travel time to get to their next ride, the information they make their decision on is essentially delayed. Thus, our queueing analysis is useful for Disney to understand how their decision to offer an app that displays waiting time information will affect the lines for rides in the park.

This paper introduces a stochastic queueing model, which describes the dynamics of customer choice with delayed information. In the queueing model, the customer receives information about the queue length which is delayed by a constant parameter Δ. Using strong approximations theory, we are able to prove fluid and diffusion limit theorems for our queueing model. We show that the fluid limit is a deterministic delay differential equation and the diffusion limit is a stochastic delay differential equation. We analyze the fluid limit in steady state and show that there exists an explicit threshold that governs whether all queues will oscillate or synchronize in steady state. Thus, when the lag in information is small, all queues will be balanced in steady state and when the delay is large enough, all queues are not balanced and have asynchronous dynamics. Our analysis combines theory from delay differential equations, customer choice models, and stability analysis of differential equations, strong approximations, and stochastic analysis.

1.1. Main Contributions of Paper

The contributions of this work can be summarized as follows:

- We use strong approximations for Poisson processes to derive fluid and diffusion limits showing for a stochastic queueing model with customer choice and delayed information. We show that the fluid limit yields a system of delay differential equations and the diffusion limit yields a system of stochastic delay differential equations. We highlight that the fluid and diffusion limits are non-trivial as they have delays and delays introduce many new complexities.

- We analyze the steady state dynamics of the fluid limit and determine the exact critical delay threshold that governs whether a Hopf bifurcation will occur. To do this, we show that we can reduce the analysis of a system of N equations to just analyzing two delay differential equations for stability purposes. We also prove that as the number of queues is increased and all other parameters stay fixed, the stability region of the delay differential system is increased.

1.2. Organization of Paper

The remainder of this paper is organized as follows. Section 2 describes a constant delay fluid model. We derive the critical delay threshold under which the queues are balanced if the delay is below the threshold and the queues are asynchronized if the delay is above the threshold. We also prove the fluid limit for our stochastic model and show that it converges to a system of delay differential equations. Section 3 establishes the existence and uniqueness of our diffusion limit and also shows that the centered rescaled stochastic queue length process converges to a system of stochastic delay differential equations. Finally in Section 4, we conclude with directions for future research related to this work.

2. Constant Delay Queueing Model

In this section, we present a new stochastic queueing model with customer choice based on the queue length with a constant delay. Thus, we begin with N infinite-server queues operating in parallel, where customers make a choice of which queue to join by taking the size of the queue length into account via a customer choice model. We assume that the total arrival rate to the system (sum of all queues) is λ and the service rate at each queue is given by μ. However, we add the twist that the queue length information that is given to the customer is delayed by a constant Δ for all of the queues. Therefore, the queue length that the customer receives is actually the queue length Δ time units in the past.

Since customers will decide on which queue to join based on the queue length information, the choice model that we use to model the customer choice dynamics is identical to that of a Multinomial Logit Model (MNL). The MNL model has an economic interpretation where we assume that the utility for being served in the i^{th} queue with delayed queue length $Q_i(t-\Delta)$ is $u_i(Q_i(t-\Delta)) = -Q_i(t-\Delta)$. Thus, in a stochastic context with N queues, the probability of going to the i^{th} queue is given by the following expression

$$p_i(Q(t),\Delta) = \frac{\exp(-\theta Q_i(t-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j(t-\Delta))} \tag{1}$$

where $Q(t) = (Q_1(t), Q_2(t), ..., Q_N(t))$.

It is evident from the above expression that if the queue length in station i is larger than the other queue lengths, then the i^{th} station has a smaller likelihood of receiving the next arrival. This decrease in likelihood as the queue length increases represents the disdain customers have for waiting in longer lines. We should also mention that the Multinomial Logit Model we present in this work can be viewed as smoothed and infinitely differentiable approximation of the join the shortest queue model. Using these probabilities for joining each queue allows us to construct the following stochastic model for the queue length process of our N dimensional system for $t \geq 0$

$$Q_i(t) = Q_i([-\Delta,0]) + \Pi_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j(s-\Delta))}ds\right) - \Pi_i^d\left(\int_0^t \mu Q_i(s)ds\right) \tag{2}$$

where each $\Pi(\cdot)$ is a unit rate Poisson process and $Q_i(s) = \varphi_i(s)$ for all $s \in [-\Delta, 0]$. In this model, for the i^{th} queue, we have that

$$\Pi_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j(s-\Delta))}ds\right) \tag{3}$$

counts the number of customers that decide to join the i^{th} queue in the time interval $(0,t]$. Note that the rate depends on the queue length at time $t-\Delta$ and not time t, hence representing the lag in information. Similarly

$$\Pi_i^d\left(\int_0^t \mu Q_i(s)ds\right) \tag{4}$$

counts the number of customers that depart the i^{th} queue having received service from an agent or server in the time interval $(0,t]$. However, in contrast to the arrival process, the service process depends on the current queue length and not the past queue length.

2.1. Large Customer Scaling and Fluid Limits

In many service systems, the arrival rate of customers is high. For example in Disneyland there are thousands of customers moving around the park and deciding on which ride they should join. Motivated by the large number of customers, we introduce the following scaled queue length process by a parameter η

$$Q_i^\eta(t) = Q_i^\eta([-\Delta,0]) + \frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right) - \frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right). \tag{5}$$

Note that we scale the rates of both Poisson processes, which is different from the many server scaling, which would only scale the arrival rate. Scaling only the arrival rate would yield a different limit than the one analyzed by Pender et al. [32] since the multinomial logit function is not a homogeneous function. Moreover, one should observe the term $Q_i^\eta([-\Delta,0])$, which highlights an important difference between delayed systems and their real-time counterparts. $Q_i^\eta([-\Delta,0])$ is a necessary function that keeps track of the past values of the queue length on the interval $[-\Delta,0]$. Unlike the case when $\Delta=0$, we need more than an initial value $Q_i^\eta(0)$ to initialize our stochastic queue length process. In fact in the delayed setting, we need an initial function to initialize our stochastic queue length process. We need these values since our arrival rate function is delayed and depends on previous queue length information. By letting the scaling parameter η go to infinity, we obtain our first result.

Theorem 1. *If $Q_i^\eta(s)\to\varphi_i(s)$ almost surely for all $s\in[-\Delta,0]$ and for all $1\le i\le N$, then the sequence of stochastic processes $\{Q^\eta(t)=(Q_1^\eta(t),Q_2^\eta(t),...,Q_N^\eta(t)\}_{\eta\in\mathbb{N}}$ converges almost surely and uniformly on compact sets of time to $(q(t)=(q_1(t),q_2(t),...,q_N(t))$ where*

$$\overset{\bullet}{q}_i(t) = \lambda\cdot\frac{\exp(-\theta q_i(t-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(t-\Delta))} - \mu q_i(t), \tag{6}$$

$q_i(s)=\varphi_i(s)$ for all $s\in[-\Delta,0]$ and for all $1\le i\le N$, and where $\varphi_i(s)$ is assumed to be a Lipschitz continuous function that keeps track of the previous values on the interval $[-\Delta,0]$.

See Appendix.

This result states that as we let η go towards infinity, the sequence of queueing processes converges to a system of **delay differential equations**. Unlike ordinary differential equations, the existence and uniqueness results for delay differential equations is much less well known. However, we provide the result of existence and uniqueness for the delay differential system that we analyze in this paper below.

Theorem 2. *Given a Lipschitz continuous initial function $\varphi_i:[-\Delta,0]\to\mathbb{R}$ for all $1\le i\le N$ and a finite time horizon $T>0$, there exists a unique Lipschitz continuous function $q(t)=\{q(t)\}_{-\Delta\le t\le T}$ that is the solution to the following delay differential equation*

$$\overset{\bullet}{q}_i(t) = \lambda\cdot\frac{\exp(-\theta q_i(t-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(t-\Delta))} - \mu q_i(t) \tag{7}$$

and $q_i(s)=\varphi_i(s)$ for all $s\in[-\Delta,0]$ and for all $1\le i\le N$.

The proof of this result can be found in Hale [17].

2.2. Hopf Bifurcations in the Constant Delay Model Similar to ordinary differential equations (ODEs), a DDE is exponentially stable if and only if all eigenvalues lie in the open left complex half-plane, see for example Hale [17]. However, a major difference between the two is that, unlike ODEs, the spectrum of DDEs has a countably infinite number of eigenvalues and are truly infinite dimensional objects. Fortunately, it can be shown in Hale [17] that there are only a finite number of eigenvalues to the right of any vertical line in the complex plane. This implies that there are only finite number of eigenvalues that yield unstable or oscillatory dynamics for the DDE.

As a result, in the one delay setting, it is of particular interest to determine for what value of the delay Δ makes the system of DDEs given in Equation 7 stable. The set of values for Δ that yield only eigenvalues in the left half of the complex plane of is referred to as the stability region of the DDE. Furthermore, the complement of the stability region is the region of instability. Thus, the point at which the DDE system switches from being stable to unstable is defined as the critical delay Δ_{cr}. This critical delay value in the single delay setting is important for a complete stability analysis to determine whether the DDE system will converge to the equilibrium or oscillate around it.

In this paper, we focus on the derivation of critical delay for the DDE system given in Equation 7. Recent work by Pender et al. [32] explores a two dimensional version of our fluid limit and uncovers that the two queues can oscillate in equilibrium when the delay Δ is large enough. Pender et al. [32] also characterizes the critical delay Δ_{cr} in terms of the model parameters and provides an exact formula for the critical delay in the two dimensional case. However, this analysis is limited and does not immediately generalize to the multi-dimensional setting. The main goal of this section is to generalize the critical delay analysis of Pender et al. [32] and derive the exact critical delay for an arbitrary number of queues.

