



International Journal of Bifurcation and Chaos, Vol. 32, No. 12 (2022) 2250174 (14 pages)
 World Scientific Publishing Company
 DOI: 10.1142/S0218127422501747

k-Nearest Neighbor Queues with Delayed Information

Jamol Pender

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

Received ; Revised

In this paper, we analyze a model called the k-nearest neighbor queue with the possibility of having delayed queue length feedback. We prove fluid limits for the stochastic queueing model and show that the fluid limit converges to a system of delay differential equations. Using the properties of circulant matrices, we derive a closed form expression for the value of the critical delay, which governs whether the delayed information will induce oscillations or a Hopf bifurcation in our queueing system.

Keywords: Queueing theory; k-nearest neighbor; delayed information; delay differential equation; stochastic process; fluid limit; Hopf bifurcation.

1. Introduction

Cloud computing services are pervasive in our society and are expanding across the world. These services are supported by very complex infrastructures in data centers. As demand for cloud computing services continues to increase it is important to understand how to manage these data centers so that they do not overload. It is not surprising that many researchers are now developing new load balancing algorithms for data center applications. Since data centers can be modeled well by queueing theory, often researchers use queueing theory to verify that their algorithms perform well. However, many data center managers must make the trade off between using storing information about current system state or component of the system load among several queues. Many algorithms have been developed to do this in the context of data centers, telecommunications networks, and even call centers.

In order to achieve load balancing many authors have analyzed stochastic queueing models that are modifications of join the shortest queue, see for example the work by [Mitzenmacher, 2001;

Byers *et al.*, 2004; Xie *et al.*, 2015; Bramson *et al.*, 2010; Lu *et al.*, 2011; Bramson *et al.*, 2013; Mukherjee *et al.*, 2016; Mukherjee *et al.*, 2018; Mukherjee *et al.*, 2017; Banerjee *et al.*, 2019; Budhiraja *et al.*, 2017; Aghajani & Ramanan, 2017; Aghajani *et al.*, 2015, 2017; Cybenko, 1989; Mitzenmacher, 2016]. Many of these variants of join the shortest queue allow for joining the shortest among d randomly selected queues. Selecting only $d \geq 2$ queues allows for less storage of all queue lengths and still achieves high quality performance.

Our paper is motivated by the load balancing literature, but is also motivated by the work in [Dong *et al.*, 2018; Ding *et al.*, 2020] where customers have the option to choose which queue they join based on the queue length they observe in a delay announcement or smartphone app. The work of Dong *et al.* [2018], Ding *et al.* [2020] explores the use of the multinomial logit choice model (MNL) as a probabilistic way for customers to choose which queue they will join. However, the current literature assumes two important, but not realistic properties about the information that customers have when

J. Pender

making their decision of which queue to join. The first assumption is that customers have information about all queues. In the context of telecommunication systems and data centers, this assumption is not realistic as it requires a substantial amount of data storage and knowledge. Thus, it is common in these settings that a queue might have information only about its neighbors. The second assumption is that customers receive information in a real-time fashion. This is also not realistic as there are many situations where either the information is delayed from a technological point of view or the decision about joining a queue must be made before actually joining the queue. In both situations, the information can be viewed as being delayed.

The queueing model that we present in this work tackles both of these gaps in the literature. First we consider a $2k$ -nearest neighbor setup where a customer who will join the i th queue also knows the queue length of the k neighbors to the left and k neighbors to the right. One might view this nearest neighbor setup as each queue knowing some local information about the queues nearest to it. Second, we consider the fact that the information about the queues is delayed and is not given in real-time. This is important to consider in our model as the rate of information about the system is not infinitely fast. Thus, the queue length information about the neighbors must be delayed.

We should also mention that our work is closely related to the work by Mitzenmacher [2000], Ding *et al.* [2020], Pender *et al.* [2020], Lipshutz & Williams [2015], Novitzky *et al.* [2020], Doldo *et al.* [2021] in that these papers either consider choice model dynamics in the construction of their queueing models or they consider delayed dynamics in the context of queueing theory. We should also mention that there is recent work by Atar and Lipshutz [2021] that considers heavy traffic limits for systems with delayed information. Thus, our work is similar in this spirit.

Finally, we would like to illustrate how this work is different from the following papers Pender *et al.* [2020], Novitzky *et al.* [2020], Doldo *et al.* [2021]. First our work considers a more general choice model where the choice function is not limited to just exponential functions. A second difference is that we are using the k -nearest neighbors selection for comparisons. This k -nearest neighbor selection is more general than the previous situation where all queues are compared. This k -nearest

neighbor feature is more practical in physical settings where a comparison to a queue far away is not possible in reality. As an example, if one thinks about a grocery store, generally it is not easy to compare lines that are far away, however, it is easy to compare lines that are nearby the one you are in. Thus, in this way the k -nearest neighbor model is more realistic. Moreover, the k -nearest neighbor model is also important because in the standard MNL model, the delay differential equations decouple reduces to analyzing just one dde. However, this is not the case k -nearest neighbor model and this model requires a more refined multidimensional analysis based on circulant matrices.

1.1. *Our contributions*

This paper makes the following two major contributions:

First, we develop a new model of k -nearest neighbor queues, where the information about each queue length is delayed by a constant Δ . Second, we prove fluid limits for a scaled version of the queue length process and show that the fluid limit is a delay differential equation. Moreover, we prove the exact threshold for when oscillations in the queue length dynamics will occur in this fluid model.

1.2. *Organization of the paper*

The rest of the paper is organized as follows:

In Sec. 2, we describe and construct the k -nearest neighbor queueing model. We prove the fluid limit of the scaled queue length process and derive an exact threshold where oscillations will occur if the delay in information is larger than the threshold. In Sec. 3, we conclude and provide new directions for future research. Finally, all proofs of our main results are given in the Appendix.

