Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Stochastic Analysis of Queues with Customer Choice and Delayed Information

Jamol Pender, Richard Rand, Elizabeth Wesson

To cite this article:
Jamol Pender, Richard Rand, Elizabeth Wesson (2020) A Stochastic Analysis of Queues with Customer Choice and Delayed Information. Mathematics of Operations Research 45(3):1104-1126. https://doi.org/10.1287/moor.2019.1024

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 45, No. 3, August 2020, pp. 1104–1126
ISSN 0364-765X (print), ISSN 1526-5471 (online)

A Stochastic Analysis of Queues with Customer Choice and Delayed Information

Jamol Pender,[a] Richard Rand,[b] Elizabeth Wesson[c]

[a] School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14850; [b] Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, New York 14850; [c] Center for Applied Mathematics, Cornell University, Ithaca, New York 14850

Contact: jjp274@cornell.edu, https://orcid.org/0000-0002-5418-6918 (JP); rand@math.cornell.edu (RR); enw27@cornell.edu (EW)

Received: February 1, 2017
Revised: December 1, 2017; August 13, 2018
Accepted: April 26, 2019
Published Online in Articles in Advance: April 10, 2020

MSC2000 Classification Codes: Primary: 60K25; secondary: 90B22 and 62L20
OR/MS Subject Classifications: Queues: Approximations; Organizational studies: Information; Queues: Diffusion models

https://doi.org/10.1287/moor.2019.1024

Abstract. 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 it is often not provided in real time. Recent work by Dong et al. [Dong J, Yom-Tov E, Yom-Tov GB (2018) The impact of delay announcements on hospital network coordination and waiting times. *Management Sci.* 65(5):1969–1994.] explores the impact of these delays in an empirical study in U.S. hospitals. Work by Pender et al. [Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *Internat. J. Bifurcation Chaos Appl. Sci. Engrg.* 27(4):1730016-1–1730016-20.] 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. [Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *Internat. J. Bifurcation Chaos Appl. Sci. Engrg.* 27(4):1730016-1–1730016-20.] 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.

Funding: This work was supported by the Division of Civil, Mechanical and Manufacturing Innovation [Grant 1751975]. The authors acknowledge the generous support of the National Science Foundation [Jamol Penders Career Award CMMI 1751975].

Keywords: 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 announcements 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 works by Allon and Bassamboo [1], Allon et al. [2], Armony and Maglaras [4], Armony et al. [5], Guo and Zipkin [15], Guo and Zipkin [16], Hassin [19], Ibrahim et al. [21], Jouini et al. [23, 24], and Whitt [37], and references therein focus on this aspect of the announcements. 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, because 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 bike-sharing network in New York City. Customers can look and see where there are bikes on an app. However, in the time that 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 [22] and Pender [30, 32] 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 works of Jennings and Pender [22] and Pender [30, 32] compare 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 4:04 p.m., whereas the time of a 12-minute wait is as of 3:44 p.m. 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, which we will show in the rest of the paper. Some empirical work by Dong et al. [10] has shown that delays in information can cause oscillations in wait times.

Another relevant application of our work is for amusement parks, like Disneyland or Six Flags. In Figure 2, 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 theme park. Customers obviously have the opportunity to choose which ride

Figure 1. (Color online) JFK Medical Center online reporting.

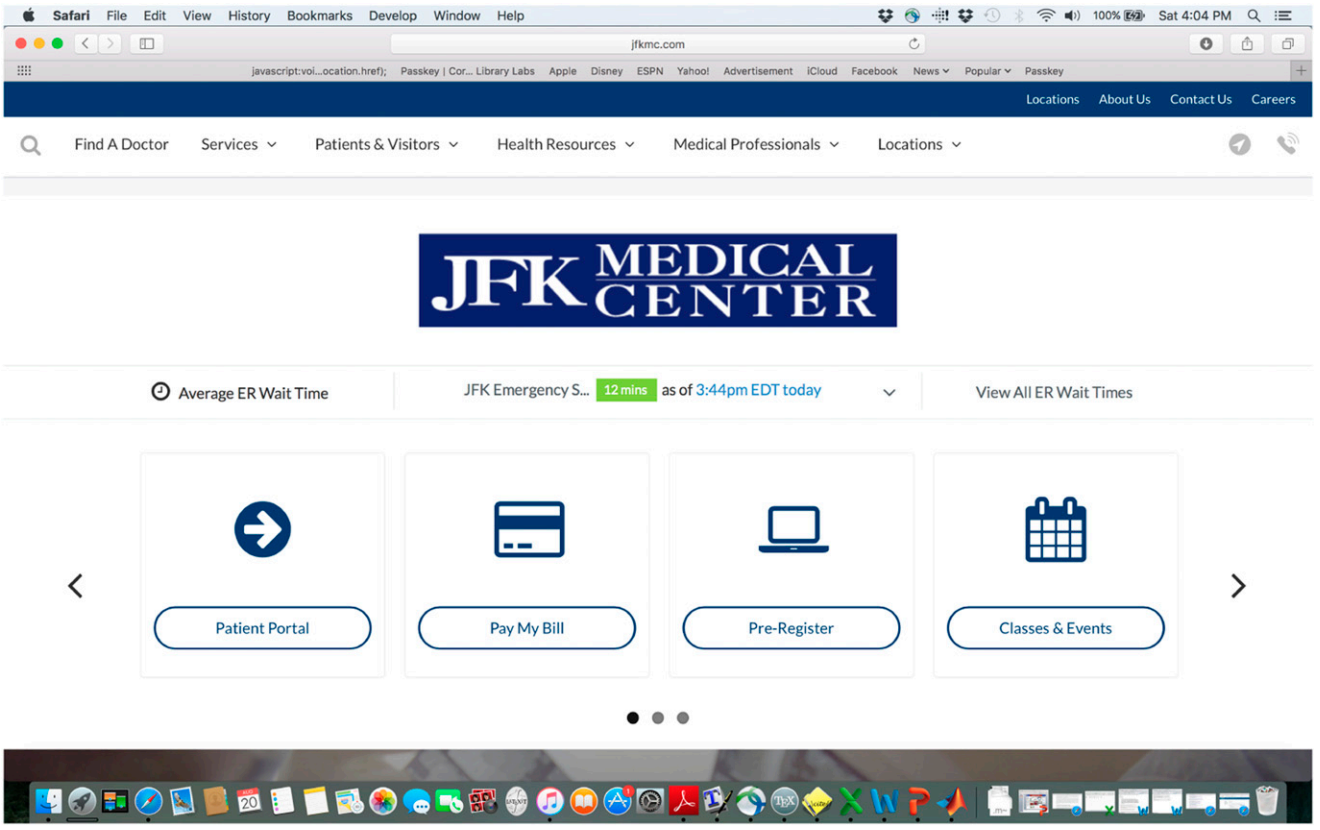


Figure 2. (Color online) Disneyland Park wait times app.