Theorem 3. *For the constant delay choice queueing model given in Equation 7 with arbitrary $N \geq 2$, the critical delay, $\Delta_{cr}(\lambda, \mu, \theta, N)$, is given by the following expression*

$$\Delta_{cr}(\lambda, \mu, \theta, N) = \frac{N \cdot \arccos\left(\frac{-\mu \cdot N}{\lambda \theta}\right)}{\sqrt{\lambda^2\theta^2 - N^2 \cdot \mu^2}}. \tag{8}$$

Proof: The first part of the proof is to compute an equilibrium for the solution to the delay differential equations. In standard ordinary differential equations, one sets the time derivative of the differential equations to zero and solve for the value of the queue length that makes it zero. This implies that we set

$$\overset{\bullet}{q}_i(t) = 0 \tag{9}$$

This further implies that we need to solve the following N nonlinear delay equations

$$\lambda \cdot \frac{\exp(-\theta q_i(t-\Delta))}{\sum_{j=1}^{N} \exp(-\theta q_j(t-\Delta))} - \mu \cdot q_i(t) = 0 \tag{10}$$

Sometimes finding the equilibrium is non-trivial in many non-linear systems. In our system, we also have the complication that the differential equations are delay differential equations and have an extra complexity. However, in our case, the delay differential equations given in Equation 10 are symmetric and this simplifies some of the analysis. In this case the N equations converge to the same point since in equilibrium each queue will receive exactly $1/N$ of the arrivals and the service rates of all of the queues are the same. Thus, we have in equilibrium that for all $1 \leq i \leq N$

$$q_i(t-\Delta) = q_i(t) = \frac{\lambda}{N\mu} \quad \text{as } t \to \infty. \tag{11}$$

To mathematically verify that this is an equilibrium for the system of equations, one can substitute $\frac{\lambda}{N\mu}$ for $q_i(t)$ and $q_i(t-\Delta)$ and make the observation that the time derivative for all of the equations are equal to zero. However, we may be unsure of whether the equilibrium is unique. We can show that the equilibrium in our setting is unique by noting that

$$\overset{\bullet}{q}_i(t) = 0 \tag{12}$$

and setting the equilibrium $q_i(\infty) = c_i$. Thus, for each i, we have that

$$\lambda \cdot \frac{\exp(-\theta c_i(t-\Delta))}{\sum_{j=1}^{N} \exp(-\theta c_j(t-\Delta))} = \mu \cdot c_i. \tag{13}$$

This implies that

$$\frac{\exp(-\theta c_i)}{c_i} = \frac{\mu}{\lambda} \cdot \sum_{j=1}^{N} \exp(-\theta c_j) = \text{constant}. \tag{14}$$

Now we observe that the function on the left $\frac{\exp(-\theta c_i)}{c_i}$ is a one-to-one function of $c_i \geq 0$. Therefore, all of the functions $\frac{\exp(-\theta c_i)}{c_i}$ are equal implies that all of the c_i terms are equal. This implies that our equilibrium is unique.

Now that we have established the unique equilibrium for Equation 7, we need to understand the stability of the delay differential equations near the equilibrium. The first step in doing this is to set each of the queue lengths to the equilibrium points plus a perturbation. With this in mind, we substitute the following values for each of the queue lengths

$$q_i(t) = \frac{\lambda}{N\mu} + u_i(t) \tag{15}$$

In this substitution, the $u_i(t)$ are pertubations about the equilibrium point $\frac{\lambda}{N\mu}$. By substituting Equation 15 into Equation 7 we get the following equations

$$\overset{\bullet}{u}_i(t) = \lambda \cdot \frac{\exp(-\theta u_i(t-\Delta))}{\sum_{j=1}^{N} \exp(-\theta u_j(t-\Delta))} - \mu u_i(t) - \frac{\lambda}{N} \tag{16}$$

Now if we linearize around the point $u_i(t) = 0$, which is equivalent to performing a Taylor expansion and keeping only the linear terms, we have that the linearized version of $u_i(t)$, which we now defined as $w_i(t)$ solve the following linear delay differential equations

$$\overset{\bullet}{w}_i(t) = -\frac{\lambda \cdot \theta \cdot (N-1)}{N^2} \cdot w_i(t-\Delta) + \sum_{j \neq i}^{N} \frac{\lambda \cdot \theta}{N^2} \cdot w_j(t-\Delta) - \mu \cdot w_i(t) \tag{17}$$

$$= -\frac{\lambda \cdot \theta}{N} \cdot w_i(t-\Delta) + \sum_{j=1}^{N} \frac{\lambda \cdot \theta}{N^2} \cdot w_j(t-\Delta) - \mu \cdot w_i(t) \tag{18}$$

This can be written as a matrix system by

$$\overset{\bullet}{w}(t) = -\frac{\lambda \cdot \theta}{N} \cdot \mathcal{I} w(t-\Delta) + \frac{\lambda \cdot \theta}{N^2} \cdot \mathcal{A} w(t-\Delta) - \mu \cdot \mathcal{I} w(t) \tag{19}$$

where $\mathcal{I}$ is an N dimensional identity matrix and $\mathcal{A}$ is a N dimensional square matrix of ones i.e.

$$\mathcal{A} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}.$$

With the representation of our linearized system in Equation 19, we can now exploit the fact that both $\mathcal{A}$ and $\mathcal{I}$ can be simultaneously diagonalized. Thus, we can write both $\mathcal{A}$ and $\mathcal{I}$ in terms of the eigenvectors of the matrix $\mathcal{A}$. If we denote S as the orthgonal matrix of the eigenvectors of $\mathcal{A}$ and denote Λ as diagonal matrix of the eigenvalues of $\mathcal{A}$, then we have that $\mathcal{A}$ and $\mathcal{I}$ can both be decomposed in terms of S, S^{-1}, and Λ as

$$\mathcal{A} = S\Lambda S^{-1} \tag{20}$$
$$\mathcal{I} = S\mathcal{I}S^{-1}. \tag{21}$$

The matrix $\mathcal{A}$ has rank 1 and therefore only has one non-zero eigenvalue. The only non-zero eigenvalue is equal to N and all other eigenvalues are equal to zero. The eigenvector corresponding to the eigenvalue N is given by $(1,1,,1)^T$. Moreover, the following eigenvectors $(1,-1,0,.....,0)^T, (1,0,-1,0,....,0)^T, ..., (1,0,....,0,....,-1)^T$ have an eigenvalue whose value is equal to zero. Using this knowledge of the matrix $\mathcal{A}$, we now we define, $v = S^{-1}w$ or $w = Sv$ and this leads us to the following delay differential system for v

$$\overset{\bullet}{w}(t) = S\overset{\bullet}{v}(t) = -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}w(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot\mathcal{A}w(t-\Delta) - \mu\cdot\mathcal{I}w(t) \tag{22}$$
$$= -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}w(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S\Lambda S^{-1}w(t-\Delta) - \mu\cdot\mathcal{I}w(t) \tag{23}$$
$$= -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}Sv(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S\Lambda S^{-1}Sv(t-\Delta) - \mu\cdot\mathcal{I}Sv(t) \tag{24}$$
$$= -\frac{\lambda\cdot\theta}{N}\cdot Sv(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S\Lambda v(t-\Delta) - \mu\cdot Sv(t) \tag{25}$$

Now by multiplying both sides by S^{-1} we have the following delay differential system for v

$$\overset{\bullet}{v}(t) = -\frac{\lambda\cdot\theta}{N}\cdot S^{-1}Sv(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S^{-1}S\Lambda v(t-\Delta) - \mu\cdot S^{-1}Sv(t) \tag{26}$$
$$= -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}v(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot\mathcal{I}\Lambda v(t-\Delta) - \mu\cdot\mathcal{I}v(t). \tag{27}$$

Thus, for the i^{th} entry of the vector v, we have the following delay differential equation

$$\overset{\bullet}{v}_i(t) = -\frac{\lambda\cdot\theta}{N}\cdot v_i(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot\Lambda_{ii}\cdot v_i(t-\Delta) - \mu\cdot v_i(t), \quad i\in 1,...N \tag{28}$$

where Λ_{ii} is the i^{th} diagonal entry of the matrix Λ. One crucial observation is that this representation shows that system of delay equations given in Equation 28 are uncoupled and can be analyzed **separately** for stability purposes. In fact, since the matrix $\mathcal{A}$ has two distinct eigenvalues N and 0, the stability of the system of our delay equations reduces to analyzing the following two delay differential equations

$$\overset{\bullet}{v}_1(t) = -\frac{\lambda\cdot\theta}{N}\cdot v_1(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot N\cdot v_1(t-\Delta) - \mu\cdot v_1(t) \tag{29}$$
$$\overset{\bullet}{v}_2(t) = -\frac{\lambda\cdot\theta}{N}\cdot v_2(t-\Delta) - \mu\cdot v_2(t). \tag{30}$$

Reducing these delay differential equations further, we have that

$$\overset{\bullet}{v}_1(t) = -\mu \cdot v_1(t) \tag{31}$$
$$\overset{\bullet}{v}_2(t) = -\frac{\lambda \cdot \theta}{N} \cdot v_2(t-\Delta) - \mu \cdot v_2(t). \tag{32}$$

To finish the proof, we observe that $v_1(t)$ is stable since μ is assumed to be positive. Therefore, it only remains to analyze the stability of the second equation for $v_2(t)$. To do this we make the ansatz $v_2(t) = e^{rt}$ and derive an equation for the variable r. This yields the following transcendental equations for r

$$r = -\frac{\lambda \cdot \theta}{N} \cdot e^{-r\Delta} - \mu. \tag{33}$$

Note that this is the real difference between ordinary differential equations and delay differential equations. These types of transcendental equations do not appear in ordinary differential equations because Δ is typically equal to zero in the ordinary differential equation context. Now we complete the proof by analyzing our transcendental equation for r. If we substitute $r = i\omega$, we obtain two equations for the real and imaginary parts respectively using Euler's identity

$$\cos(\omega\Delta) = -\frac{N \cdot \mu}{\lambda\theta} \tag{34}$$
$$\sin(\omega\Delta) = \frac{N \cdot \omega}{\lambda\theta}. \tag{35}$$

Now by squaring both sides and adding the two equations together we arrive at the following equation

$$\cos^2(\omega\Delta) + \sin^2(\omega\Delta) = 1 = \frac{N^2 \cdot (\mu^2 + \omega^2)}{\lambda^2\theta^2} \tag{36}$$

By moving all terms of Equation 36 that do not involve ω to the right, we can isolate an expression for ω. Thus, solving for ω, we arrive at the following expression

$$\omega = \frac{1}{N}\sqrt{\lambda^2\theta^2 - N^2 \cdot \mu^2}. \tag{37}$$

Using this expression for ω, we can finally invert Equation 34 since it does not contain ω on the right hand side unlike Equation 35 to solve for the critical value of Δ. We find that our threshold Δ is equal to

$$\Delta_{cr}(\lambda, \mu, , \theta, N) = \frac{N \cdot \arccos\left(\frac{-\mu \cdot N}{\lambda\theta}\right)}{\sqrt{\lambda^2\theta^2 - N^2 \cdot \mu^2}}. \tag{38}$$

Thus our proof is complete. ■

Theorem 3 provides a complete local characterization of the oscillation behavior of an arbitrary queueing system with N queues. If the delay Δ is larger than the critical delay $\Delta_{cr}(\lambda, \mu, N)$, then we should expect that the N queues should oscillate in equilibrium. However, if the delay Δ is smaller than the critical delay $\Delta_{cr}(\lambda, \mu, N)$, then we should expect that the N queues should converge to the limit $\frac{\lambda}{\mu N}$ and not oscillate around the equilibrium. In Figures 3 - 4 we plot the critical threshold as a function of λ and N. From observation, it is clear that as N is increased, the critical delay is also increased, which means that the region of stability becomes larger. We prove that this phenomenon is true in the following proposition.

