Queueing Systems (2020) 95:281–330
https://doi.org/10.1007/s11134-020-09657-9

Limiting the oscillations in queues with delayed information through a novel type of delay announcement

Sophia Novitzky[1] · Jamol Pender[2] · Richard H. Rand[3] · Elizabeth Wesson[4]

Received: 3 July 2019 / Revised: 21 March 2020 / Published online: 16 June 2020

Abstract
Many service systems use technology to notify customers about their expected waiting times or queue lengths via delay announcements. However, in many cases, either the information might be delayed or customers might require time to travel to the queue of their choice, thus causing a lag in information. In this paper, we construct a neutral delay differential equation model for the queue length process and explore the use of *velocity* information in our delay announcement. Our results illustrate that using velocity information can have either a beneficial or detrimental impact on the system. Thus, it is important to understand how much velocity information a manager should use. In some parameter settings, we show that velocity information can eliminate oscillations created by delays in information. We derive a fixed point equation for determining the optimal amount of velocity information that should be used and find closed-form upper and lower bounds on its value. When the oscillations cannot be eliminated altogether, we identify the amount of velocity information that minimizes the amplitude of the oscillations. However, we also find that using too much velocity information can create oscillations in the queue lengths that would otherwise be stable.

Keywords Neutral delay-differential equation · Hopf bifurcation · Perturbations method · Operations research · Queueing theory · Fluid limits · Delay announcement · Velocity

Mathematics Subject Classification 34K40 · 34K18 · 41A10 · 37G15 · 34K27

1 Introduction

Many corporations and services eagerly adopt new technologies that allow service managers to interact with their customers. One highly important aspect of the communication is the delay announcement, which informs the customers of their estimated waiting time or queue length. Delay announcements are used in variety of service sys-

Extended author information available on the last page of the article

tems. For example, some hospital emergency rooms display their expected waiting times online. Telephone call centers warn the customers that are placed on hold about extensive waiting times. Amusement parks update the waiting times for different rides, and some transportation networks warn about heavy traffic or delays via road signs.

The popularity of delay announcements among service managers extends beyond customer satisfaction. Informing customers about their waiting times allows managers to influence customer decisions, and the overall system dynamics, which are crucial to a company's productivity and underlying revenue. As a result, many important questions arise when a service decides to implement a delay announcement for its customers: What type of information, if any, should the service provider give to customers? Are there circumstances when the delay announcement hurts the service provider? How long does it take for the service provider to calculate the delay announcement and disseminate it to customers?

The existing literature explores different ways to give a delay announcement, as well as different response behaviors of the customers. For example, Ibrahim et al. [21] study a service system with applications to telephone call centers. Upon calling and receiving the delay announcement, customers have the option to join the queue, balk (leave immediately), or abandon the queue after spending some time waiting. The authors develop methods for determining the accuracy of the last-to-enter-service (LES) delay announcement, which estimates the waiting time for an incoming customer as the waiting time of the most recent customer who entered service. Under the same options for customer behavior, Jouini et al. [23] consider providing different percentiles of the waiting time distribution as information to their customers. They determine the amount of information that maximizes the number of customers who end up receiving service. Armony and Maglaras [3] model a queue where, upon arrival, the customers are told the steady-state expected waiting time and in addition are given an option to request a call back. The authors propose a staffing rule that picks the minimum number of service agents that satisfies a set of operational constraints on the performance of the system. Guo and Zipkin [15,16] allow customers to either join the queue or balk, when the customers are presented either with no information, partial information, or full information about the queues. The authors discuss for what situations the extra information is beneficial, and when the addition information can hurt the customers or the service provider.

This paper also explores the impact of the delay announcement on the dynamics of the queueing process. However, the current literature focuses only on services that give the delay announcements to their customers in real-time, while we consider scenarios when the information itself is delayed. Lags in information are common in services that inform their customers about the waiting times prior to customers' arrival to the service. Such services are prominent in the context of hospital emergency rooms, highway transportation, amusement park rides, and internet buffer sizing [1,9,28,33]. One specific example is the Citibike bike-sharing network in New York City [14,37]. Riders can search the availability of bikes on a smartphone app, as shown in Fig. 1. However, in the time that it takes for the riders to leave their home and get to a station, all of the bikes could have been taken from that station. Thus, the information they used is delayed and is somewhat unreliable by the time they arrive at the station.