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 that 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 that 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 nontrivial, because 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 that 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.

Because 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 (MNL) model. The MNL model has an economic interpretation where we assume that the utility for being served in the ith 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 ith queue is given by the expression

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

where $Q(t) = (Q_1(t), Q_2(t), \ldots, Q_N(t))$. In this choice function, the parameter θ represents the sensitivity of customers to the queue length. The larger θ is, the more likely a customer is to join the shortest queue. In fact, if one sends $\theta \to \infty$, then the choice function converges to join the shortest queue function.

It is evident from the above expression that, if the queue length in station i is larger than the other queue lengths, then the ith station has a smaller likelihood of receiving the next arrival. This decrease in likelihood as the queue length increases represents the disdain that customers have for waiting in longer lines. We should also mention that the multinomial logit model that we present in this work can be viewed as a smoothed and infinitely differentiable approximation of the join of 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 $t \geq 0$:

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

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

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

counts the number of customers who decide to join the ith 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 ith 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:

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

Note that we scale the rates of both Poisson processes; this 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. [36], because 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, because 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. *Let $\varphi_i(s)$ be a Lipschitz continuous function that keeps track of the previous values on the interval $[-\Delta, 0]$. Then, 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 $\{Q^\eta(t) = (Q_1^\eta(t), Q_2^\eta(t), \ldots, Q_N^\eta(t)\}_{\eta\in\mathbb{N}}$ converges almost surely and uniformly on compact sets of time to $\big(q(t) = (q_1(t), q_2(t), \ldots, q_N(t))$, where*

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

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

See the appendix.

This result states that, as we let η go toward infinity, the sequence of queueing processes converges to a system of delay differential equations. Unlike ordinary differential equations (ODEs), the existence and uniqueness results for delay differential equations are 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:*

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

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 ODEs, a delay differential equation (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 they are truly infinite-dimensional objects. Fortunately, it is 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 numbers 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 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. [36] 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. [36] also characterize the critical delay Δ_{cr} in terms of the model parameters and provide 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. [36] 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 solves for the value of the queue length that makes it zero. This implies that we set

$$\dot{q}_i(t) = 0. \tag{9}$$

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

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

Sometimes, finding the equilibrium is nontrivial in many nonlinear 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, because 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 is 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

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

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

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

This implies that

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

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, which 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 perturbations about the equilibrium point $\frac{\lambda}{N\mu}$. By substituting Equation (15) into Equation (7), we get the following equations:

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

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 define as $w_i(t)$, solves the following linear delay differential equations:

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

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

This can be written as a matrix system by

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

where $\mathcal{I}$ is an N-dimensional identity matrix and $\mathcal{A}$ is an N-dimensional square matrix of ones: that is,

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

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 orthogonal 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, it only has one nonzero eigenvalue. The only nonzero 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 $(1, -1, 0, \dots, 0)^T, (1, 0, -1, 0, \dots, 0)^T, \dots,$

$(1, 0, \ldots, 0, \ldots, -1)^T$ have an eigenvalue with value that is equal to 0. Using this knowledge of the matrix $\mathcal{A}$, we now define $v = S^{-1}w$ or $w = Sv$, and this leads us to the following delay differential system for v:

$$\dot{w}(t) = S\,\dot{v}(t) = -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}w(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot\mathcal{A}w(t-\Delta) - \mu\cdot\mathcal{I}w(t) \tag{22}$$

$$= -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}w(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S\Lambda S^{-1}w(t-\Delta) - \mu\cdot\mathcal{I}w(t) \tag{23}$$

$$= -\frac{\lambda\cdot\theta}{N}\cdot\mathcal{I}Sv(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S\Lambda S^{-1}Sv(t-\Delta) - \mu\cdot\mathcal{I}Sv(t) \tag{24}$$

$$= -\frac{\lambda\cdot\theta}{N}\cdot Sv(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S\Lambda v(t-\Delta) - \mu\cdot Sv(t). \tag{25}$$

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

$$\dot{v}(t) = -\frac{\lambda\cdot\theta}{N}\cdot S^{-1}Sv(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot S^{-1}S\Lambda v(t-\Delta) - \mu\cdot S^{-1}Sv(t) \tag{26}$$

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

Thus, for the ith entry of the vector v, we have the following delay differential equation:

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

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

$$\dot{v}_1(t) = -\frac{\lambda\cdot\theta}{N}\cdot v_1(t-\Delta) + \frac{\lambda\cdot\theta}{N^2}\cdot N\cdot v_1(t-\Delta) - \mu\cdot v_1(t) \tag{29}$$

$$\dot{v}_2(t) = -\frac{\lambda\cdot\theta}{N}\cdot v_2(t-\Delta) - \mu\cdot v_2(t). \tag{30}$$

Reducing these delay differential equations further, we have that

$$\dot{v}_1(t) = -\mu\cdot v_1(t) \tag{31}$$

$$\dot{v}_2(t) = -\frac{\lambda\cdot\theta}{N}\cdot v_2(t-\Delta) - \mu\cdot v_2(t). \tag{32}$$

To finish the proof, we observe that $v_1(t)$ is stable, because 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:

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

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), because 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 and 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*

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

Moreover, when $N + 1 > \frac{\lambda\theta}{\mu}$, we have that $\Delta_{cr}(\lambda, \mu, \theta, N + 1) = \infty$.