Proposition 1. *For all $N \geq 2$ and $N+1 < \frac{\lambda\theta}{\mu}$, we have that*

$$\Delta_{cr}(\lambda,\mu,\theta,N) \leq \Delta_{cr}(\lambda,\mu,\theta,N+1). \tag{39}$$

Proof: *Take the derivative of $\Delta_{cr}(\lambda,\mu,N)$ or*

$$\frac{N \cdot \arccos\left(\frac{-\mu \cdot N}{\lambda\theta}\right)}{\sqrt{\lambda^2\theta^2 - N^2 \cdot \mu^2}} \tag{40}$$

with respect to N. For the values of N in the assumed region, the derivative is given by

$$\frac{\arccos\left(\frac{-\mu \cdot N}{\lambda\theta}\right)}{\sqrt{\lambda^2\theta^2 - N^2 \cdot \mu^2}} + \frac{N\mu}{\lambda^2\theta^2 - N^2 \cdot \mu^2} + \frac{N^2 \cdot \mu^2 \cdot \arccos\left(\frac{-\mu \cdot N}{\lambda\theta}\right)}{\sqrt{\lambda^2\theta^2 - N^2 \cdot \mu^2}^3}. \tag{41}$$

This quantity is positive in our assumed region and therefore it suggests the stability region gets larger as we increase N and all other parameters remain fixed. This also proves our claim. ■

Moreover, in Figure 5, we plot the critical delay value as a function of λ and μ. From this plot, we observe that the critical delay value appears to be monotonically decreasing as λ increases and monotonically increasing as μ is increased. This makes sense since increasing both parameters have an opposite affect on the queue length behavior; increasing λ increases the queue length, while increasing μ decreases the queue length. To further illustrate our results, in the sequel we compare our analytical result given in Theorem 1 with a numerical integration of the delay differential equations and a simulation of the stochastic queueing process.

2.3. Numerical Results for Fluid Limits

In this section, we describe some numerical results that compare the scaled stochastic queue length processes with their delay differential equation counterparts. At first they were surprising, however, after further inspection, we noticed a new phenomenon that we have not observed in the queueing literature before where the scaling needed for convergence is a true function of the time interval of observation. However, before describing our results, we describe via references how we perform the simulations of our delayed information queues. Work by Bratsun et al. [8] considers methods based on Gillespie's direct method Gillespie [13] or the next jump method of Gibson and Bruck [12]. The reader is encouraged to read the simulation section of Anderson and Kurtz [3] or the appendix of [24] for more details of the method.

In Figure 6, we plot the case of when N=2, $\eta = 10$ and $\Delta = .25$. On the left of Figure 6, we compare the simulated first queue with its fluid limit and on the right of Figure 6, we compare the second queue with its fluid limit. In both plots, we observe that the fluid limit approximates the mean dynamics quite well. Since we have that $\Delta = .25 < \Delta_{cr} = .3614$, we should expect that the two queues should synchronize and it is apparent from Figure 6 that they do exactly that.

In Figure 7, we plot the same queue length process, however, this time we make $\Delta = .45 > \Delta_{cr} = .3614$. Unlike Figure 6, we see in Figure 7 that the delay differential equation does not seem to approximate the mean stochastic dynamics well at all. However, we do observe in Figure 7 that the fluid limit and the mean of the stochastic queueing model are matching quite well until $t = 3$. Initially, this seems like that the limit theorem is wrong and it does not predict the right behavior of the stochastic model. However, as we will see later, the scaling parameter η needed to show proper convergence actually depends heavily on the time interval being considered. In other words, if one wants to show convergence on $[0,T_1]$ and $[0,T_2]$ where $T_1 < T_2$, one will need to choose a larger η value for T_2 given that one wants the same value of accuracy.

In Figure 8, we explore the Hopf bifurcation dynamics. We see that both queues are not synchronized and oscillate. However, on the left of Figure 8, which models the mean of the stochastic

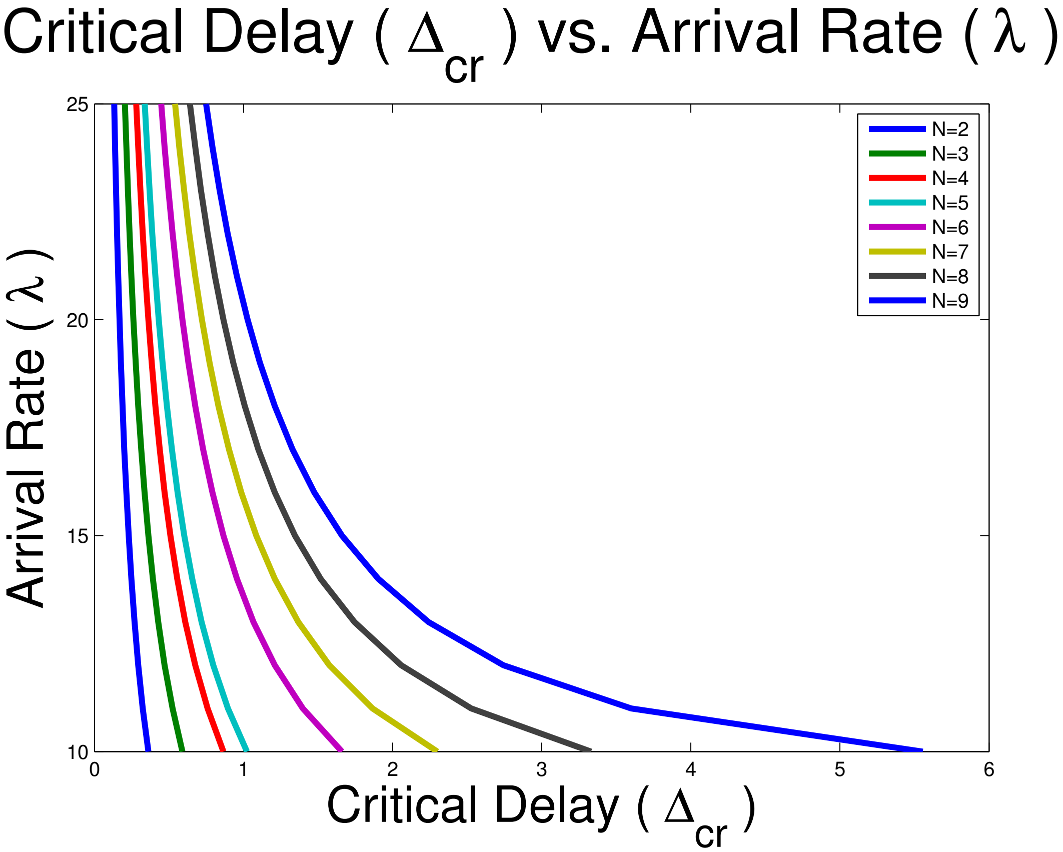


FIGURE 3. Plot of the critical threshold as a function of λ and N. $\lambda \in [10, 25]$, $\mu = 1$.

system with finite η, the oscillations are damped and on the right of Figure 8 the oscillations are not damped and remain for all time. This decaying of the oscillations in the stochastic model highlights the difference between steady state dynamics where $T = \infty$ and compact sets of time. It also highlights the difference between finite η and when $\eta = \infty$. To explore these concepts further, in Figure 9 we scale up η by a factor of 10 and keep all of the other parameters identical. Unlike Figure 7, Figure 9 actually shows convergence to the fluid limit on a larger time interval. Thus, this shows that our results really only hold for compact sets of time since it is clear that as you let time go towards infinity, the system of delay differential equations and the simulated mean of the stochastic queueing model will not match for any fixed value of η. What our numerical results also show is that suppose one would like the supremum of the absolute value of the simulated process to differ from the fluid limit by a constant $\varepsilon = .05$ i.e. $\sup_{t \le T} |Q_i^\eta(t) - q_i(t)| < \varepsilon = .05$, then Figure 7 demonstrates that $\eta = 10$ is enough on the time interval $[0, 3]$, but one will need a higher value of η for larger time intervals. However, Figure 9 suggests that $\eta = 100$ is enough on the time interval $[0, 12]$, but one will need a higher value of η for larger time intervals. Thus, in order to achieve a constant accuracy for longer periods of time, we need to consider larger and larger values of η.

3. Diffusion Limits Proving the fluid limit for our queueing model with delays allows us to gain knowledge about the average sample path dynamics of the queue length process. Since this limiting object is deterministic, it does not provide us with any insight on the flucuations of the queue length process around the mean. In order to study the flucuations about the mean, we need to study the diffusion limit. By exploiting the fluid limit, we can center the queue length process by the fluid limit and rescale to prove the diffusion limit. Like in the fluid limit, the initial condition

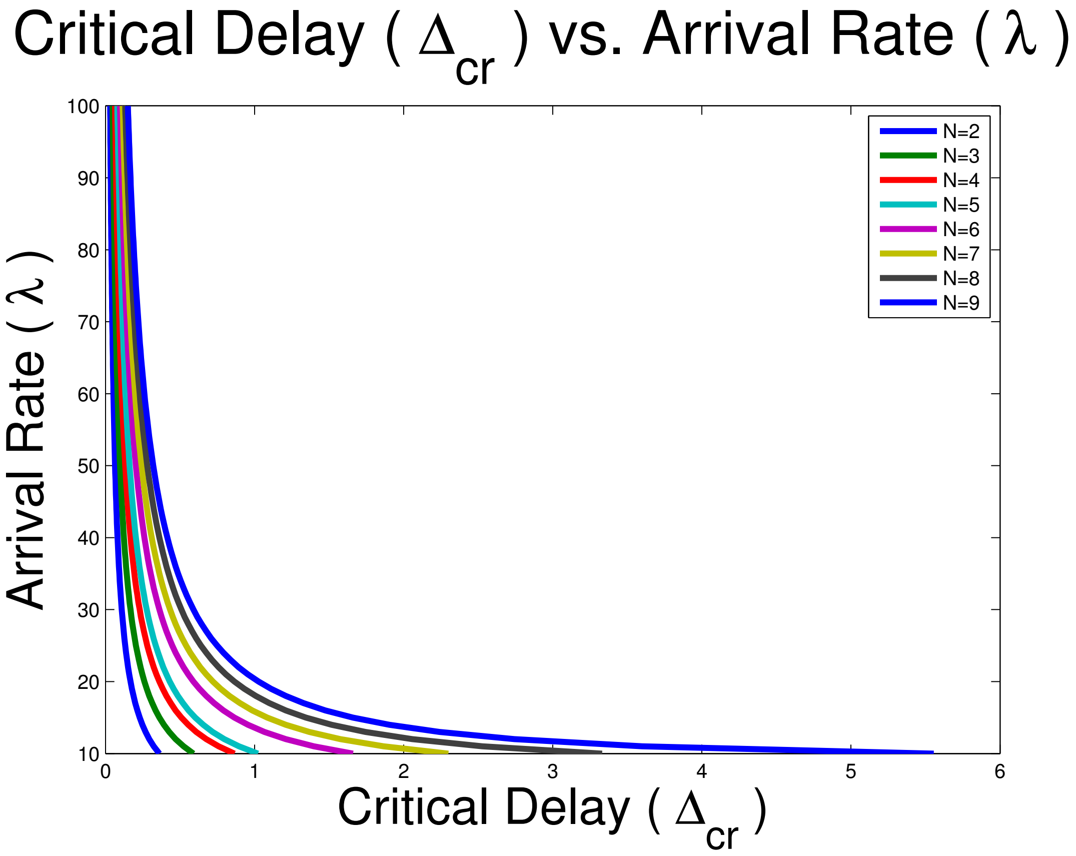


FIGURE 4. Plot of the critical threshold as a function of λ and N. $\lambda \in [10,100]$, $\mu = 1$.

is no longer a single point, but instead is a function over the time interval $[-\Delta, 0]$. Thus, we must take care of this issue by defining the appropriate Banach spaces and operators for our diffusion limit. Moreover, we also need to establish the existence and uniqueness of our stochastic differential equation with delay as these results are much less common and we would like to keep the paper self-contained. This is given by our next theorem.