2. k -Nearest Neighbor Queueing Model

In this section, we present a new stochastic queueing model where customers that would arrive to the i th queue are allowed to also join any of the k neighbors to the right or to the left of the i th queue. Thus, any arriving customer will have the option of joining $2k + 1$ queues. This choice reflects the fact that often in load balancing settings, one might not have

the information of all of the possible queues one could join since it can be computationally expensive to store all of this information. In settings where queue length information is provided to customers via smartphone apps like in bike-sharing networks [Schuijbroek *et al.*, 2017; Faghih-Imani *et al.*, 2017] or waiting times at Disneyland [Nirenberg *et al.*, 2018], a smartphone app will only indicate the nearest possible queues you can retrieve or return a bike. Thus, the information that is given to the customer is limited to the nearest stations to them. 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. In addition to only being able to know the queue length of one's k -nearest neighbors to the left and to the right, we also add the complication 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 not in real-time, which is commonly assumed. In fact, the queue length information that a customer actually receives is the queue length Δ time units in the past.

Thus, in a stochastic context with N queues, the probability of joining the i th queue is given by the following expression

$$p_i(Q(t), \Delta) = \frac{g(Q_i(t - \Delta))}{\sum_{j=1}^k g(Q_{i-j}(t - \Delta)) + g(Q_{i+j}(t - \Delta)) + g(Q_i(t - \Delta))}, \quad (1)$$

where $Q(t) = (Q_1(t), Q_2(t), \dots, Q_N(t))$. We also make the following assumptions regarding the function $g(x)$.

- The function $g(x)$ maps from $\mathbb{R}_+ \rightarrow \mathbb{R}_+/\{0\}$ and is a continuously differentiable function with a uniformly bounded derivative. Moreover, we assume the function is strictly nonincreasing.
- $g(\frac{\lambda}{\mu(2k+1)})$ is bounded away from zero $\forall k \in \mathbb{N}$ and $k \leq \frac{N-1}{2}$.
- $g'(x) < 0, \forall x \geq 0$.

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. 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 $t \geq 0$

$$\begin{aligned} Q_i(t) = & Q_i(0) + \Pi_i^a \int_0^t \lambda \cdot p_i(Q(s), \Delta) ds \\ & - \Pi_i^d \int_0^t \mu Q_i(s) ds, \end{aligned}$$

where each $\Pi(\cdot)$ is a unit rate Poisson process and $Q_i(s) = \phi_i(s)$ for all $s \in [-\Delta, 0]$. In this model, for

the i th queue, we have that

$$\Pi_i^a \int_0^t \lambda p_i(Q(s), \Delta) ds \quad (2)$$

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

$$\Pi_i^d \int_0^t \mu Q_i(s) ds \quad (3)$$

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. Special cases of the model

In this section, we specify some forms for the function $g(x)$. In the case that $g(x) = e^{-\theta x}$, we have the choice model is given by a k -nearest neighbor version of the multinomial logit choice model (MNL). In the MNL choice model, there is an analogous utility perspective about the choices that customers make. In fact the utility for being served in the i th queue with delayed queue length $Q_i(t - \Delta)$

J. Pender

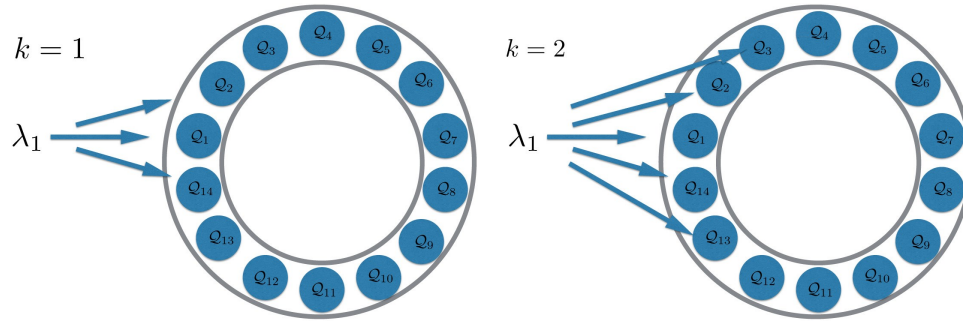


Fig. 1. k-nearest neighbor queue with $k = 1$ (left) and $k = 2$ (right).

is $u_i(Q_i(t - \Delta)) = -\vartheta Q_i(t - \Delta)$. Thus, there is an economic interpretation of the MNL model. We should also mention that the MNL model has two important asymptotic regimes, which are of independent interest. In the setting where we let the sensitivity parameter $\vartheta \rightarrow 0$, we find that our probabilities of choosing a queue to join converge to a uniform distribution over the k-nearest neighbors. Moreover, in the setting where we let the sensitivity parameter $\vartheta \rightarrow \infty$, we find that our probabilities of choosing a queue to join converge to an indicator function for shortest queue. If there is a tie, then it is uniform over those queues that are identical for the smallest queue length. As a result, we can view the MNL model outside of those two extremes as a smoothed and infinitely differentiable approximation of the join the shortest queue model.

Another function that could be used is the polynomial $g(x) = \frac{\gamma}{(\gamma x^\alpha + c)^\sigma}$. This function is commonly used in the context of utility maximization and is a version of the Cobb–Douglas family of utilities used in economics. The interested reader for additional functions can see the many references in [Hassin & Haviv, 2003; Hassin, 2016].

Finally, we can also suggest that any tail cdf of a continuous non-negative distribution could also work as a potential function for $g(x)$. For example $g(x) = e^{-(x\vartheta)^k}$ is the tail cdf of a Weibull distribution. Another distribution potentially is the folded normal distribution where the tail cdf is given by the function $g(x) = \frac{1}{2}[\text{erf}(\frac{x+\mu}{2\sigma^2}) + \text{erf}(\frac{x-\mu}{2\sigma^2})]$ for $x \geq 0$ and where $\text{erf}()$ is the error function.

2.2. Fluid 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

$$Q_i^\eta(t) = Q_i^\eta(0) + \Pi_i^a \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \frac{1}{\eta} \Pi_i^d \int_0^t \mu Q_i^\eta(s) ds. \quad (4)$$

Note that we write $Q_i^\eta(0)$ for the initial condition, however, it really should be interpreted as an initial function since it will contain all of the information needed to know the queue length in the past Δ time units. Now by letting the scaling parameter η go to infinity gives us our first result.