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

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

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

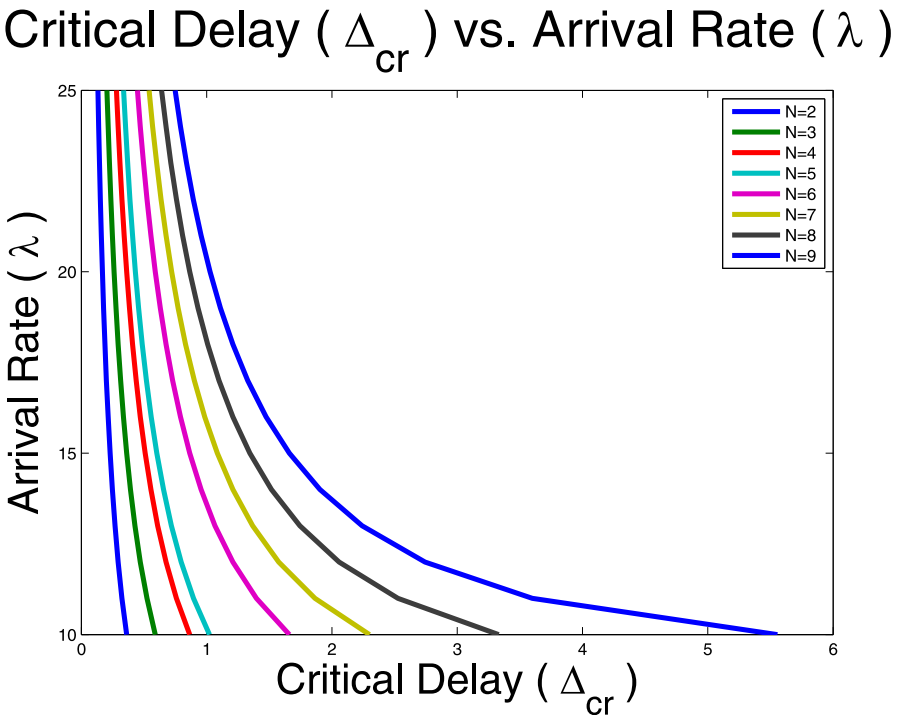
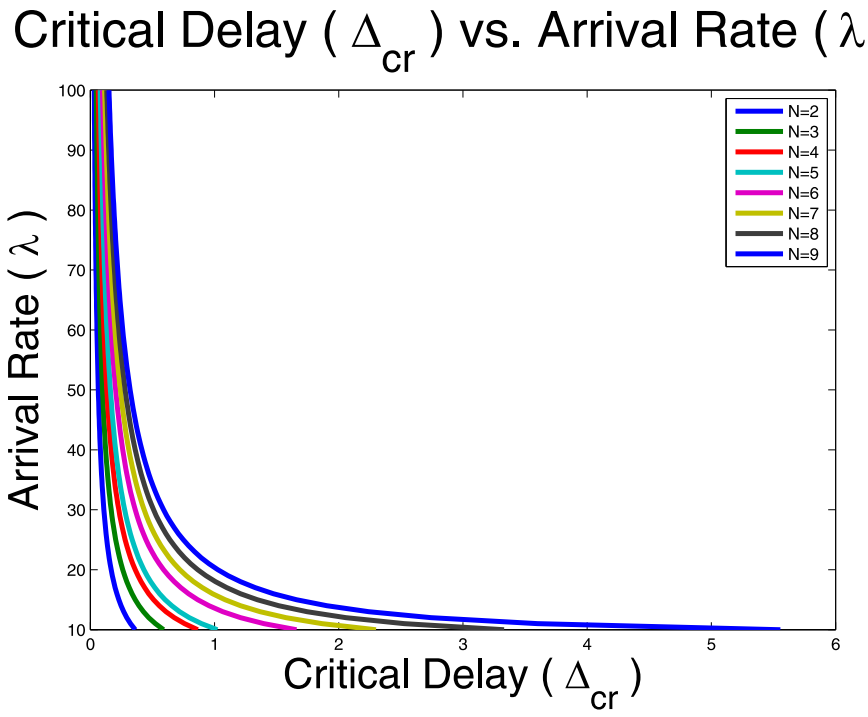


Figure 4. (Color online) Plot of the critical threshold as a function of λ and N. $\lambda \in [10, 100]$ and $\mu = 1$.



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 that the stability region gets larger as we increase N and all other parameters remain fixed. Finally, the critical delay is infinite when $N + 1 > \frac{\lambda\theta}{\mu}$, because the ω_{cr} is no longer real. This 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, because increasing both parameters has an opposite effect on the queue length behavior; increasing λ increases the queue length, whereas 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.

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

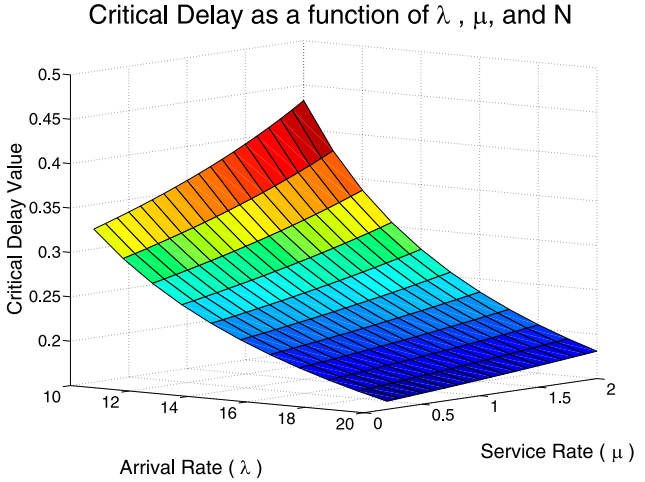
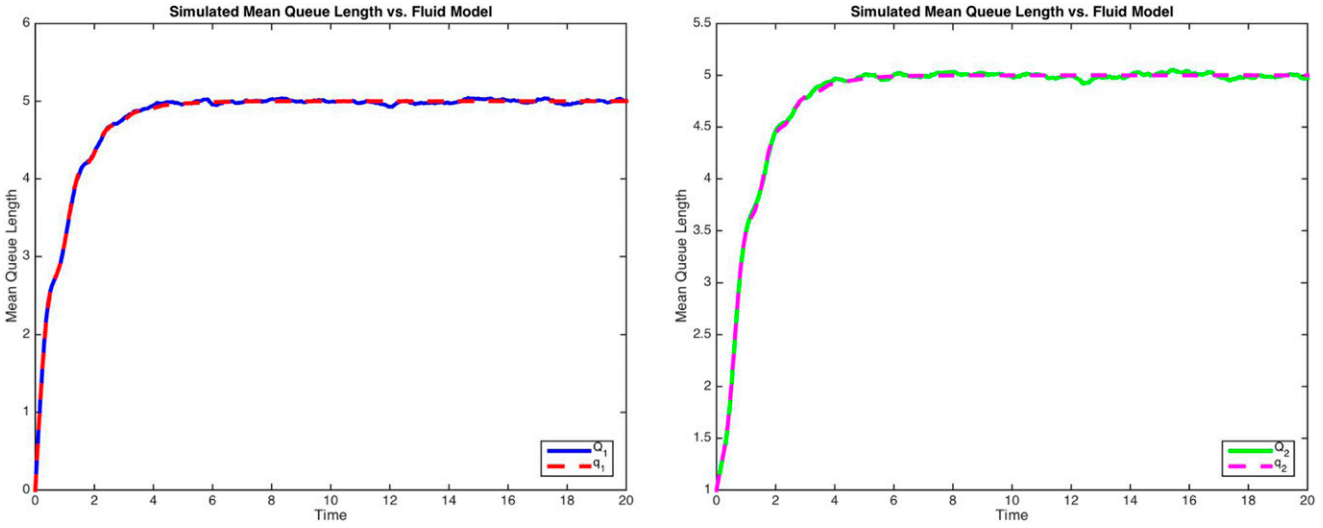