THEOREM 4. *There exists an almost surely unique pathwise solution* $(\tilde{D}(t) = (\tilde{D}_1(t), \tilde{D}_2(t), ..., \tilde{D}_N(t)))$ *to the stochastic delay integral equations*

$$\tilde{D}_i(t) = \int_0^t \lambda \cdot \theta \cdot \sum_{j \neq i}^{N} \frac{\exp(-\theta(q_i(u-\Delta) + q_j(u-\Delta)))}{\left(\sum_{k=1}^N \exp(-\theta q_k(u-\Delta))\right)^2} \cdot \tilde{D}_j(u) du - \int_0^t \mu \cdot \tilde{D}_i(u) du \tag{42}$$

$$- \int_0^t \lambda \cdot \theta \cdot \frac{\sum_{j \neq i}^{N} \exp(-\theta(q_i(u-\Delta) + q_j(u-\Delta)))}{\left(\sum_{k=1}^N \exp(-\theta q_k(u-\Delta))\right)^2} \cdot \tilde{D}_i(u) du + V_i(t) \tag{43}$$

where we assume that $\tilde{D}_i(t) = 0$ *for all* $t \in [-\Delta, 0]$,

$$V_i(t) = \mathcal{B}_i^a \left(\int_0^t \frac{\lambda \cdot \exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))} ds \right) + \mathcal{B}_i^d \left(\int_0^t \mu \cdot q_i(s) ds \right), \tag{44}$$

and $\mathcal{B}_i^d, \mathcal{B}_i^a$ *are mutually independent standard Brownian motions.* ***Proof:*** *To prove the existence of a global solution, we must show two results. First we must show the existence and uniqueness of*

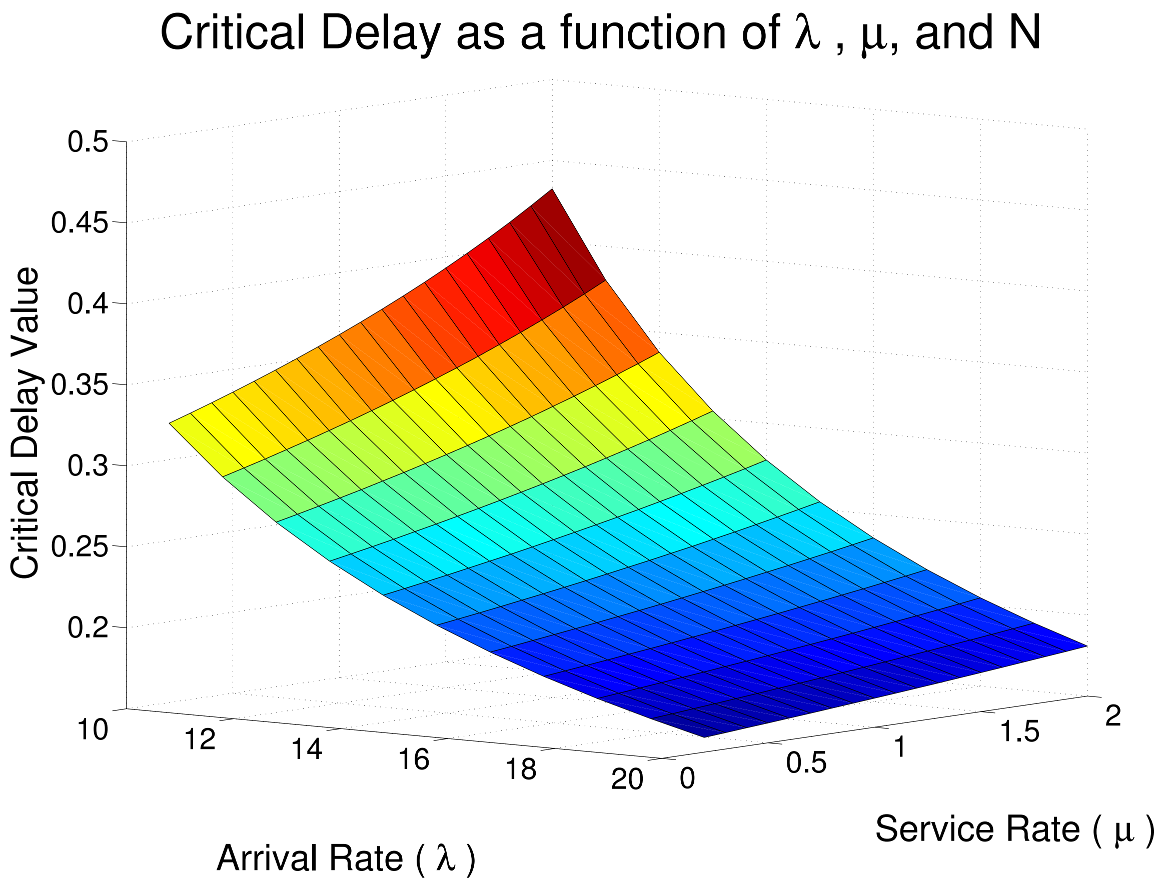


FIGURE 5. Plot of the critical threshold as a function of λ, μ, and N=2. $\lambda \in [10, 20]$, $\mu \in [.2, 2]$.

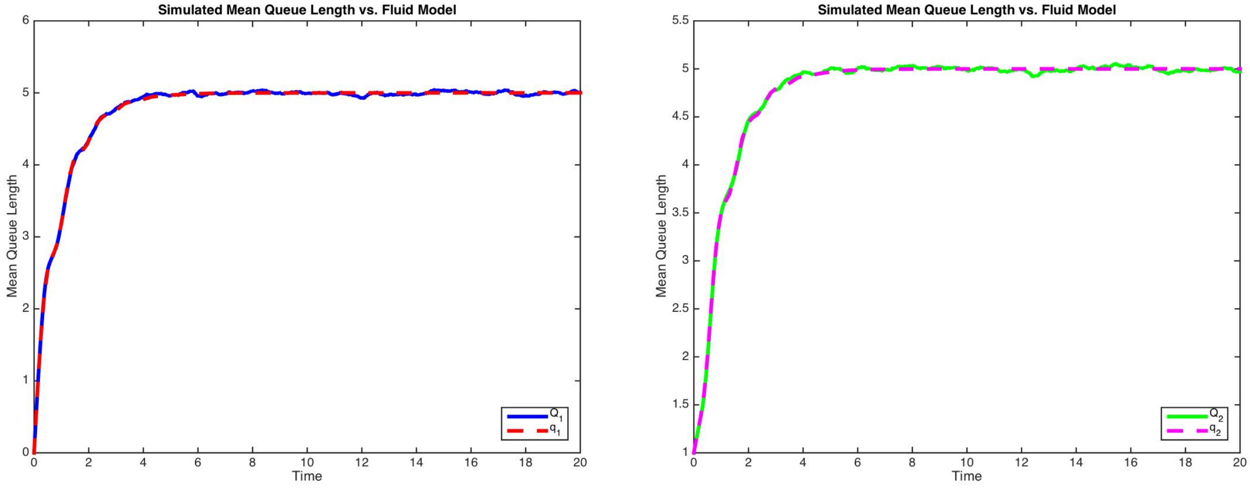


FIGURE 6. $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .25$, $\boldsymbol{\eta} = \mathbf{10}$. First Queue (Left) Second Queue (Right).

a local solution on a time interval $[0, \delta]$ for a sufficiently small $\delta > 0$. After establishing existence and uniqueness of a local solution, we show in a second step that the local solution can be extended

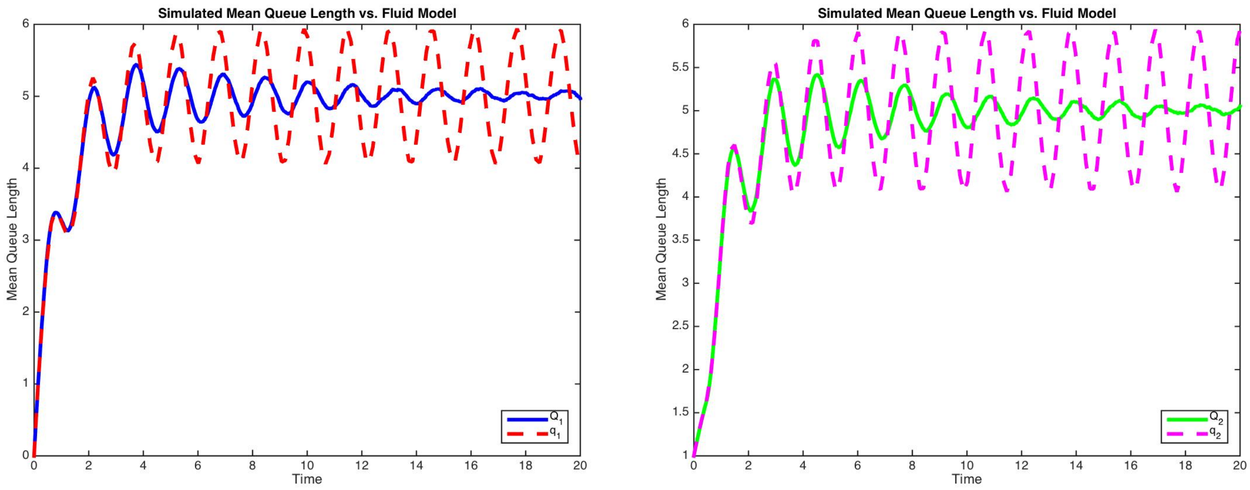


FIGURE 7. $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, $\boldsymbol{\eta} = \mathbf{10}$. First Queue (Left) Second Queue (Right).