Theorem 1. *Let $\gamma_i(s)$ be a non-negative Lipschitz continuous function that knows all of the queue length values on the interval $[-\Delta, 0]$. Then, if $Q_i^\eta(s) \rightarrow \gamma_i(s)$ almost surely for all $s \in [-\Delta, 0]$ and for all $1 \leq i \leq N$, then the sequence of stochastic processes $\{Q^\eta(t) = (Q_1^\eta(t), Q_2^\eta(t), \dots, Q_N^\eta(t))\}_{\eta \in \mathbb{N}}$ converges almost surely and uniformly on compact sets of time to $(q(t) = (q_1(t), q_2(t), \dots, q_N(t)))$ where*

$$\dot{q}_i(t) = \lambda \cdot p_i(q(t), \Delta) - \mu q_i(t) \quad (5)$$

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

Proof. 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 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 non-negative Lipschitz continuous initial function $\phi_i : [-\Delta, 0] \rightarrow \mathbb{R}_+$ for all $1 \leq i \leq N$ and a finite time horizon $T > 0$, there exists a unique Lipschitz continuous function $q(t) = \{q_i(t)\}_{-\Delta \leq t \leq T}$ that is the solution to the following delay differential equation*

$$\dot{q}_i(t) = \lambda \cdot p_i(q(t), \Delta) - \mu q_i(t) \quad (6)$$

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

Proof. The proof of this result can be found in [Hale, 1971] as our model satisfies the Lipschitz continuity conditions of the right-hand side.

2.3. Oscillations in the *k*-nearest neighbor model

Unlike ordinary differential equations, delay differential equations are truly infinite dimensional and the smallest of delays can cause surprising dynamics. Recent work by Pender *et al.* [2017] explores a two-dimensional version of our fluid limit and uncovers that the two queues can oscillate in equilibrium when the delay is large enough. Pender *et al.* [2017] also characterizes the threshold in terms of the model parameters and provides an exact formula for the threshold in the two-dimensional case. However, this analysis is limited and does not immediately generalize to the multidimensional setting. The main goal of this section is to generalize the critical delay analysis of Pender *et al.* [2017] and derive the exact threshold for an arbitrary number of queues. Before we state the formal theorem, we will need a lemma, which is stated below. This lemma provides an explicit expression for the sum of cosines in terms of the Dirichlet kernel. We will need the lemma for analyzing the stability of the delay differential equations that arise from our fluid model.

Lemma 1

$$\sum_{n=1}^N \cos(n\theta) = -\frac{1}{2} + \frac{\sin\left(N + \frac{1}{2}\theta\right)}{2 \sin\left(\frac{\theta}{2}\right)}. \quad (7)$$

Proof. Start with the series

$$f(x) = \frac{1}{2} + \sum_{k=1}^N \cos(kx). \quad (8)$$

Then one should multiply both sides of the above by $2 \sin(x/2)$ and use the trigonometric identity

$$\cos(a) \sin(b) = \frac{\sin(a+b) - \sin(a-b)}{2} \quad (9)$$

to reduce the right-hand side to

$$\sin\left(n + \frac{1}{2}\right) x. \quad (10)$$

This completes the proof.

Lemma 2. *The equilibrium for the delay differential system of equation given in Eq. (6) is unique.*

Proof. 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 derivatives for all 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

$$\dot{q}_i(t) = 0 \quad (11)$$

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

$$\lambda \cdot p(c_i, \Delta) = \mu \cdot c_i. \quad (12)$$

This implies that

$$\begin{aligned} \frac{g(c_i)}{c_i} &= \frac{\mu}{\lambda} \cdot \frac{g(c_i)}{\sum_{j=1}^N g(c_{i-j}) + g(c_{i+j}) + g(c_i)} \\ &= \text{const.} \end{aligned} \quad (13)$$

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

Theorem 3. *For the constant delay choice queueing model given in Eq. (6) with arbitrary $N \geq 2$,*

J. Pender

the critical delay, $\Delta_{\text{cr}}(\lambda, \mu, N)$, is given by the following

$$\Delta_{\text{cr}}(\lambda, \mu, N, k) = \frac{\arccos \frac{\mu}{\min_{1 \leq j \leq N} \alpha_j}}{\min_{1 \leq j \leq N} \alpha_j^2 - \mu^2}, \quad (14)$$

where α_j is given by

$$\alpha_j = \frac{\sin \left(\frac{\pi j}{N} \right) \cdot \frac{\sin \left(\frac{\pi (2k+1)j}{N} \right)}{\sin \left(\frac{\pi j}{N} \right)} \cdot \frac{\lambda \cdot g^0}{(2k+1)\mu} + \frac{\lambda}{(2k+1)\mu}}{\lambda} \quad (15)$$

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

$$\dot{q}_i(t) = 0. \quad (16)$$

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

$$\lambda \cdot p_i(q(t), \Delta) - \mu q_i(t) = 0. \quad (17)$$

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 Eq. (17) 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 $1/(2k+1)$ of the arrivals and the service rates of all queues are the same. Thus, we have in equilibrium that for all $1 \leq i \leq N$

$$q_i(t - \Delta) = q_i(t) = \frac{\lambda}{(2k+1)\mu} \quad \text{as } t \rightarrow \infty. \quad (18)$$

To mathematically verify that this is an equilibrium for the system of equations, one can substitute

$\frac{\lambda}{(2k+1)\mu}$ for $q_i(t)$ and $q_i(t - \Delta)$ and make the observation that the time derivative for all equations are equal to zero.