Figure 6. (Color online) $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .25$, and $\eta = \mathbf{10}$. First queue (left panel) and second queue (right panel).



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. However, before describing our results, we describe via references how we perform the simulations of our delayed information queues. Work by Bratsun et al. [9] considers methods based on Gillespie's direct method [14] or the next jump method of Gibson and Bruck [13]. The reader is encouraged to read the simulation section of Anderson and Kurtz [3] or [27, appendix] for more details of the method.

In Figure 6, we plot the case of when $n = 2$, $\eta = 10$, and $\Delta = 0.25$. In the left panel of Figure 6, we compare the simulated first queue with its fluid limit, and in the right panel 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. Because we have that $\Delta = 0.25 < \Delta_{cr} = 0.3614$, 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 $\Delta = 0.45 > \Delta_{cr} = 0.3614$. 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 the limit theorem is wrong, and it does not predict the right behavior of the stochastic model. However, as we will see later, the

Figure 7. (Color online) $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, and $\eta = \mathbf{10}$. First queue (left panel) and second queue (right panel).

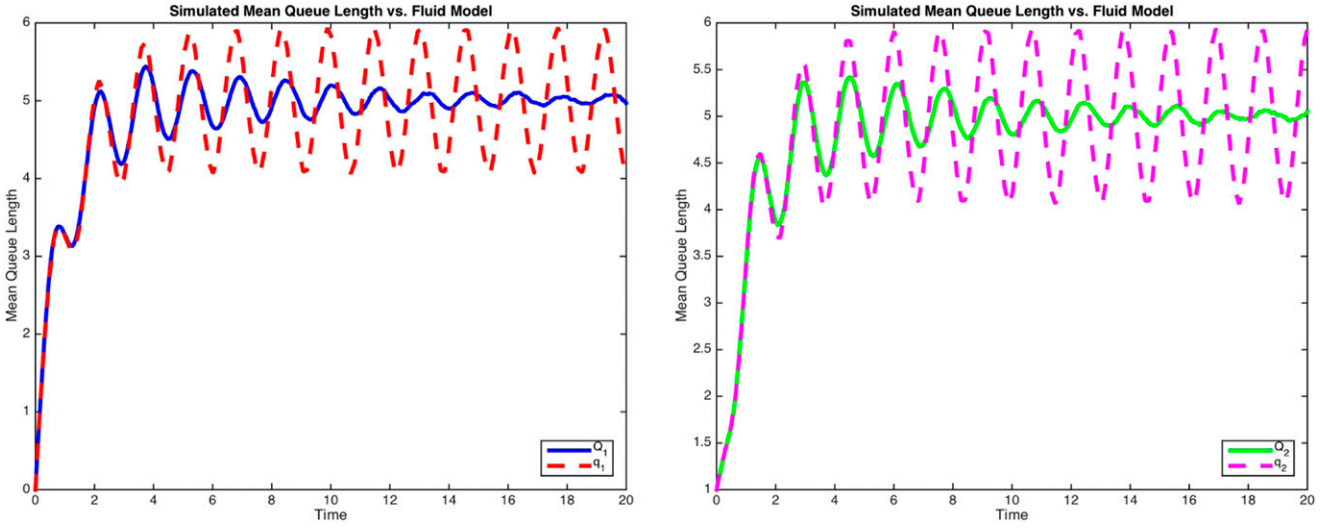
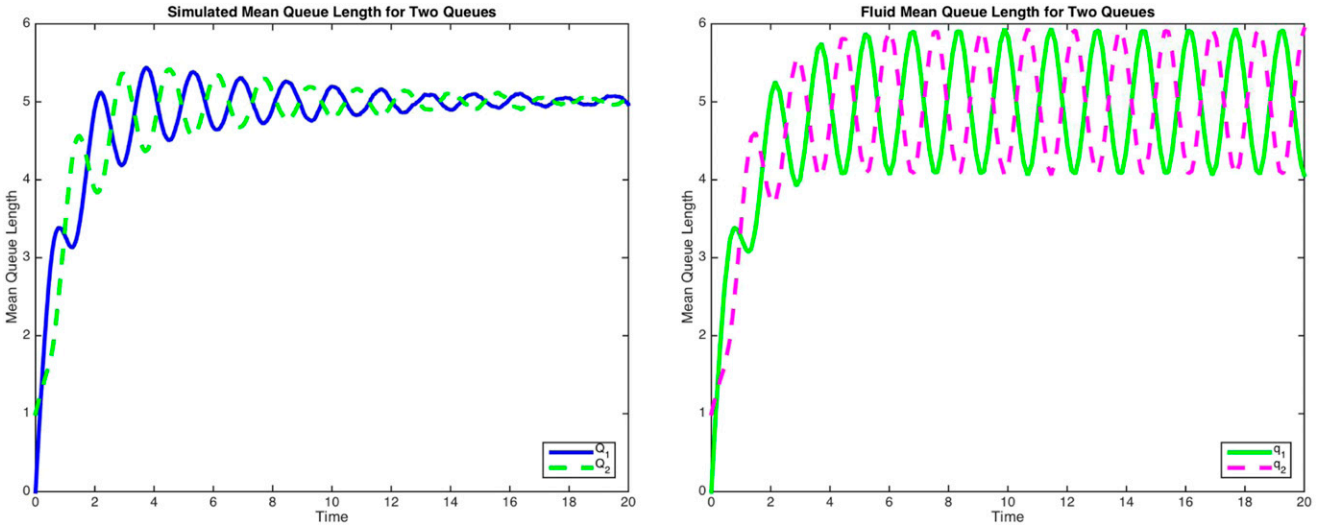


Figure 8. (Color online) $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, and $\eta = \mathbf{10}$. Stochastic simulation (left panel) and fluid limits (right panel).