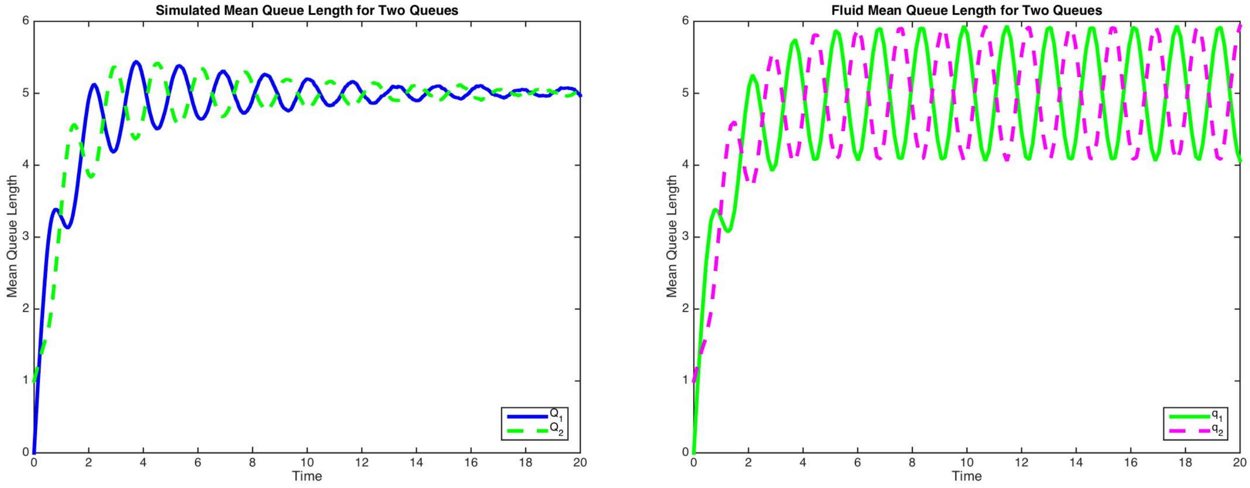


FIGURE 8. $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, $\boldsymbol{\eta} = \mathbf{10}$. Stochastic Simulation (Left) Fluid Limits (Right).

to a solution on $[0,T]$ where T is bounded. We should point out that this second step is similar to the method of steps in the delay differential equation context.

We begin by showing existence and uniqueness on a sufficiently small time interval. To show the existence and uniqueness on a short time interval, we begin by defining $\mathcal{C}_T$ as the Banach space of all continuous N-dimensional functions on the interval $[-\Delta, 0]$. We equip the Banach space $\mathcal{C}_0$ with the standard sup-norm $||\cdot||_\infty$. In addition, we also need to define the continuous initial function $\varphi(s) = (\varphi_1(s), \varphi_2(s), ..., \varphi_N(s))$ where $\varphi_i : [-\Delta, 0] \to \mathbb{R}$ and an almost surely continuous sample path $\mathcal{Z}_t(\omega)_{t\geq 0}$. The continuous initial function φ_i highlights one of the major differences between delay equations and their non-delayed counterparts. A function is needed in the delay differential equation setting while only a point is needed in the ordinary differential equation setting. Finally, we define the following two mappings $\Gamma_i(t)$ and $\mathcal{H}_i(t,z)$ for $1 \leq i \leq N$ as

$$\gamma_i(t) = \begin{cases} \varphi_i(t), \quad t \in [-\Delta, 0] \\ \varphi_i(0) + \mathcal{Z}_t(\omega), \quad t \in [0, \delta] \end{cases} \tag{45}$$

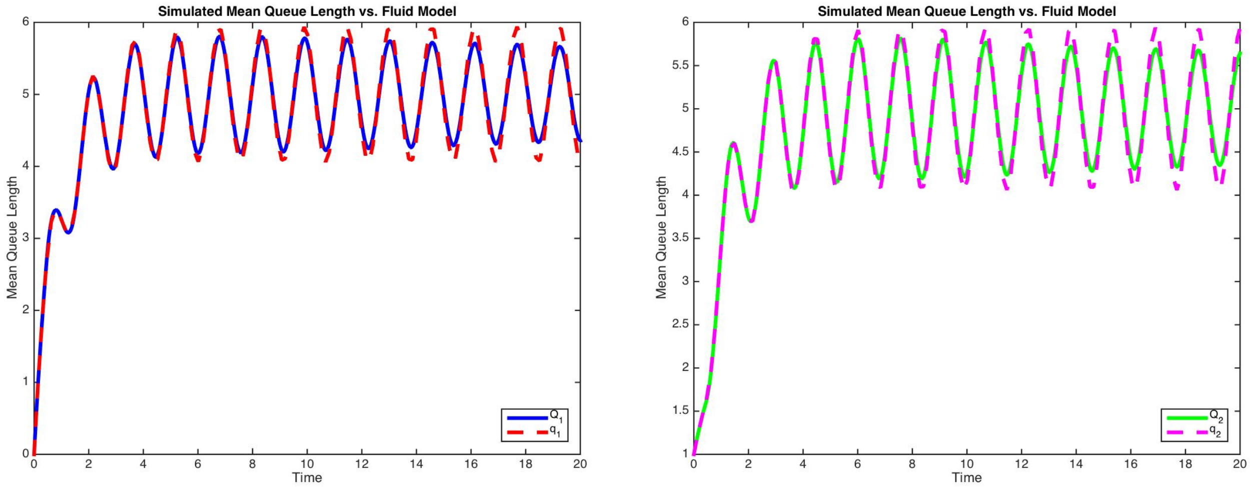


FIGURE 9. $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, $\boldsymbol{\eta} = \mathbf{100}$. First Queue (Left) Second Queue (Right).

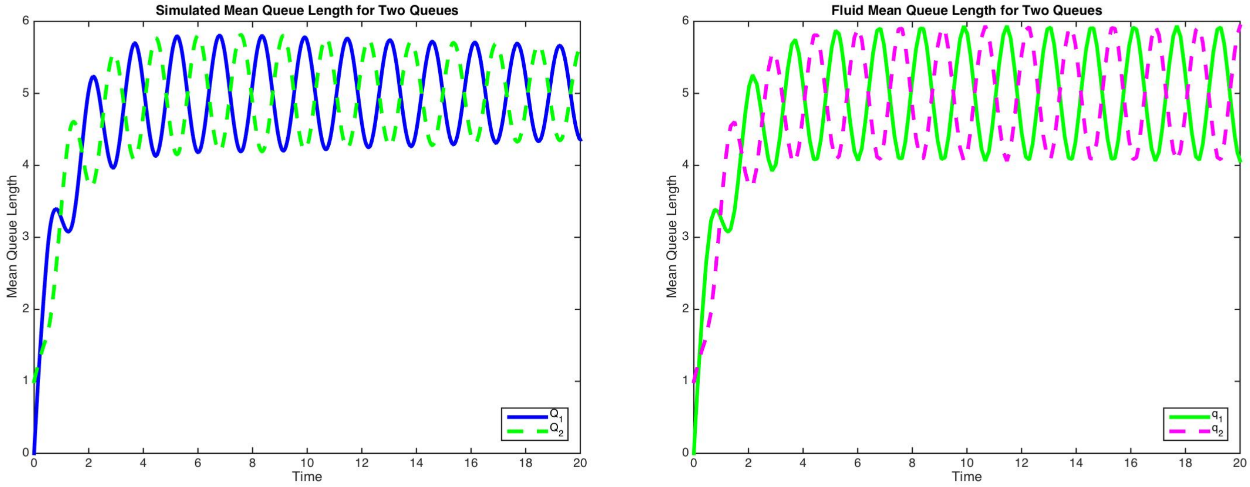


FIGURE 10. $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, $\boldsymbol{\eta} = \mathbf{100}$. First Queue (Left) Second Queue (Right).

and

$$\begin{aligned}\mathcal{H}_i(t,z) = \mathcal{H}_i(t,z_1,...,z_N) \;=\; & \lambda \cdot \theta \cdot \sum_{j \neq i}^{N} \frac{\exp(-\theta(q_i(t-\Delta)+q_j(t-\Delta)))}{\left(\sum_{k=1}^{N}\exp(-\theta q_k(t-\Delta))\right)^2} \cdot z_j(t) - \mu \cdot z_i(t) \\ & -\lambda \cdot \theta \cdot \frac{\sum_{j\neq i}^{N}\exp(-\theta(q_i(t-\Delta)+q_j(t-\Delta)))}{\left(\sum_{k=1}^{N}\exp(-\theta q_k(t-\Delta))\right)^2} \cdot z_i(t). \end{aligned} \tag{46}$$

Now we exploit the properties of the arrival and service process of our queueing system. In this case, we exploit the boundedness and continuity property of the derivative of the arrival and service rate functions. Since the partial derivatives of the rate functions are bounded and continuous, we know that the map $t \to \mathcal{H}(\cdot, \phi)$ *is almost surely continuous and bounded. This implies that the sup-norm of* $\mathcal{H}$ *is bounded i.e.* $||\mathcal{H}(\cdot,\gamma)||_\infty \leq M$. *However, the bound* M *is a random bound that can* *depend on the sample path of $\mathcal{Z}_t(\omega)$. Let us now fix a positive constant ζ. For a given $\delta > 0$ we construct a closed subset of the Banach space $\mathcal{C}_T$ by*

$$\mathcal{J}_\delta = \{\psi \in \mathcal{C}_\delta : ||\psi - \gamma||_\infty \leq \zeta \text{ and } \psi = \varphi \text{ on } [-\Delta, 0]\}. \tag{47}$$

This implies the following bound on the mapping $\mathcal{H}(t,z)$

$$\begin{aligned} |\mathcal{H}(t,z)| &= |\mathcal{H}(t,z) - \mathcal{H}(t,\gamma) + \mathcal{H}(t,\gamma)| && (48)\\ &\leq |\mathcal{H}(t,z) - \mathcal{H}(t,\gamma)| + |\mathcal{H}(t,\gamma)| && (49)\\ &\leq C \cdot ||z-\gamma||_\infty + M && (50)\\ &\leq C \cdot \zeta + M. && (51) \end{aligned}$$

Since the constants M and and ζ are not dependent on the parameter δ, the following operator

$$\mathcal{G}(z)(t) = \begin{cases} \varphi(t), & t \in [-\Delta, 0] \\ \varphi(0) + \int_0^t \mathcal{H}(u,z)du + \mathcal{Z}_t(\omega), & t \in [0,T]. \end{cases} \tag{52}$$

maps the closed set $\mathcal{J}_\delta$ into itself when δ is small enough. Thus, this implies that we can bound the following difference

$$\begin{aligned} |\mathcal{G}(z)(t) - \mathcal{G}(y)(t)| &\leq \int_0^t |\mathcal{H}(u,z) - \mathcal{H}(u,y)| du && (53)\\ &\leq C \cdot \delta \max_{-\Delta \leq u \leq \delta} |z(u) - y(u)| && (54) \end{aligned}$$

and derive a bound for the maximum of the difference of the two operators as

$$\max_{-\Delta \leq u \leq \delta} |\mathcal{G}_i(z)(u) - \mathcal{G}_i(y)(u)| \leq C \cdot \delta \max_{-\Delta \leq u \leq \delta} |z(u) - y(u)|. \tag{55}$$

Thus, for almost every sample path of $\mathcal{Z}_t(\omega)$, we have the existence of a δ small enough such that the operator $\mathcal{G} : \mathcal{J}_\delta \to \mathcal{J}_\delta$ is contraction. Consequently, by the contraction mapping principle or Banach's fixed point theorem Bharucha-Reid et al. [6], we have that the operator $\mathcal{G}$ has a unique fixed point. Thus, we have shown that our stochastic delay differential equation has a unique solution on the interval $[0,\delta]$. Now we need to join together several of these intervals together to build a solution on the compact set [0,T]. This is quite standard and to do this, we follow the same procedure as given in Ge and Zhu [11].