Now that we have established the equilibrium for Eq. (6), we need to understand the stability of the delay differential equations near the equilibrium. The first step in doing this is to set each queue length to the equilibrium points plus a perturbation. With this in mind, we substitute the following values for each queue length

$$q_i(t) = \frac{\lambda}{(2k+1)\mu} + u_i(t). \quad (19)$$

In this substitution, the $u_i(t)$ are perturbations about the equilibrium point $\frac{\lambda}{(2k+1)\mu}$. By substituting Eq. (19) into Eq. (6) we obtain the following equations

$$\begin{aligned} \dot{u}_i(t) = & \lambda \cdot p_i \left(\frac{\lambda}{(2k+1)\mu} + u(t), \Delta \right) \\ & - \mu u_i(t) - \frac{\lambda}{(2k+1)\mu}. \end{aligned} \quad (20)$$

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

$$\dot{w}_i(t) = \frac{2k\lambda \cdot g^0}{(2k+1)^2 g} \frac{\lambda}{(2k+1)\mu} w_i(t - \Delta) - \sum_{j=1}^{b(k+1)/2-1} \frac{\lambda \cdot g^0}{(2k+1)^2 g} \frac{\lambda}{(2k+1)\mu} w_{i-j}(t - \Delta)$$

$$\begin{aligned}
 & - \sum_{j=1}^{b(k+1)/2c-1} \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{w_{i+j}(t-\Delta) - \mu \cdot w_i(t)} (2k+1)^2 g (2k+1)\mu \\
 & = \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} \cdot w_i(t-\Delta) - \sum_{j=1}^{b(k+1)/2c-1} \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \cdot w_{i-j}(t-\Delta) \\
 & - \sum_{j=0}^{b(k+1)/2c-1} \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{w_{i+j}(t-\Delta) - \mu \cdot w_i(t)} (2k+1)^2 g (2k+1)\mu
 \end{aligned}$$

This can be written as a matrix system by

$$\begin{aligned}
 \dot{w}(t) &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} I \cdot w(t-\Delta) \\
 & - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} C w(t-\Delta) \\
 & - \mu \cdot I w(t),
 \end{aligned} \tag{21}$$

where I is an N -dimensional identity matrix and C is a N -dimensional symmetric circulant matrix. Circulant matrices are ideal since much is known about their eigenvalues. Thus, C has the following representation

$$C = \begin{bmatrix} c_0 & c_1 & \dots & c_{N/2-1} & c_{N/2} & c_{N/2-1} & \dots & c_1 & c_0 \\ c_1 & c_0 & & & c_{N/2-1} & c_{N/2} & & & c_1 \\ c_2 & c_1 & c_0 & \ddots & \vdots & \vdots & \ddots & c_1 & c_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{N/2-1} & c_{N/2-2} & \dots & c_1 & c_0 & c_1 & \dots & c_{N/2-2} & c_{N/2-1} \end{bmatrix}$$

With the representation of our linearized system in Eq. (21), we can now exploit the fact that both C and I can be simultaneously diagonalized. Thus, we can write both C and I in terms of the eigenvectors of the matrix C . If we denote S as the orthogonal matrix of the eigenvectors of C and denote Λ as diagonal matrix of the eigenvalues of C , then we have that C and I can both be decomposed in terms

of S , U^{-1} , and Λ as

$$C = U \Lambda U^{-1} \tag{22}$$

$$I = U I U^{-1}. \tag{23}$$

The matrix C has rank $N - 2k$ and therefore has $N - 2k$ distinct eigenvalues. The eigenvalues of any real symmetric matrix are real. Thus, the corresponding eigenvalues of our circulant matrix have explicit expressions in terms of the roots of unity i.e.

$$\begin{aligned}
 \lambda_j &= c_0 + 2c_1 \cos \omega_j + 2c_2 \cos \omega_j^2 \\
 &+ \dots + 2c_{n/2-1} \cos \omega_j^{n/2-1} + c_{n/2} \omega_j^{n/2}
 \end{aligned} \tag{24}$$

for n even, and

$$\begin{aligned}
 \lambda_j &= c_0 + 2c_1 \cos \omega_j + 2c_2 \cos \omega_j^2 \\
 &+ \dots + 2c_{(n-1)/2} \cos \omega_j^{(n-1)/2}
 \end{aligned} \tag{25}$$

for odd n . This can be further simplified by using the roots of unity identity that

$$\cos \omega_k = \cos \frac{2\pi jk}{N}. \tag{26}$$

Thus, the j th eigenvalue of the circulant matrix is given by the following expression

$$\begin{aligned}
 \lambda_j &= 1 + \sum_{m=1}^k 2 \cos \frac{2\pi jm}{N} \\
 &= \frac{1}{\sin \frac{\pi j}{N}} \cdot \frac{1}{2} \cdot \frac{2\pi j}{N} \sin \frac{2\pi j}{N} \tag{27}
 \end{aligned}$$

J. Pender

Using this knowledge of the eigenvalues of the matrix C , we now define, $v = U^{-1}w$ or $w = Uv$ and this leads us to the following delay differential system for v

$$\begin{aligned}
 \dot{w}(t) &= U\dot{v}(t) \\
 &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} \cdot w(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \cdot Cw(t - \Delta) - \mu \cdot w(t) \\
 &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} \cdot w(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \cdot U\Lambda U^{-1}w(t - \Delta) - \mu \cdot w(t) \\
 &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} \cdot Uv(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \cdot U\Lambda U^{-1}Uv(t - \Delta) - \mu \cdot Uv(t) \\
 &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} \cdot Uv(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \cdot U\Lambda v(t - \Delta) - \mu \cdot Uv(t). \tag{28}
 \end{aligned}$$

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

$$\begin{aligned}
 \dot{v}(t) &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} U^{-1}Uv(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} U^{-1}U\Lambda v(t - \Delta) - \mu \cdot U^{-1}Uv(t) \\
 &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} v(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \Lambda v(t - \Delta) - \mu \cdot v(t). \tag{29}
 \end{aligned}$$