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, in the left panel of Figure 8, which models the mean of the stochastic system with finite η, the oscillations are damped, and in the right panel of Figure 8, the oscillations are not damped and remain for all time. To explore additional scaling 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. 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 $\varepsilon = 0.05$: that is, $\sup_{t \le T} |Q_i^{\eta}(t) - q_i(t)| < \varepsilon = 0.05$; 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 (Figure 10).

Figure 9. (Color online) $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, and $\eta = \mathbf{100}$. First queue (left panel) and second queue (right panel).

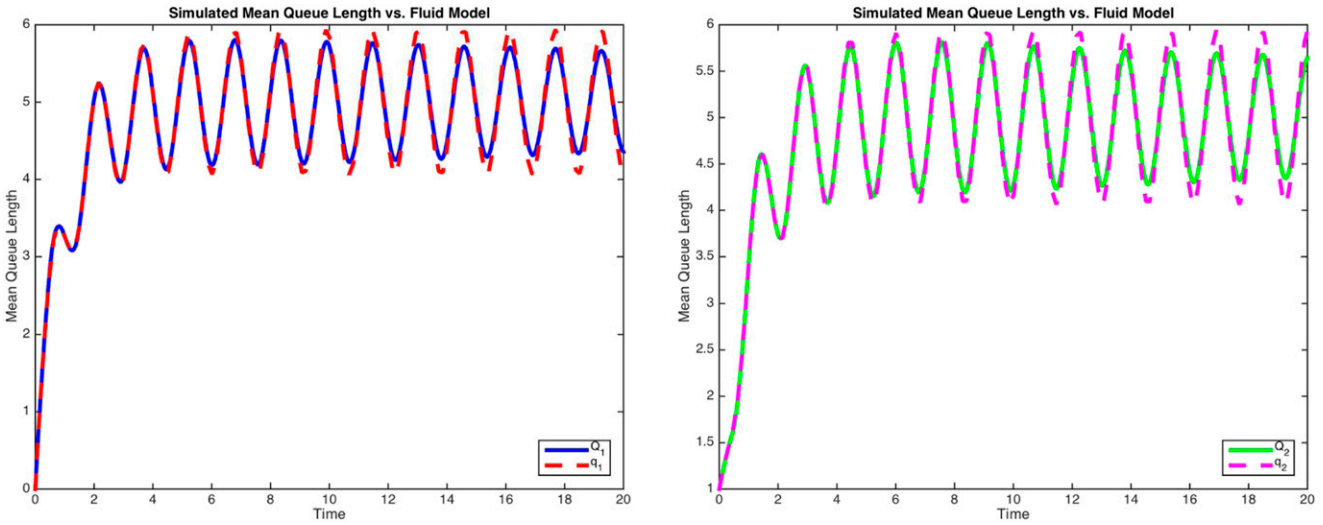
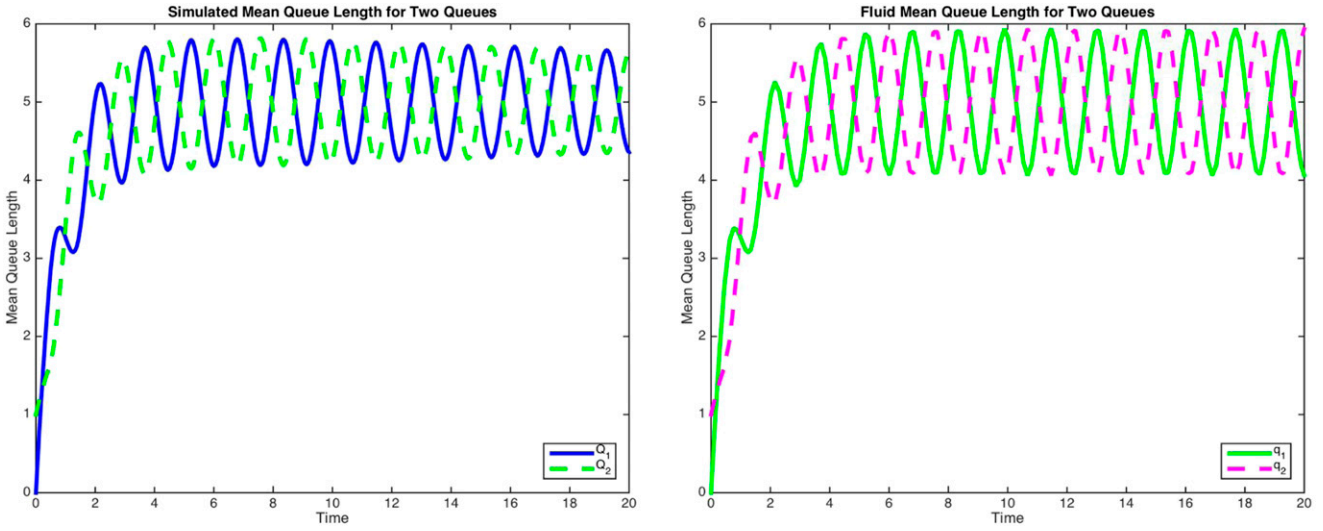


Figure 10. (Color online) $\lambda = 10$, $\mu = 1$, $\Delta_{cr} = .3614$, $\Delta = .45$, and $\eta = \mathbf{100}$. First queue (left panel) and second queue (right panel).