To extend the solution to the entire interval $[0,T]$, we assume that $[0,\delta]$, $[\delta, 2\delta]$, ... , $[k\delta, T]$ are subsets of $[0,T]$ with $k\delta < T < (k+1)\delta$. It follows from the above analysis that we can construct a unique solution $Q^i(t)$ on the interval $[i\delta,(i+1)\delta]$, which implies that we can construct a unique solution on the interval $[0,T]$ by setting

$$Q(t) = \begin{cases} Q^1(t), & t \in [0,\delta] \\ Q^2(t), & t \in [\delta, 2\delta] \\ \dots\ , & \\ Q^k(t), & t \in [(k-1)\delta, k\delta] \\ Q^{k+1}(t), & t \in [k\delta, T]. \end{cases} \tag{56}$$

Thus, our proof is complete. ■

Now that we proved that the stochastic differential delay equation has a unique solution on the interval $[0,T]$, we are ready to prove that the centered and rescaled queue length process $\tilde{D}^{\eta}_i(t)$ given by

$$\tilde{D}^{\eta}(t) = \sqrt{\eta}\cdot\left(\tilde{Q}^{\eta}(t) - q(t)\right) \tag{57}$$

converges to a stochastic delay differential equation that exists and has a unique solution and where the convergence is in the space $\mathbb{D}_T$ of functions that are right continuous and with left limits on $[0,T]$, equipped with the Skorokhod J_1 topology. The following theorem provides the proof of this convergence result.

Theorem 5. *The sequence of stochastic processes* $\{\tilde{D}^{\eta}(t) = (\tilde{D}^{\eta}_1(t), \tilde{D}^{\eta}_2(t), ..., \tilde{D}^{\eta}_N(t))\}_{\eta\in\mathbb{N}}$ *converges in distribution to the stochastic delay integral equations* $(\tilde{D}(t) = (\tilde{D}_1(t), \tilde{D}_2(t), ..., \tilde{D}_N(t))$ *where*

$$\tilde{D}_i(t) = \int_0^t \lambda\cdot\theta\cdot\sum_{j\neq i}^{N}\frac{\exp(-\theta(q_i(u-\Delta)+q_j(u-\Delta)))}{\left(\sum_{k=1}^{N}\exp(-\theta q_k(u-\Delta))\right)^2}\cdot\tilde{D}_j(u)du - \int_0^t \mu\cdot\tilde{D}_i(u)du \tag{58}$$

$$-\int_0^t \lambda\cdot\theta\cdot\frac{\sum_{j\neq i}^{N}\exp(-\theta(q_i(u-\Delta)+q_j(u-\Delta)))}{\left(\sum_{k=1}^{N}\exp(-\theta q_k(u-\Delta))\right)^2}\cdot\tilde{D}_i(u)du + V_i(t) \tag{59}$$

and $\tilde{D}_i(s) = 0$ *for all* $s\in[-\Delta,0]$ *and for all* $1\le i\le N$.

Proof: *See Appendix.* ■

4. Conclusion and Future Research

In this paper, we analyze a new N-dimensional stochastic queueing model that incorporates customer choice and delayed queue length information. Our model considers the customer choice as a multinomial logit model where the queue length information given to the customer is delayed by a constant Δ. For our model, we use strong approximations for Poisson processes to prove fluid and diffusion limit theorems. Our fluid and diffusion limits are different from the current literature in that it converges to a delay differential equation and the diffusion limit is a stochastic delay differential equation. For the fluid limit, which we determine is a delay differential equation, we derive a closed form expression for the critical delay threshold where below the threshold the all queues are balanced and converge to the equilibrium $\lambda/(N\mu)$. However, when Δ is larger than the threshold, then all queues have asynchronous dynamics and the equilibrium point is unstable. It is important for businesses and managers to determine and know these thresholds since using delayed information can have such a large impact on the dynamics of the business. Even small delays can cause oscillations and it is of great importance for managers of these service systems to understand when oscillations can arise based on the arrival and service parameters.

Since our analysis is the first of its kind in the queueing literature, there are many extensions that are worthy of future study. One extension that we would like to explore is the impact of nonstationary arrival rates in the spirit of Engblom and Pender [10], Pender [25, 27, 26, 30], Pender and Massey [31], Pender et al. [33]. This is important not only because arrival rates of customers are not constant over time, but also because it is important to know how to distinguish and separate the impact of the time varying arrival rate from the impact of the delayed information given to the customer. The proof of the limit theorems for the nonstationary setting does not really change, however, the analysis for the stability of the delay equations is a challenging problem.

Other extensions include the use of different customer choice functions and incorporating customer preferences in the model, however, once again the main limitation is the bifurcation and stability analysis and not the limit theorems. With regard to customer preferences, this is non-trivial problem because the equilibrium solution is no longer a simple expression, but the solution to a transcendental equation. This presents new challenges for deriving analytical formulas that determine synchronous or asynchronous dynamics. Another major extension that is important is the analysis of other queueing models such as the Erlang-A model. This is not only complicated in the bifurcation analysis, but also it is complicated from the limit theorem perspective. Our results in this work heavily rely on the differentiability of the rate functions and new analysis would be needed to analyze models with non-differentiable rate functions like the Erlang-A. A detailed analysis of these extensions will provide a better understanding how the information that operations managers provide to their customers will affect the dynamics of these real world systems. We plan to explore these extensions in subsequent work.

Appendix Before we begin the proof, we present two lemmas that are vital to understanding and constructing the proof via strong approximation theory.

Lemma 1 **(Kurtz 1978)**. *A standard Poisson process $\{\Pi(t)\}_{t\geq 0}$ can be realized on the same probability space as a standard Brownian motion $\{W(t)\}_{t\geq 0}$ in such a way that the almost surely finite random variable*

$$Z \equiv \sup_{t\geq 0} \frac{|\Pi(t) - t - W(t)|}{\log(2 \vee t)}$$

has finite moment generating function in the neighborhood of the origin and in particular finite mean.

Lemma 2 **(Kurtz 1978)**. *For any standard Brownian motion $\{W(t)\}_{t\geq 0}$ and any $\epsilon > 0$, $n \in \mathbb{N}$, and $T > 0$*

$$\tilde{M} \equiv \sup_{u,v,\leq n\epsilon T} \frac{|W(u) - W(v)|}{\sqrt{|u-v|\left(1 + \log\left(n\epsilon T / |u-v|\right)\right)}} < \infty \quad a.s.$$

4.1. Proof of Fluid Limit In this section we prove Theorem 1, which shows the convergence of the scaled queueing process to our system of delay differential equations.

Proof of Theorem 1

$$\begin{aligned} Q_i^\eta(t) &= Q_i^\eta(0) + \frac{1}{\eta}\Pi_i^a\left(\eta \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds\right) \\ &- \frac{1}{\eta}\Pi_i^d\left(\eta \int_0^t \mu Q_i^\eta(s) ds\right) \end{aligned} \tag{60}$$

First we need to represent the difference of the scaled stochastic queue length minus the fluid limit. This is given by the following expressions

$$\begin{aligned} Q_i^\eta(t) - q_i(t) &= Q_i^\eta(0) - q_i(0) + \frac{1}{\eta}\Pi_i^a\left(\eta \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds\right) \\ &- \int_0^t \lambda \cdot \frac{\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))} ds \\ &- \frac{1}{\eta}\Pi_i^d\left(\eta \int_0^t \mu Q_i^\eta(s) ds\right) + \int_0^t \mu q_i(s) ds \\ &= Q^\eta(0) - q(0) \\ &+ \frac{1}{\eta}\Pi_i^a\left(\eta \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds\right) - \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds \end{aligned}$$

$$
\begin{aligned}
&\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds - \int_0^t \lambda\cdot\frac{\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds \\
&-\frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) + \int_0^t \mu Q_i^\eta(s)ds \\
&-\int_0^t \mu Q_i^\eta(s)ds + \int_0^t \mu q_i(s)ds.
\end{aligned}
$$

Now we have a representation of the queue length in terms of centered time changed Poisson processes and a deterministic part, we can now apply the strong approximations theory to the absolute value of the difference.

$$
\begin{aligned}
&|Q_i^\eta(t) - q_i(t)| \\
&\le \left| Q_i^\eta(0) - q_i(0) \right| \\
&\quad+\left| \frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right) - \int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds \right| \\
&\quad+\left| \int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds - \int_0^t \lambda\cdot\frac{\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds \right| \\
&\quad+\left| \frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) - \int_0^t \mu Q_i^\eta(s)ds \right| \\
&\quad+\left| \int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds \right|.
\end{aligned}
$$

By the Lemma 1, we have the following strong approximation representation of the queue length as

$$
\begin{aligned}
Q_i^\eta(t) &= \int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds + \frac{1}{\eta}\mathcal{B}_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right) \\
&\quad-\int_0^t \mu Q_i^\eta(s)ds - \frac{1}{\eta}\mathcal{B}_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) + \mathcal{O}\frac{\log\eta}{\eta} \qquad (61) \\
&= \int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds + \frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right) \\
&\quad-\int_0^t \mu Q_i^\eta(s)ds - \frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right) + \mathcal{O}\frac{\log\eta}{\eta}. \qquad (62)
\end{aligned}
$$

Using the strong approximation representation, we now have that the difference between the scaled queue length and the fluid limit is bounded by

$$
\begin{aligned}
&|Q_i^\eta(t) - q_i(t)| \\
&\le \left| Q_i^\eta(0) - q_i(0) \right| + \left| \frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right) \right| \\
&\quad+\left| \int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds - \int_0^t \lambda\cdot\frac{\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds \right| \\
&\quad+\left| \frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right) \right| + \left| \int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds \right| + \mathcal{O}\frac{\log\eta}{\eta}.
\end{aligned}
$$

Now it remains to show that

$$\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)\right| = 0 \tag{63}$$

and

$$\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| = 0 \tag{64}$$

For the first Brownian motion term we have that

$$\begin{aligned}\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)\right| &\le \lim_{\eta\to\infty}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\lambda\cdot T\right)\right| &(65)\\ &= \lim_{\eta\to\infty}\left|\mathcal{B}_i^a\left(\frac{1}{\eta}\cdot\lambda\cdot T\right)\right| &(66)\\ &= 0. &(67)\end{aligned}$$

For the second Brownian motion term we have that

$$\begin{aligned}\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| &\le \lim_{\eta\to\infty}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left((Q^\eta(0)+\lambda)\cdot\mu\cdot T\right)\right| &(68)\\ &= \lim_{\eta\to\infty}\left|\mathcal{B}_i^d\left(\frac{1}{\eta}\cdot(Q^\eta(0)+\lambda)\cdot\mu\cdot T\right)\right| &(69)\\ &= 0. &(70)\end{aligned}$$