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

$$\begin{aligned}
 \dot{v}_i(t) &= \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)g \frac{\lambda}{(2k+1)\mu}} v_i(t - \Delta) - \frac{\lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \Lambda_{ii} \cdot v_i(t - \Delta) - \mu \cdot v_i(t) \\
 &= \frac{(2k+1 - \Lambda_{ii}) \cdot \lambda \cdot g^0 \frac{\lambda}{(2k+1)\mu}}{(2k+1)^2 g \frac{\lambda}{(2k+1)\mu}} \cdot v_i(t - \Delta) - \mu \cdot v_i(t), \tag{30}
 \end{aligned}$$

where Λ_{ii} is the i th diagonal entry of the matrix Λ . One crucial observation is that this representation shows that the system of delay equations given in Eq. (30) are uncoupled and can be analyzed individually

for stability purposes. To finish the proof, we observe that it only remains to analyze the stability of the each equation for $v_i(t)$. To do this we make the ansatz $v_i(t) = e^{rt}$ and derive an equation for the variable r . This yields the following transcendental equations for r

$$r = \alpha_i \cdot e^{-r\Delta} - \mu. \quad (31)$$

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

$$\cos(\omega\Delta) = \frac{\mu}{\alpha_i} \quad (32)$$

$$\sin(\omega\Delta) = -\frac{\omega}{\alpha_i}. \quad (33)$$

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

$$\cos^2(\omega\Delta) + \sin^2(\omega\Delta) = 1 = \frac{(\mu^2 + \omega^2)}{\alpha_i^2}. \quad (34)$$

By moving all terms of Eq. (34) 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

$$\omega = \frac{\alpha_i}{\alpha_i^2 - \mu^2}. \quad (35)$$

Using this expression for ω , we can finally invert Eq. (32) since it does not contain ω on the right-hand side unlike Eq. (33) to solve for the critical value of Δ . We find that our threshold Δ is equal to

$$\Delta_{cr}^i(\lambda, \mu, N, k) = \frac{\arccos \frac{\mu}{\alpha_i}}{\frac{\alpha_i}{\alpha_i^2 - \mu^2}}. \quad (36)$$

Thus our proof is complete.

Theorem 3 provides a 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 equilibrium point $\frac{\lambda}{\mu N}$ and not oscillate around the equilibrium.

2.4. Numerical results for fluid limits

In this section, we describe some numerical results that compare the fluid limits before and after the critical delay values. On the left of Fig. 2, we plot

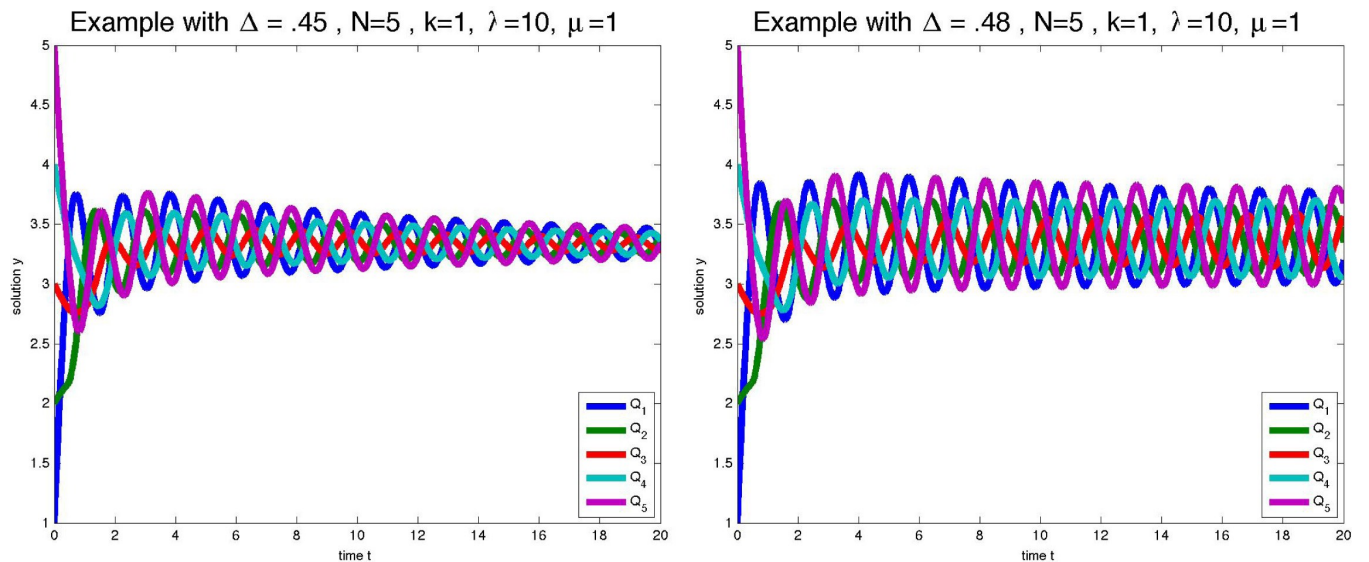


Fig. 2. $\Delta_{cr} = 0.468$.