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. Because this limiting object is deterministic, it does not provide us with any insight on the fluctuations of the queue length process around the mean. In order to study the fluctuations 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 is no longer a single point, but instead, it 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, because 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* $(\tilde{D}(t) = (\tilde{D}_1(t), \tilde{D}_2(t), \ldots, \tilde{D}_N(t)))$ *to the stochastic delay integral equations*

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

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

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

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

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. We first must show the existence and uniqueness of 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 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 $\varphi(s) = (\varphi_1(s), \varphi_2(s), \ldots, \varphi_N(s))$, where

$\varphi_i : [-\Delta, 0] \to \mathbb{R}$ and an almost surely continuous sample path $\mathcal{X}_t(\omega)_{t\geq 0}$. The continuous initial function φ_i highlights one of the major differences between delay equations and their nondelayed counterparts. A function is needed in the delay differential equation setting, whereas 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

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

and

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

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. Because 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: that is, $\|\mathcal{H}(\cdot, \gamma)\|_\infty \leq M$. However, the bound M is a random bound that can depend on the sample path of $\mathcal{X}_t(\omega)$. 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 $\mathcal{H}(t,z)$:

$$|\mathcal{H}(t,z)| = |\mathcal{H}(t,z) - \mathcal{H}(t,\gamma) + \mathcal{H}(t,\gamma)| \tag{48}$$

$$\leq |\mathcal{H}(t,z) - \mathcal{H}(t,\gamma)| + |\mathcal{H}(t,\gamma)| \tag{49}$$

$$\leq C \cdot \|z - \gamma\|_\infty + M \tag{50}$$

$$\leq C \cdot \zeta + M. \tag{51}$$

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

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

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

$$|\mathcal{G}(z)(t) - \mathcal{G}(y)(t)| \leq \int_0^t |\mathcal{H}(u,z) - \mathcal{H}(u,y)|du \tag{53}$$

$$\leq C \cdot \delta \max_{-\Delta\leq u\leq\delta} |z(u) - y(u)| \tag{54}$$

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 $\mathcal{X}_t(\omega)$, 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 [7]), 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 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 [12].

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 have proven 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 $\tilde{D}_i^\eta(t)$ 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, 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* $\{\tilde{D}^\eta(t) = (\tilde{D}_1^\eta(t), \tilde{D}_2^\eta(t), \dots, \tilde{D}_N^\eta(t))\}_{\eta\in\mathbb{N}}$ *converges in distribution to the stochastic delay integral equations* $(\tilde{D}(t) = (\tilde{D}_1(t), \tilde{D}_2(t), \dots, \tilde{D}_N(t))$, *where*

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

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

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

Proof. See the 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 they converge 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, 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, because 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.

Because 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 [11], Pender [28, 29, 31, 33], Pender and Massey [34], and Pender et al. [35]. 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 is not the limit theorems. With regard to customer preferences, this is a nontrivial 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 nondifferentiable 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 A.1 (Kurtz [26]). *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 A.2 (Kurtz [26]). *For any standard Brownian motion $\{W(t)\}_{t\geq 0}$ and any $\epsilon > 0$, $n \in \mathbb{N}$, and $T > 0$,*

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

A.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.

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

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

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

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:

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

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

$$\begin{aligned}
Q_i^\eta(t) &= \int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds + \frac{1}{\eta}\mathcal{B}_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right) \\
&\quad - \int_0^t \mu Q_i^\eta(s)ds - \frac{1}{\eta}\mathcal{B}_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) + \mathbb{O}\frac{\log\eta}{\eta} \qquad (A.2) \\
&= \int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds + \frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right) \\
&\quad - \int_0^t \mu Q_i^\eta(s)ds - \frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right) + \mathbb{O}\frac{\log\eta}{\eta}. \qquad (A.3)
\end{aligned}$$

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

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

Now, it remains to show that

$$\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right)\right| = 0 \qquad (A.4)$$

and

$$\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| = 0\,. \qquad (A.5)$$

For the first Brownian motion term, we have that

$$\begin{aligned}
\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right)\right| &\le \lim_{\eta\to\infty}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a(\lambda\cdot T)\right| \qquad (A.6) \\
&= \lim_{\eta\to\infty}\left|\mathcal{B}_i^a\left(\frac{1}{\eta}\cdot\lambda\cdot T\right)\right| \qquad (A.7) \\
&= 0. \qquad (A.8)
\end{aligned}$$

For the second Brownian motion term, we have that

$$\lim_{\eta\to\infty}\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| \le \lim_{\eta\to\infty}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\big((Q^\eta(0)+\lambda)\cdot\mu\cdot T\big)\right| \tag{A.9}$$

$$= \lim_{\eta\to\infty}\left|\mathcal{B}_i^d\left(\frac{1}{\eta}\cdot(Q^\eta(0)+\lambda)\cdot\mu\cdot T\right)\right| \tag{A.10}$$

$$= 0. \tag{A.11}$$

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

$$\left|Q_i^\eta(0) - q_i(0)\right| \le \epsilon/4, \tag{A.12}$$

$$\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp\big(-\theta Q_i^\eta(s-\Delta)\big)}{\sum_{j=1}^N\exp\big(-\theta Q_j^\eta(s-\Delta)\big)}ds\right)\right| \le \epsilon/4, \tag{A.13}$$

$$\sup_{t\le T}\left|\frac{1}{\sqrt{\eta}}\mathcal{B}_i^d\left(\int_0^t \mu Q_i^\eta(s)ds\right)\right| \le \epsilon/4, \tag{A.14}$$

and

$$\mathbb{O}\frac{\log\eta}{\eta} \le \epsilon/4 \tag{A.15}$$

so that we have

$$\begin{aligned}\left|Q_i^\eta(t) - q_i(t)\right| &\le \left|\int_0^t \frac{\lambda\cdot\exp\big(-\theta Q_i^\eta(s-\Delta)\big)}{\sum_{j=1}^N\exp\big(-\theta Q_j^\eta(s-\Delta)\big)}ds - \int_0^t \lambda\cdot\frac{\exp\big(-\theta q_i(s-\Delta)\big)}{\sum_{j=1}^N\exp\big(-\theta q_j(s-\Delta)\big)}ds\right| \\ &\quad + \left|\int_0^t \mu Q_i^\eta(s)ds - \int_0^t \mu q_i(s)ds\right| + \epsilon \qquad (A.16)\\ &\le \int_0^t\left|\frac{\lambda\cdot\exp\big(-\theta Q_i^\eta(s-\Delta)\big)}{\sum_{j=1}^N\exp\big(-\theta Q_j^\eta(s-\Delta)\big)} - \lambda\cdot\frac{\exp\big(-\theta q_i(s-\Delta)\big)}{\sum_{j=1}^N\exp\big(-\theta q_j(s-\Delta)\big)}\right|ds \\ &\quad + \int_0^t\left|\mu Q_i^\eta(s) - \mu q_i(s)\right|ds + \epsilon. \qquad (A.17)\end{aligned}$$

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

$$\left|Q_i^\eta(t) - q_i(t)\right| \le C\int_0^t \sup_{-\Delta\le r\le s}|Q_i^\eta(r) - q_i(r)|ds + \epsilon \tag{A.18}$$

$$\le C\cdot\left(\int_0^t \sup_{0\le r\le s}|Q_i^\eta(r) - q_i(r)|ds + t\cdot\sup_{-\Delta\le r\le 0}|Q_i^\eta(r) - q_i(r)|\right) + \epsilon. \tag{A.19}$$