Thus, for every $\epsilon > 0$ there exists an η^* such that for all $\eta \ge \eta^*$

$$\left|Q_i^\eta(0) - q_i(0)\right| \le \epsilon/4, \tag{71}$$

$$\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)\right| \le \epsilon/4, \tag{72}$$

$$\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| \le \epsilon/4, \tag{73}$$

and

$$\mathcal{O}\frac{\log\eta}{\eta} \le \epsilon/4 \tag{74}$$

so that we have

$$\begin{aligned}|Q_i^\eta(t) - q_i(t)| \le &\left|\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds - \int_0^t \lambda\cdot\frac{\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds\right|\\ &+\left|\int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds\right| + \epsilon &(75)\\ \le &\int_0^t \left|\frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} - \lambda\cdot\frac{\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}\right|ds\\ &+\int_0^t \left|\mu Q_i^\eta(s) - \mu q_i(s)\right|ds + \epsilon &(76)\end{aligned}$$

Now because the multi-nomial logit probability function and the linear departure function are differentiable functions with uniformly bounded first derivatives, there exists a constant C such that

$$\begin{aligned} |Q_i^\eta(t) - q_i(t)| &\le C \int_0^t \sup_{-\Delta \le r \le s} \left| Q_i^\eta(r) - q_i(r) \right| ds + \epsilon && (77) \\ &\le C \cdot \left(\int_0^t \sup_{0 \le r \le s} \left| Q_i^\eta(r) - q_i(r) \right| ds + t \cdot \sup_{-\Delta \le r \le 0} \left| Q_i^\eta(r) - q_i(r) \right| \right) + \epsilon. && (78) \end{aligned}$$

Now we exploit the fact that we assumed that $Q_i^\eta(t) = q_i(t)$ for $t \in [-\Delta, 0]$ for our initial condition. This assumption yields the following new bound for the difference of the scaled queue length and the fluid limit by

$$|Q_i^\eta(t) - q_i(t)| \le C \int_0^t \sup_{0 \le r \le s} \left| Q_i^\eta(r) - q_i(r) \right| ds + \epsilon. \tag{79}$$

Note that the difference between the two equations above is the interval of the supremum inside the integral. Now by invoking Gronwall's lemma in Hale [16], we have that

$$\sup_{0 \le t \le T} |Q_i^\eta(t) - q_i(t)| \le \epsilon \cdot e^{CT} \tag{80}$$

and since ϵ is arbitrary, we can let it go towards zero and this proves the fluid limit. ■

4.2. Proof of Diffusion Limit In this section we prove Theorem 5, which shows the convergence of the centers and rescaled queueing process to our system of stochastic delay differential equations.

Proof of Theorem 5

$$Q_i^\eta(t) = Q_i^\eta(0) + \frac{1}{\eta}\Pi_i^a\left(\eta \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds\right) - \frac{1}{\eta}\Pi_i^d\left(\eta \int_0^t \mu Q_i^\eta(s) ds\right) \tag{81}$$

First we need to represent the difference of the scaled stochastic queue length minus the fluid limit. This is given by the following expressions

$$\begin{aligned} \sqrt{\eta}\left(Q_i^\eta(t) - q_i(t)\right) &= \sqrt{\eta}\left(Q_i^\eta(0) - q_i(0)\right) + \sqrt{\eta} \cdot X_i^\eta(t) && (82) \\ &\quad + \sqrt{\eta} \cdot \int_0^t \left(F_i(s, Q^\eta(s-\Delta), Q^\eta(s)) - F_i(s, q(s-\Delta), q(s))\right) ds \end{aligned}$$

where

$$\begin{aligned} X_i^\eta(t) &= \frac{1}{\eta}\Pi_i^a\left(\eta \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds\right) - \int_0^t \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} ds \\ &\quad + \frac{1}{\eta}\Pi_i^d\left(\eta \int_0^t \mu Q_i^\eta(s) ds\right) - \int_0^t \mu Q_i^\eta(s) ds && (83) \end{aligned}$$

and

$$F_i(s, Q^\eta(s)) = \frac{\lambda \cdot \exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))} - \mu \cdot Q_i^\eta(s). \tag{84}$$

PROPOSITION 2. *Let $V_i^\eta(t)$ be defined by the following equation*

$$V_i^\eta(t) = \mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right) + \mathcal{B}_i^d\left(\int_0^t \mu\cdot Q_i^\eta(s)ds\right), \tag{85}$$

then

$$\lim_{\eta\to\infty}\sup_{0\le t\le T}|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)| = 0 \quad \text{in distribution.} \tag{86}$$

Proof: *We will show the result for one of the Brownian motion terms and one of the centered Poisson processes. The proof for the remaining terms will follow in a similar manner and are therefore omitted. Using the strong approximation result of Lemma 1, we obtain*

$$\sup_{t\ge0}\frac{1}{\sqrt{\eta}}\frac{\left|\overline{\Pi}_i^a\left(\eta\int_0^t\frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N\exp(-\theta Q_j^\eta(s-\Delta))}ds\right) - \mathcal{B}_i^a\left(\eta\int_0^t\frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N\exp(-\theta Q_j^\eta(s-\Delta))}ds\right)\right|}{\log\left(2\vee\eta\int_0^t\frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N\exp(-\theta Q_j^\eta(s-\Delta))}ds\right)} \le \frac{C_i^a}{\sqrt{\eta}} \tag{87}$$

where the distribution of C_i^a is independent η and $\overline{\Pi}_i^a$ is a centered Poisson process. Using Lemma 1 and the fact that the arrival rate function is bounded above by a constant K, then we have that

$$\sup_{0\le t\le T}|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)| \le \log(2\vee\eta KT)\sup_{0\le t\le T}\frac{\left|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)\right|}{\log(2\vee\eta Kt)} \tag{88}$$

$$\le \log(2\vee\eta KT)\frac{C_i^a}{\sqrt{\eta}} \tag{89}$$

Since the distribution of C_i^a is independent η and we have that

$$\lim_{\eta\to\infty}\frac{\log(2\vee\eta KT)}{\sqrt{\eta}} = 0,$$

it implies that as $\eta\to\infty$ we have that

$$\sup_{0\le t\le T}|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)| \Rightarrow 0 \quad \text{in distribution as } \eta\to\infty. \tag{90}$$

All the other terms can be proved similarly with the same technique. ■

PROPOSITION 3. *The sequence of stochastic processes $V_i^\eta(t)$ converges in distribution to the process $V_i(t)$ where*

$$V_i(t) = \mathcal{B}_i^a\left(\int_0^t\frac{\lambda\cdot\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N\exp(-\theta q_j(s-\Delta))}ds\right) + \mathcal{B}_i^d\left(\int_0^t\mu\cdot q_i(s)ds\right). \tag{91}$$

Proof: *In order to prove the convergence of the scaled Brownian motions, we will use Lemma 2. Moreover, we will provide the full proof for the arrival process for an arbitrary queue and the proofs for the remaining terms follow analagously. We now define a new function $\gamma_i^\eta(t)$ as follows.*

$$\gamma_i^\eta(t) \equiv \left|\int_0^t\frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N\exp(-\theta Q_j^\eta(s-\Delta))}ds - \int_0^t\frac{\lambda\cdot\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N\exp(-\theta q_j(s-\Delta))}ds\right|$$

and

$$\overline{\gamma}_i^\eta \equiv \sup_{0\le t\le T}\gamma_i^\eta(t).$$

This implies that

$$\begin{aligned}&\left|\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)-\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds\right)\right|\\ &=\frac{\left|\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)-\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds\right)\right|}{\sqrt{\gamma^\eta(t)\cdot(1+\log(KT/\gamma^\eta(t)))}}\cdot\sqrt{\gamma^\eta(t)\cdot(1+\log(KT/\gamma^\eta(t)))}.\end{aligned}$$

However, from Lemma 2, we obtain

$$\begin{aligned}&\left|\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)-\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds\right)\right|\\ &\le \tilde{M}_i^a\cdot\sqrt{\gamma_i^\eta(t)\cdot(1+\log(KT/\gamma_i^\eta(t)))}\end{aligned}$$

From the Lipschitz continuity of the rate functions, we have that

$$\overline{\gamma}_i^\eta \le KT\cdot\sup_{0\le t\le T}|Q_i^\eta(t)-q_i(t)|.$$

Therefore, by convergence of the fluid limit, we have that

$$\overline{\gamma}_i^\eta \Rightarrow 0.$$

By observing that the distribution of $\tilde{M}_i^a$ is independent of η and that the following limit

$$\lim_{\delta\to 0}\sqrt{\delta\cdot(1+\log(KT/\delta))}=0,$$

we conclude that

$$\tilde{M}_i^a\cdot\sqrt{\overline{\gamma}_i^\eta\cdot(1+\log(KT/\overline{\gamma}_i^\eta))}\Rightarrow 0$$

and therefore,

$$\lim_{\eta\to\infty}\sup_{0\le t\le T}\left|\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta Q_i^\eta(s-\Delta))}{\sum_{j=1}^N \exp(-\theta Q_j^\eta(s-\Delta))}ds\right)-\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp(-\theta q_i(s-\Delta))}{\sum_{j=1}^N \exp(-\theta q_j(s-\Delta))}ds\right)\right|\Rightarrow 0.$$

The remaining terms for other queues and the departures can be shown to converge by identical arguments and therefore, we do not provide their proofs. ■

The following lemma shows that the sequence $\tilde{D}_i^\eta(t)$ is stochastically bounded for all $1\le i\le N$.