J. Pender

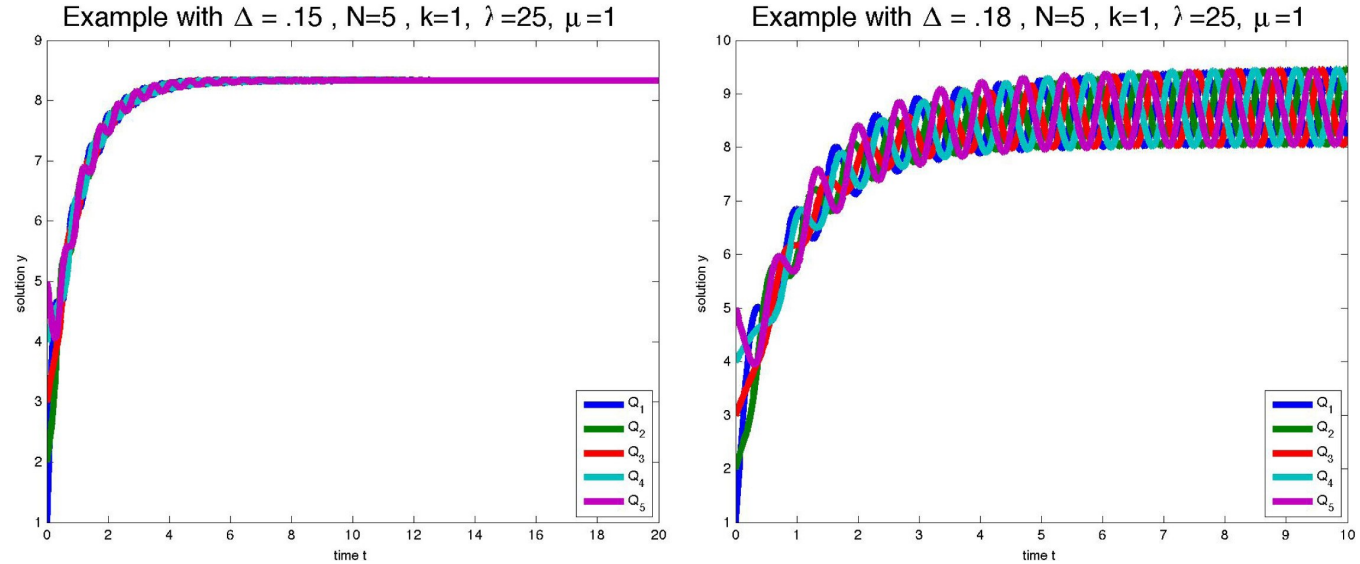


Fig. 3. $\Delta_{cr} = 0.167$.

an example of $N = 5$ queues and we let $k = 1$. Since $\Delta < \Delta_{cr} = 0.468$, we observe that the queues converge to the equilibrium and the initial oscillations vanish over time. However, on the right of Fig. 2, we observe that because $\Delta > \Delta_{cr} = 0.468$, the queues oscillate around the equilibrium. Moreover, it is important to note that even though each queue length oscillates around the equilibrium, the size of the oscillations are not all equal. Thus, there is an asymmetry in the size of the oscillations.

In Fig. 3, we plot an example of $N = 5$ queues and we let $k = 1$, but we increase the arrival rate. Since $\Delta < \Delta_{cr} = 0.167$, we observe that the queues converge to the equilibrium and the initial oscillations vanish over time. However, on the right of Fig. 3, we observe that because $\Delta > \Delta_{cr} = 0.167$, the queues oscillate around the equilibrium. Moreover, it is important to note that even though each queue length oscillates around the equilibrium, the size of the oscillations are all actually equal this time. Thus, the size of the oscillations appears to be symmetric. This is a clear difference from the previous figures.

3. 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 in a k -nearest neighbor fashion. Our model considers the customer choice as a generalized multinomial logit choice model where the queue length

information given to the customer is delayed by an amount of size Δ . We prove fluid limit theorems for our queueing process and show that the fluid limit is a system of delay differential equations. We also prove that the resulting fluid limit can experience a Hopf bifurcation. Using the properties of circulant matrices, we compute exactly when this Hopf bifurcation occurs in terms of our queueing model parameters and verify numerically our results. Although we consider an infinite server system, the analysis of multiserver queues like the Erlang-A queues can also be analyzed in an identical fashion yielding similar results. In fact, the results are identical to the infinite server case as long as the equilibrium does not linger around the number of servers, see for example [Pender & Ko, 2017; Ko & Pender, 2018].

There are many remaining questions for research. For one, what if the delay were nonstationary and not a constant? What if the topology of the network is more complicated? What if each queue has a different delay? We intend to answer these important questions in future work.

Acknowledgments

Jamol Pender acknowledges the generous support of NSF Career Grant CMMI # 1751975.

References

Aghajani, R., Li, X. & Ramanan, K. [2015] ‘‘Mean-field dynamics of load-balancing networks with