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

$$\left|Q_i^\eta(t) - q_i(t)\right| \le C\int_0^t \sup_{0\le r\le s}|Q_i^\eta(r) - q_i(r)|ds + \epsilon. \tag{A.20}$$

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 [18], we have that

$$\sup_{0\le t\le T}\left|Q_i^\eta(t) - q_i(t)\right| \le \epsilon\cdot e^{CT}, \tag{A.21}$$

and because ϵ is arbitrary, we can let it go toward zero; this proves the fluid limit. ■

A.2. Proof of Diffusion Limit

In this section, we prove Theorem 5, which shows the convergence of the centered and rescaled queueing process to our system of stochastic delay differential equations. We leverage ideas from the papers of Bayraktar et al. [6] and Horst and Rothe [20].

Proof of Theorem 5.

$$Q_i^\eta(t) = Q_i^\eta(0) + \frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right) - \frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right). \tag{A.22}$$

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

$$\begin{aligned}\sqrt{\eta}\left(Q_i^\eta(t) - q_i(t)\right) &= \sqrt{\eta}\left(Q_i^\eta(0) - q_i(0)\right) + \sqrt{\eta}\cdot X_i^\eta(t)\\ &\quad + \sqrt{\eta}\cdot\int_0^t \left(F_i(s, Q^\eta(s-\Delta), Q^\eta(s)) - F_i(s, q(s-\Delta), q(s))\right)ds,\end{aligned} \tag{A.23}$$

where

$$\begin{aligned}X_i^\eta(t) &= \frac{1}{\eta}\Pi_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right) - \int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\\ &\quad + \frac{1}{\eta}\Pi_i^d\left(\eta\int_0^t \mu Q_i^\eta(s)ds\right) - \int_0^t \mu Q_i^\eta(s)ds\end{aligned} \tag{A.24}$$

and

$$F_i(s, Q^\eta(s)) = \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)} - \mu\cdot Q_i^\eta(s). \tag{A.25}$$

Proposition A.1. *Let $V_i^\eta(t)$ be defined by the following equation:*

$$V_i^\eta(t) = \mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right) + \mathcal{B}_i^d\left(\int_0^t \mu\cdot Q_i^\eta(s)ds\right), \tag{A.26}$$

and then,

$$\lim_{\eta\to\infty}\sup_{0\le t\le T}\left|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)\right| = 0 \quad \textit{in distribution.} \tag{A.27}$$

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

$$\sup_{t\ge 0}\frac{1}{\sqrt{\eta}}\frac{\left|\overline{\Pi}_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right) - \mathcal{B}_i^a\left(\eta\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right)\right|}{\log\left(2\vee\eta\int_0^t \frac{\lambda\cdot\exp\left(-\theta Q_i^\eta(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta Q_j^\eta(s-\Delta)\right)}ds\right)} \le \frac{C_i^a}{\sqrt{\eta}}, \tag{A.28}$$

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

$$\sup_{0\le t\le T}\left|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)\right| \le \log(2\vee\eta KT)\sup_{0\le t\le T}\frac{\left|\sqrt{\eta}\cdot X_i^\eta(t) - V_i^\eta(t)\right|}{\log(2\vee\eta Kt)} \tag{A.29}$$

$$\le \log(2\vee\eta KT)\frac{C_i^a}{\sqrt{\eta}}. \tag{A.30}$$

Because 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

$$\sup_{0\le t\le T}\left|\sqrt{\eta}\cdot X_i^{\eta}(t) - V_i^{\eta}(t)\right| \Rightarrow 0 \quad \text{in distribution as } \eta \to \infty. \tag{A.31}$$

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

Proposition A.2. *The sequence of stochastic processes $V_i^{\eta}(t)$ converges in distribution to the process $V_i(t)$, where*

$$V_i(t) = \mathcal{B}_i^a\left(\int_0^t \frac{\lambda\cdot\exp\left(-\theta q_i(s-\Delta)\right)}{\sum_{j=1}^N \exp\left(-\theta q_j(s-\Delta)\right)}ds\right) + \mathcal{B}_i^d\left(\int_0^t \mu\cdot q_i(s)ds\right). \tag{A.32}$$

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

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

and

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

This implies that

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

However, from Lemma A.2, we obtain

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

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

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

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

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

and therefore,

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

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

$$\mathbb{P}\left(\sup_{0\le t\le T} |\tilde{D}_i^\eta(t)| > K\right) < \epsilon \quad \text{for all } \eta \ge \eta^*. \tag{A.33}$$

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

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

We know that each $V_i^\eta(t)$ is tight, because 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, Billingsley [8, section 15] or Whitt [38, section 3]. Moreover, by using the Lipschitz continuity of the rate functions, we have that

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

for some Lipschitz constant L. Thus, by Gronwall's inequality in Karatzas and Shreve [25, problem 2.7], 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 A.4. *If $\{f^\eta(t), \eta \in \mathbb{N}, t \in \mathbb{R}_+\}$ is a sequence of nonnegative random processes such that*

$$\lim_{\eta\to\infty}\int_0^T f^\eta(s)ds = 0 \quad \text{in probability}, \tag{A.34}$$

then for all $\delta > 0$,

$$\lim_{\eta\to\infty}\mathbb{P}\left(\sup_{0\le t\le T}\left|\int_0^t f^\eta(s)\tilde{D}_i^\eta(s)ds\right| > \delta\right) = 0. \tag{A.35}$$

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

$$\int_0^T f^\eta(s)ds < \epsilon/2 \quad \text{on } \Omega_{\eta,1} \text{ and such that } \mathbb{P}(\Omega_{\eta,1}) \ge 1 - \epsilon/2, \tag{A.36}$$

and

$$\sup_{0\le t\le T}\left|\tilde{D}_i^\eta(t)\right| < K \quad \text{on } \Omega_{\eta,2} \text{ and such that } \mathbb{P}(\Omega_{\eta,2}) \ge 1 - \epsilon/2 \tag{A.37}$$

Therefore, we have that

$$\sup_{0\le t\le T}\left|\int_0^t f^\eta(s)\tilde{D}_i^\eta(s)ds\right| \le \sup_{0\le t\le T}\left|\tilde{D}_i^\eta(t)\right|\int_0^T f^\eta(s)ds < K\epsilon \quad \text{on } \Omega_{\eta,1}\cap\Omega_{\eta,2}. \tag{A.38}$$