Lemma 3. *For any $\epsilon>0$, there exists $\eta^*\in\mathbb{N}$ and $K<\infty$ such that*

$$\mathbb{P}\left(\sup_{0\le t\le T}|\tilde{D}_i^\eta(t)|>K\right)<\epsilon \quad \text{for all } \eta\ge\eta^*. \tag{92}$$

Proof: *The strong approximation for unit rate Poisson processes gives us the following representation for the centered and rescaled queue length process as*

$$\tilde{D}_i^\eta(t)=\sqrt{\eta}\int_0^t (F_i(s,Q^\eta(s))-F_i(s,q(s)))\,ds+V_i^\eta(t).$$

We know that each $V_i^\eta(t)$ is tight since it converges to a time changed Brownian motion, which is a continuous stochastic process. Therefore, the tightness of $V_i^\eta(t)$ implies that it is bounded in *probability, see for example Section 15 of Billingsley [7] or Section 3 of Whitt et al. [35]. Moreover, by using the Lipschitz continuity of the rate functions we have that*

$$\sup_{0\le t\le T} |\tilde{D}_i^\eta(t)| \le L\int_0^T \sup_{0\le t\le s} |\tilde{D}_i^\eta(s)|ds + \sup_{0\le t\le T} |V_i^\eta(t)|$$

for some Lipschitz constant L. Thus, by Gronwall's inequality in Problem 2.7 of Karatzas and Shreve [23] we have almost surely that

$$\sup_{0\le t\le T} \tilde{D}_i^\eta(t) \le e^{LT} \sup_{0\le t\le T} V_i^\eta(t)$$

and this concludes the proof. ■

Lemma 4. *If $\{f^\eta(t), \eta\in\mathbb{N}, t\in\mathbb{R}_+\}$ be a sequence of non-negative random processes such that*

$$\lim_{\eta\to\infty}\int_0^T f^\eta(s)ds = 0 \quad \text{in probability,} \tag{93}$$

then, for all $\delta > 0$,

$$\lim_{\eta\to\infty} \mathbb{P}\left(\sup_{0\le t\le T}\left|\int_0^t f^\eta(s)\tilde{D}_i^\eta(s)ds\right| > \delta\right) = 0. \tag{94}$$

Proof: *If we fix $\epsilon > 0$, then we know that there exists a constant $\eta^*\in\mathbb{N}$ such that for all $\eta > \eta^*$, there exists sets $\Omega_{\eta,1}$ and $\Omega_{\eta,2}$ such that*

$$\int_0^T f^\eta(s)ds < \epsilon/2 \qquad \text{on } \Omega_{\eta,1} \text{ and such that } \mathbb{P}(\Omega_{\eta,1}) \ge 1-\epsilon/2, \tag{95}$$

and

$$\sup_{0\le t\le T} |\tilde{D}_i^\eta(t)| < K \qquad \text{on } \Omega_{\eta,2} \text{ and such that } \mathbb{P}(\Omega_{\eta,2}) \ge 1-\epsilon/2, \tag{96}$$

Therefore, we have that

$$\sup_{0\le t\le T}\left|\int_0^t f^\eta(s)\tilde{D}_i^\eta(s)ds\right| \le \sup_{0\le t\le T}\left|\tilde{D}_i^\eta(t)\right|\int_0^T f^\eta(s)ds < K\epsilon \quad \text{on } \Omega_{\eta,1}\cap\Omega_{\eta,2}. \tag{97}$$

The result follows since ϵ was choosen arbitrarily. ■

For a function $g\in\mathcal{C}_T$ and the continuous initial function $\varphi: [-\Delta, 0]\to\mathbb{R}^N$, we define the operator $\mathcal{G}(g) = (\mathcal{G}^1(g), \mathcal{G}^2(g), ..., \mathcal{G}^N(g))$ be to the unique function that satisfies the following integral equation

$$\mathcal{G}_t^i(g) = \begin{cases} \varphi_i(t), & t\in[-\Delta, 0] \\ \int_0^t \langle\nabla F_i(s, q(s)), \mathcal{G}_s^i(g)\rangle ds + g_i(t), & t\in[0,T]\end{cases} \tag{98}$$

where ∇F_i is the gradient of F_i and $\langle\cdot,\cdot\rangle$ is the inner product of two vectors. Using this operator, it is obvious to see that $\mathcal{G}(V(t)) = \tilde{D}$, where $\tilde{D}$ is the stochastic delay differential equation defined in Equation 59. Since the arrival rate function and the service rate functions are continuously differentiable and the derivative is bounded, then we can show that $\mathcal{G}$ is a continuous operator using Gronwall's lemma of Karatzas and Shreve [23]. Moreover, we know that V^η converges to V in probability and this implies that

$$\lim_{\eta\to\infty} ||\mathcal{G}(V^\eta) - \tilde{D}||_\infty = 0. \tag{99}$$

Therefore, if we can show that the following difference

$$\lim_{\eta\to\infty} \sup_{0\le t\le T} ||\tilde{D}^\eta(t) - \mathcal{G}(V^\eta)(t)||_\infty = 0. \tag{100}$$

converges to zero in probability, then we will have completed our proof for the diffusion limit. To prove this, we define the difference between the two processes as

$$\begin{aligned}\mathcal{E}_i^\eta(t) &= \tilde{D}_i^\eta(t) - \mathcal{G}^i(V^\eta)(t)\\ &= \sqrt{\eta}\int_0^t \left(F_i(s,Q^\eta(s)) - F_i(s,q(s))\right) ds + V_i^\eta(t) - \left(\int_0^t \langle \nabla F_i(s,q(s)), \tilde{D}^\eta(s)\rangle ds + V_i^\eta(t)\right)\\ &= \sqrt{\eta}\int_0^t \left(F_i(s,Q^\eta(s)) - F_i(s,q(s))\right) ds - \int_0^t \langle \nabla F_i(s,q(s)), \tilde{D}^\eta(s)\rangle ds\\ &= \int_0^t \langle \nabla F_i(s,q(s)), \mathcal{E}^\eta(s)\rangle ds + \sqrt{\eta}\int_0^t \left(F_i(s,Q^\eta(s)) - F_i(s,q(s))\right) ds\\ &\quad - \int_0^t \langle \nabla F_i(s,q(s)), D^\eta(s)\rangle ds\end{aligned}$$

Thus, by the mean value theorem and the fact that the arrival rate and service rate functions are continuously differentiable, there exists a vector $\xi^\eta(s)$ that is in between $q(s)$ and $Q^\eta(s)$ such that

$$\begin{aligned}F_i\left(s,Q^\eta(s)\right) - F_i(s,q(s)) &= \langle \nabla F_i(s,\xi^\eta(s)), (Q^\eta(s) - q(s))\rangle\\ &= \langle \nabla F_i(s,\xi^\eta(s)), \frac{1}{\sqrt{\eta}}\cdot\sqrt{\eta}\left(Q^\eta(s) - q(s)\right)\rangle\\ &= \frac{1}{\sqrt{\eta}}\langle \nabla F_i(s,\xi^\eta(s)), D^\eta(s)\rangle.\end{aligned}$$

From this equivalence provided by the mean value theorem, it now implies that

$$\mathcal{E}^\eta(t) = \int_0^t \langle \left(\nabla F_i(s,\xi^\eta(s)) - \nabla F_i(s,q(s))\right), D^\eta(s)\rangle ds + \int_0^t \langle \nabla F_i(s,q(s)), \mathcal{E}^\eta(s)\rangle ds.$$

We also know that

$$\lim_{\eta\to\infty} \sup_{0\le t\le T} \|\nabla F_i(t,\xi^\eta(t)) - \nabla F_i(t,q(t))\| = 0 \quad a.s \tag{101}$$

in lieu of the fluid limit convergence and the continuity of the function $\partial F_i(\cdot,\xi^\eta(\cdot))$. Moreover, since $D^\eta(u)$ is bounded in probability and Lemma 4 is true, we have that the process

$$\lim_{\eta\to\infty} \sup_{0\le t\le T} \int_0^t \langle \left(\nabla F_i(s,\xi^\eta(s)) - \nabla F_i(s,q(s))\right), D^\eta(s)\rangle ds = 0 \quad \text{in probability}.$$

Finally by the application of Gronwall's inequality in Problem 2.7 of Karatzas and Shreve [23] and Lemma 4, we obtain our diffusion limit result since i was chosen arbitrarily.

■

References

[1] Gad Allon and Achal Bassamboo. The impact of delaying the delay announcements. *Operations research*, 59(5):1198–1210, 2011.

[2] Gad Allon, Achal Bassamboo, and Itai Gurvich. "we will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations research*, 59(6):1382–1394, 2011.

[3] David F Anderson and Thomas G Kurtz. Continuous time markov chain models for chemical reaction networks. In *Design and analysis of biomolecular circuits*, pages 3–42. Springer, 2011.

[4] Mor Armony and Constantinos Maglaras. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004.

[5] Mor Armony, Nahum Shimkin, and Ward Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81, 2009.

[6] AT Bharucha-Reid et al. Fixed point theorems in probabilistic analysis. *Bulletin of the American Mathematical Society*, 82(5):641–657, 1976.

[7] Patrick Billingsley. *Convergence of probability measures.* John Wiley &amp; Sons, 2013.

[8] Dmitri Bratsun, Dmitri Volfson, Lev S Tsimring, and Jeff Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14593–14598, 2005.

[9] Jing Dong, Elad Yom-Tov, and Galit B Yom-Tov. The impact of delay announcements on hospital network coordination and waiting times. Technical report, Working Paper, 2015.

[10] Stefan Engblom and Jamol Pender. Approximations for the moments of nonstationary and state dependent birth-death queues. 2014.

[11] Xintong Ge and Yuanguo Zhu. Existence and uniqueness theorem for uncertain delay differential equations. *Journal of Computational Information Systems*, 8(20):8341–8347, 2012.

[12] Michael A Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9):1876–1889, 2000.

[13] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[14] Pengfei Guo and Paul Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53(6):962–970, 2007.

[15] Pengfei Guo and Paul Zipkin. The impacts of customers'delay-risk sensitivities on a queue with balking. *Probability in the engineering and informational sciences*, 23(03):409–432, 2009.

[16] Jack K Hale. Ordinary differential equations. *Pure and Applied Mathematics*, 21, 1969.

[17] Jack K Hale. Functional differential equations. In *Analytic theory of differential equations*, pages 9–22. Springer, 1971.

[18] Refael Hassin. Information and uncertainty in a queuing system. *Probability in the Engineering and Informational Sciences*, 21(03):361–380, 2007.

[19] Rouba Ibrahim, Mor Armony, and Achal Bassamboo. Does the past predict the future? the case of delay announcements in service systems, 2015.

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

[21] Oualid Jouini, Yves Dallery, and Zeynep Akşin. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics*, 120(2):389–399, 2009.

[22] Oualid Jouini, Zeynep Aksin, and Yves Dallery. Call centers with delay information: Models and insights. *Manufacturing &amp; Service Operations Management*, 13(4):534–548, 2011.

[23] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science &amp; Business Media, 2012.

[24] William A. Massey and Jamol Pender. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75(2-4):243–277, February 2013.

[25] Jamol Pender. Gram charlier expansion for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics*, 74(4):1238–1265, 2014.

[26] Jamol Pender. Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Informational Sciences*, 29(01):27–49, 2015.

[27] Jamol Pender. An analysis of nonstationary coupled queues. *Telecommunication Systems*, pages 1–16, 2015.

[28] Jamol Pender. Heavy traffic limits for unobservable queues with clearing times. 2015.

[29] Jamol Pender. The impact of dependence on unobservable queues. 2015.

[30] Jamol Pender. The truncated normal distribution: Applications to queues with impatient customers. *Operations Research Letters*, 43(1):40–45, 2015.

[31] Jamol Pender and William A Massey. Approximating and stabilizing dynamic rate jackson networks with abandonment. *Probability in the Engineering and Informational Sciences*, 31(1):1–42, 2017.

[32] Jamol Pender, Richard H Rand, and Elizabeth Wesson. Managing information in queues: The impact of giving delayed information to customers. *arXiv preprint arXiv:1610.01972*, 2016.

[33] Jamol Pender, Richard H Rand, and Elizabeth Wesson. An asymptotic analysis of queues with delayed information and time varying arrival rates. *arXiv preprint arXiv:1701.05443*, 2017.

[34] Ward Whitt. Improving service by informing customers about anticipated delays. *Management science*, 45(2):192–207, 1999.

[35] Ward Whitt et al. Proofs of the martingale fclt. *Probab. Surv*, 4:268–302, 2007.

Acknowledgments.