- general service distributions,” arXiv preprint arXiv:1512.05056.
- Aghajani, R. & Ramanan, K. [2017] “The hydrodynamic limit of a randomized load balancing network,” arXiv preprint arXiv:1707.02005.
- Aghajani, R., Li, X. & Ramanan, K. [2017] “The pde method for the analysis of randomized load balancing networks,” *Proc. ACM Measur. Anal. Comput. Syst.* 1, 38.
- Atar, R. & Lipschutz, D. [2021] “Heavy traffic limits for join-the-shortest-estimated-queue policy using delayed information,” *Math. Operat. Res.* 46, 268–300.
- Banerjee, S., Mukherjee, D. *et al.* [2019] “Join-the-shortest queue diffusion limit in Halfin–Whitt regime: Tail asymptotics and scaling of extrema,” *Ann. Appl. Probab.* 29, 1262–1309.
- Bramson, M., Lu, Y. & Prabhakar, B. [2010] “Randomized load balancing with general service time distributions,” *ACM SIGMETRICS Performance Evaluation Review* (ACM), pp. 275–286.
- Bramson, M., Lu, Y., Prabhakar, B. *et al.* [2013] “Decay of tails at equilibrium for fifo join the shortest queue networks,” *Ann. Appl. Probab.* 23, 1841–1878.
- Budhiraja, A., Mukherjee, D. & Wu, R. [2017] “Supermarket model on graphs,” arXiv preprint arXiv:1712.07607.
- Byers, J. W., Considine, J. & Mitzenmacher, M. [2004] “Geometric generalizations of the power of two choices,” *Proc. Sixteenth Ann. ACM Symp. Parallelism in Algorithms and Architectures* (ACM), pp. 54–63.
- Cybenko, G. [1989] “Dynamic load balancing for distributed memory multiprocessors,” *J. Paral. Distrib. Comput.* 7, 279–301.
- Ding, Y., Nagarajan, M. & Zhang, Z. G. [2020] “Parallel queues with discrete-choice arrival pattern: Empirical evidence and asymptotic characterization,” Available at SSRN 3584880.
- Doldo, P., Pender, J. & Rand, R. [2021] “Breaking the symmetry in queues with delayed information,” *Int. J. Bifurcation and Chaos* 31, 2130027.
- Dong, J., Yom-Tov, E. & Yom-Tov, G. B. [2018] “The impact of delay announcements on hospital network coordination and waiting times,” *Manag. Sci.*
- Faghih-Imani, A., Hampshire, R., Marla, L. & Eluru, N. [2017] “An empirical analysis of bike sharing usage and rebalancing: Evidence from barcelona and seville,” *Trans. Res. Part A: Policy and Practice* 97, 177–191.
- Hale, J. K. [1969] “Ordinary differential equations,” *Pure Appl. Math.* 21.
- Hale, J. K. [1971] “Functional differential equations,” *Analytic Theory of Differential Equations* (Springer), pp. 9–22.
- Hassin, R. & Haviv, M. [2003] *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, Vol. 59 (Springer Science & Business Media).
- Hassin, R. [2016] *Rational Queueing* (Chapman and Hall/CRC).
- Ko, Y. M. & Pender, J. [2018] “Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes,” *Stoch. Mod.* 34, 186–206.
- Lipschutz, D. & Williams, R. J. [2015] “Existence, uniqueness, and stability of slowly oscillating periodic solutions for delay differential equations with nonnegativity constraints,” *SIAM J. Math. Anal.* 47, 4467–4535.
- Lu, Y., Xie, Q., Kliot, G., Geller, A., Larus, J. R. & Greenberg, A. [2011] “Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services,” *Perform. Eval.* 68, 1056–1071.
- Mitzenmacher, M. [2000] “How useful is old information?” *IEEE Trans. Paral. Distrib. Syst.* 11, 6–20.
- Mitzenmacher, M. [2001] “The power of two choices in randomized load balancing,” *IEEE Trans. Paral. Distrib. Syst.* 12, 1094–1104.
- Mitzenmacher, M. [2016] “Analyzing distributed join-idle-queue: A fluid limit approach,” *2016 54th Ann. Allerton Conf. Communication, Control, and Computing* (Allerton) (IEEE), pp. 312–318.
- Mukherjee, D., Borst, S., Van Leeuwen, J. & Whiting, P. [2016] “Universality of power-of-d load balancing schemes,” *ACM SIGMETRICS Perform. Eval. Rev.* 44, 36–38.
- Mukherjee, D., Dhara, S., Borst, S. C. & van Leeuwen, J. S. [2017] “Optimal service elasticity in large-scale distributed systems,” *Proc. ACM Measur. Anal. Comput. Syst.* 1, 25.
- Mukherjee, D., Borst, S. C., Van Leeuwen, J. S. & Whiting, P. A. [2018] “Universality of power-of-d load balancing in many-server systems,” *Stoch. Syst.* 8, 265–292.
- Nirenberg, S., Daw, A. & Pender, J. [2018] “The impact of queue length rounding and delayed app information on disney world queues,” *2018 Winter Simulation Conf. (WSC)* (IEEE), pp. 3849–3860.
- Novitzky, S., Pender, J., Rand, R. H. & Wesson, E. [2020] “Limiting the oscillations in queues with delayed information through a novel type of delay announcement,” *Queueing Syst.* 95, 281–330.
- Pender, J. & Ko, Y. M. [2017] “Approximations for the queue length distributions of time-varying many-server queues,” *INFORMS J. Comput.* 29, 688–704.
- Pender, J., Rand, R. H. & Wesson, E. [2017] “Queues with choice via delay differential equations,” *Int. J. Bifurcation and Chaos* 27, 1730016.
- Pender, J., Rand, R. & Wesson, E. [2020] “A stochastic analysis of queues with customer choice and delayed information,” *Math. Operat. Res.* 45, 1104–1126.

J. Pender

Schuijbroek, J., Hampshire, R. C. & Van Hoes, W.-J. [2017] “Inventory rebalancing and vehicle routing in bike sharing systems,” *Europ. J. Operat. Res.* 257, 992–1004.

Xie, Q., Dong, X., Lu, Y. & Srikant, R. [2015] “Power of choices for large-scale bin packing: A loss model,” *ACM SIGMETRICS Perform. Eval. Rev.* 43, 321–334.

Appendix

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

Lemma 3 [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*

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

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

Lemma 4 [Kurtz, 1978]. *For any standard Brownian motion $\{W(t)\}_{t \geq 0}$ and any $\epsilon > 0$, $n \in \mathbb{N}$, and $T > 0$*

$$\tilde{M} \equiv \sup_{u, v, \leq nT} \frac{|W(u) - W(v)|}{|u - v|^{1 + \log \frac{nT}{|u - v|}}} < \infty \text{ a.s.}$$

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

Proof

$$Q_i^\eta(t) = Q_i^\eta(0) + \frac{1}{\eta} \Pi_i^a \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \frac{1}{\eta} \Pi_i^d \int_0^t \mu Q_i^\eta(s) ds. \quad (\text{A.1})$$

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

$$\begin{aligned} Q_i^\eta(t) - q_i(t) &= Q_i^\eta(0) - q_i(0) + \frac{1}{\eta} \Pi_i^a \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \int_0^t \lambda \cdot p_i(q(s), \Delta) ds \\ &\quad - \frac{1}{\eta} \Pi_i^d \int_0^t \mu Q_i^\eta(s) ds + \int_0^t \mu q_i(s) ds \\ &= Q_i^\eta(0) - q_i(0) + \frac{1}{\eta} \Pi_i^a \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ &\quad \times \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \int_0^t \lambda \cdot p_i(q(s), \Delta) ds \\ &\quad - \frac{1}{\eta} \Pi_i^d \int_0^t \mu Q_i^\eta(s) ds + \int_0^t \mu Q_i^\eta(s) ds - \int_0^t \mu Q_i^\eta(s) ds + \int_0^t \mu q_i(s) ds. \end{aligned}$$

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.