The result follows, because ϵ was chosen 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 $\mathcal{G}(g) = (\mathcal{G}^1(g), \mathcal{G}^2(g), \ldots, \mathcal{G}^N(g))$ to the unique function that satisfies the following integral equation

$$\mathcal{G}_t^i(g) = \begin{cases} \varphi_i(t), & t \in [-\Delta, 0], \\ \displaystyle\int_0^t \langle \nabla F_i(s, q(s)), \mathcal{G}_s^i(g)\rangle ds + g_i(t), & t \in [0, T], \end{cases} \tag{A.39}$$

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). Because 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 [25]. Moreover, we know that V^η converges to V in probability, and this implies that

$$\lim_{\eta\to\infty} \left\| \mathcal{G}(V^\eta) - \tilde{D} \right\|_\infty = 0. \tag{A.40}$$

Therefore, if we can show that the following difference

$$\lim_{\eta\to\infty} \sup_{0\le t\le T} \left\| \tilde{D}^\eta(t) - \mathcal{G}(V^\eta)(t) \right\|_\infty = 0 \tag{A.41}$$

converges to 0 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

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

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

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

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

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

We also know that

$$\lim_{\eta\to\infty} \sup_{0\le t\le T} \|\nabla F_i(t, \xi^\eta(t)) - \nabla F_i(t, q(t))\| = 0 \quad \text{a.s} \tag{A.42}$$

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

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

Finally, by the application of Gronwall's inequality in Karatzas and Shreve [25, problem 2.7] and Lemma A.4, we obtain our diffusion limit result, because i was chosen arbitrarily. ■

References

[1] Allon G, Bassamboo A (2011) The impact of delaying the delay announcements. *Oper. Res.* 59(5):1198–1210.
[2] Allon G, Bassamboo A, Gurvich I (2011) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.
[3] Anderson DF, Kurtz TG (2011) Continuous time markov chain models for chemical reaction networks. *Design and Analysis of Biomolecular Circuits* (Springer, New York), 3–42.
[4] Armony M, Maglaras C (2004) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52(2):271–292.
[5] Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1): 66–81.
[6] Bayraktar E, Horst U, Sircar R (2006) A limit theorem for financial markets with inert investors. *Math. Oper. Res.* 31(4):789–810.

[7] Bharucha-Reid AT (1976) Fixed point theorems in probabilistic analysis. *Bull. Amer. Math. Soc* 82(5):641–657.
[8] Billingsley P (2013) *Convergence of Probability Measures* (John Wiley & Sons, New York).
[9] Bratsun D, Volfson D, Tsimring LS, Hasty J (2005) Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. USA* 102(41):14593–14598.
[10] Dong J, Yom-Tov E, Yom-Tov GB (2018) The impact of delay announcements on hospital network coordination and waiting times. *Management Sci.* 65(5):1969–1994.
[11] Engblom S, Pender J (2014) Approximations for the moments of nonstationary and state dependent birth-death queues.
[12] Ge X, Zhu Y (2012) Existence and uniqueness theorem for uncertain delay differential equations. *J. Comput. Inform. Systems* 8(20):8341–8347.
[13] Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Physical Chemistry A* 104(9):1876–1889.
[14] Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chemical Physics* 115(4):1716–1733.
[15] Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53(6):962–970.
[16] Guo P, Zipkin P (2009) The impacts of customers'delay-risk sensitivities on a queue with balking. *Probab. Engrg. Inform. Sci.* 23(03):409–432.
[17] Hale JK (1971) Functional differential equations. *Analytic Theory of Differential Equations* (Springer, Berlin, Heidelberg), 9–22.
[18] Hale JK (1980) *Ordinary differential equations*, 2nd ed. (Krieger Publishing, Malabar, FL).
[19] Hassin R (2007) Information and uncertainty in a queuing system. *Probab. Engrg. Inform. Sci.* 21(03):361–380.
[20] Horst U, Rothe C (2008) Queuing, social interactions, and the microstructure of financial markets. *Macroeconomics Dynamics* 12(2):211–233.
[21] Ibrahim R, Armony M, Bassamboo A (2016) Does the past predict the future? the case of delay announcements in service systems. *Management Sci.* 63(6):1762–1780.
[22] Jennings OB, Pender J (2016) Comparisons of ticket and standard queues. *Queueing Systems* 84(1-2):145–202.
[23] Jouini O, Dallery Y, Akşin Z (2009) Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *Internat. J. Production Econom.* 120(2):389–399.
[24] Jouini O, Aksin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing Service Oper. Management* 13(4): 534–548.
[25] Karatzas I, Shreve S (2012) Brownian *Motion and Stochastic Calculus*, vol. 113 (Springer Science & Business Media, New York).
[26] Kurtz TG (1978) Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Appl.* 6(3):223–240.
[27] Massey WA, Pender J (2013) Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* 75(2-4):243–277.
[28] Pender J (2014) Gram charlier expansion for time varying multiserver queues with abandonment. *SIAM J. Appl. Math.* 74(4):1238–1265.
[29] Pender J (2015) An analysis of nonstationary coupled queues. *Telecomm. Systems* 1–16.
[30] Pender J (2015) Heavy traffic limits for unobservable queues with clearing times. Technical Report 1, Cornell University, Ithaca, NY.
[31] Pender J (2015) Nonstationary loss queues via cumulant moment approximations. *Probab. Engrg. Inform. Sci.* 29(1):27–49.
[32] Pender J (2015) The impact of dependence on unobservable queues. Technical Report 2, Cornell University, Ithaca, NY.
[33] Pender J (2015) The truncated normal distribution: Applications to queues with impatient customers. *Oper. Res. Lett.* 43(1):40–45.
[34] Pender J, Massey WA (2017) Approximating and stabilizing dynamic rate jackson networks with abandonment. *Probab. Engrg. Inform. Sci.* 31(1):1–42.
[35] Pender J, Rand RH, Wesson E (2018) An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynamics* 91(4):2411–2427.
[36] Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *Internat. J. Bifurcation Chaos Appl. Sci. Engrg.* 27(4):1730016-1–1730016-20.
[37] Whitt W (1999) Improving service by informing customers about anticipated delays. *Management Sci.* 45(2):192–207.
[38] Whitt W (2007) Proofs of the martingale FCLT. *Probab. Surveys* 4:268–302.