$$\begin{aligned} |Q_i^\eta(t) - q_i(t)| &\leq |Q_i^\eta(0) - q_i(0)| + \frac{1}{\eta} \Pi_i^a \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ &\quad + \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \int_0^t \lambda \cdot p_i(q(s), \Delta) ds \\ &\quad + \left| \frac{1}{\eta} \Pi_i^d \int_0^t \mu Q_i^\eta(s) ds - \int_0^t \mu Q_i^\eta(s) ds + \int_0^t \mu Q_i^\eta(s) ds - \int_0^t \mu q_i(s) ds \right|. \end{aligned}$$

By Lemma 3, we have the following strong approximation representation of the queue length as

$$\begin{aligned} Q_i^\eta(t) = & \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ & - \frac{1}{\sqrt{\eta}} \int_0^t B_i^a \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ & - \int_0^t \mu Q_i^\eta(s) ds - \frac{1}{\sqrt{\eta}} \int_0^t B_i^d \mu Q_i^\eta(s) ds \\ & + O\left(\frac{\log \eta}{\eta}\right). \end{aligned} \quad (\text{A.2})$$

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

$$\begin{aligned} |Q_i^\eta(t) - q_i(t)| \leq & |Q_i^\eta(0) - q_i(0)| \\ & + \frac{1}{\sqrt{\eta}} \int_0^t B_i^a \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ & + \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ & - \int_0^t \lambda \cdot p_i(q(s), \Delta) ds \\ & + \frac{1}{\sqrt{\eta}} \int_0^t B_i^d \mu Q_i^\eta(s) ds \\ & + \int_0^t \mu Q_i^\eta(s) ds - \int_0^t \mu q_i(s) ds \\ & + O\left(\frac{\log \eta}{\eta}\right). \end{aligned}$$

Now it remains to show that

$$\lim_{\eta \rightarrow \infty} \sup_{t \leq T} \frac{1}{\sqrt{\eta}} \int_0^t B_i^a \lambda \cdot p_i(Q^\eta(s), \Delta) ds = 0 \quad (\text{A.3})$$

and

$$\lim_{\eta \rightarrow \infty} \sup_{t \leq T} \frac{1}{\sqrt{\eta}} \int_0^t B_i^d \mu Q_i^\eta(s) ds = 0. \quad (\text{A.4})$$

For the first Brownian motion term we have that

$$\begin{aligned} \lim_{\eta \rightarrow \infty} \sup_{t \leq T} \frac{1}{\sqrt{\eta}} \int_0^t B_i^a \lambda \cdot p_i(Q^\eta(s), \Delta) ds \\ \leq \lim_{\eta \rightarrow \infty} \frac{1}{\sqrt{\eta}} B_i^a (\lambda \cdot T) = \\ \lim_{\eta \rightarrow \infty} B_i^a \cdot \lambda \cdot \frac{1}{\sqrt{\eta}} = 0. \end{aligned}$$

For the second Brownian motion term we have that

$$\begin{aligned} \lim_{\eta \rightarrow \infty} \sup_{t \leq T} \frac{1}{\sqrt{\eta}} \int_0^t B_i^d \mu Q_i^\eta(s) ds \\ \leq \lim_{\eta \rightarrow \infty} \frac{1}{\sqrt{\eta}} B_i^d ((Q^\eta(0) + \lambda) \cdot \mu \cdot T) \\ = \lim_{\eta \rightarrow \infty} B_i^d \cdot (Q^\eta(0) + \lambda) \cdot \mu \cdot \frac{1}{\sqrt{\eta}} = 0. \end{aligned}$$

Thus, for every $\epsilon > 0$ there exists an $\eta^\square$ such that for all $\eta \geq \eta^\square$

$$|Q_i^\eta(0) - q_i(0)| \leq \frac{\epsilon}{4}, \quad (\text{A.5})$$

$$\sup_{t \leq T} \frac{1}{\sqrt{\eta}} \int_0^t B_i^a \lambda \cdot p_i(Q^\eta(s), \Delta) ds \leq \frac{\epsilon}{4}, \quad (\text{A.6})$$

$$\sup_{t \leq T} \frac{1}{\sqrt{\eta}} \int_0^t B_i^d \mu Q_i^\eta(s) ds \leq \frac{\epsilon}{4} \quad (\text{A.7})$$

and

$$O\left(\frac{\log \eta}{\eta}\right) \leq \frac{\epsilon}{4} \quad (\text{A.8})$$

so that we have

$$\begin{aligned} |Q_i^\eta(t) - q_i(t)| \leq & \int_0^t \lambda \cdot p_i(Q^\eta(s), \Delta) ds - \int_0^t \lambda \cdot p_i(q(s), \Delta) ds \\ & + \int_0^t \mu Q_i^\eta(s) ds - \int_0^t \mu q_i(s) ds + 0 \end{aligned}$$

J. Pender

$$\begin{aligned} & \leq \int_0^t |\lambda \cdot p_i(Q_i^\eta(s), \Delta) - \lambda \cdot p_i(q(s), \Delta)| ds \\ & \quad + \int_0^t |\mu Q_i^\eta(s) - \mu q_i(s)| ds + . \end{aligned}$$

Since we assumed that the function $g(x)$ was continuously differentiable with uniformly bounded first derivatives, there exists a constant C such that

$$\begin{aligned} & |Q_i^\eta(t) - q_i(t)| \\ & \leq C \int_0^t \sup_{-\Delta \leq r \leq s} |Q_i^\eta(r) - q_i(r)| ds + \\ & \leq C \cdot \int_0^t \sup_{0 \leq r \leq s} |Q_i^\eta(r) - q_i(r)| ds + \\ & \quad + C \cdot t \cdot \sup_{-\Delta \leq r \leq 0} |Q_i^\eta(r) - q_i(r)| + . \end{aligned} \tag{A.9}$$

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

$$|Q_i^\eta(t) - q_i(t)| \leq C \int_0^t \sup_{0 \leq r \leq s} |Q_i^\eta(r) - q_i(r)| ds + . \tag{A.10}$$

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, 1969], we have that

$$\sup_{0 \leq t \leq T} |Q_i^\eta(t) - q_i(t)| \leq . \cdot e^{CT} \tag{A.11}$$

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