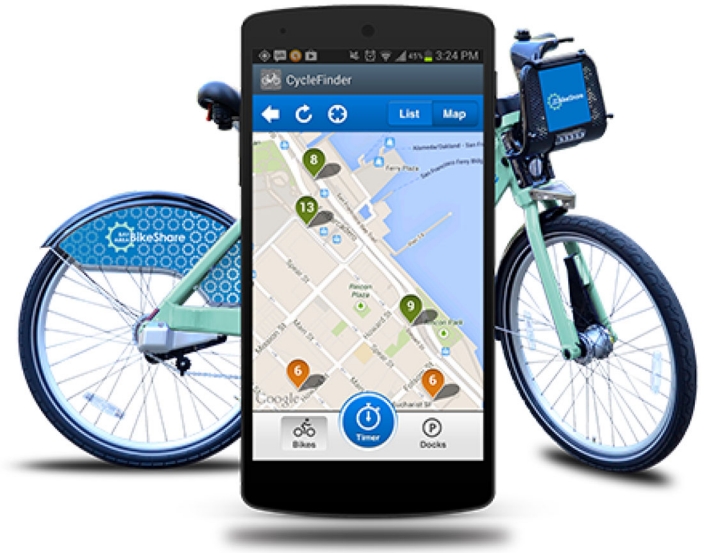

Fig. 1 Bike sharing network app

In this paper, we present a deterministic fluid-like model of queues. This may seem a counter-intuitive choice given that queues are usually comprised of discrete units such as the number of people, jobs, or automobiles. However, a queueing system with a heavy traffic flow can be well approximated by a fluid model. These approximations are common in queueing literature [4,21], where the queues are modeled as stochastic processes and then shown to converge in some limiting regime to deterministic equations. Fluid models are especially useful in settings like manufacturing systems, traffic networks, cloud-computing jobs, busy call centers, and crowds at Disneyland parks where the demand for service is large [2,4,20,28,32].

1.1 Contributions of paper

We present a fluid model of N queues where customers choose which queue to join, giving preference to the shorter queue based on delayed information. Similar models were previously considered by the authors in [29–31]. The size of delay in information determines whether the queues approach a stable equilibrium or Hopf bifurcations occur and the queues oscillate indefinitely. The threshold at which the queues become unstable can be affected by the type of information that is revealed to customers. This paper analyzes what kind of information the service managers should provide to customers in order to distribute the workload evenly among the queues. This benefits both the customers who will avoid excessive waits at the longer queues and the servers who will avoid being overworked or underworked. In many settings, the operator knows not only the current queue lengths, but also the rate at which the queues are changing, namely the queue velocity.

- We develop a new queueing model, where customers are told a weighted sum of the queue length and the queue velocity.
- We show that the queueing system can undergo a Hopf bifurcation if the delay due to the customer travel time is large. We derive the exact point where the Hopf bifurcation occurs.
- We specify how the weight coefficient of velocity information should be chosen so that queues can maintain their stability under greater lags in time. Specifically, we prove that there exists an optimal weight that maximizes the delay where the bifurcation occurs.
- We derive a fixed point equation for the optimal weight, as well as closed-form expressions for upper and lower bounds on that weight. We also provide upper and lower bounds on the maximum delay where the bifurcation occurs.
- When the oscillations in queues cannot be prevented, we use the second-order approximation of amplitude via Lindstedt's method to determine the weight of velocity information that minimizes the amplitude of oscillations.
- When the weights are chosen inadequately, the velocity information can be harmful to the system. We specify the threshold for the weight coefficient where the adverse effects take place.

1.2 Organization of paper

The remainder of the paper is organized as follows: Section 3 presents a mathematical model for N queues and describes the qualitative behavior of the queueing system. In particular, we prove the existence and uniqueness of the equilibrium and give conditions under which the equilibrium is locally stable. We show that for certain values of the parameters, infinitely many Hopf bifurcations may occur.

For some parameters, the queues converge to an equilibrium for sufficiently small delay in information, but as the delay exceeds a certain threshold Δ_{cr}, the equilibrium becomes unstable. Section 4 discusses how the velocity information affects Δ_{cr}, and since the queues are stable only when the delay is less than Δ_{cr}, it becomes our objective to maximize the threshold delay to provide.

Section 5 considers a queueing system with two queues, which is a special case of our N-queue model. We prove that all Hopf bifurcations are supercritical. We use a perturbations technique to develop a highly accurate approximation of the amplitude near the bifurcation point and show that the amplitude of oscillations in queues can be decreased with the right choice of the velocity information weight parameter.

2 Literature review

In this section, we provide a review of the literature that is relevant to this work as the delayed information space is relatively new in the context of this work.

The first paper [30] analyzed a similar model to the one presented in this paper. This particular paper considered a fluid model of two queues and derived an explicit formula for the Hopf bifurcation under the setting of two queues. Our current work

differs in many ways from this first paper. First, our paper considers an N-dimensional system, which is non-trivial. One reason is that the two-dimensional system of delay differential equations reduces to a one dimensional delay differential equation, which follows from the fact that the sum of all of the queues is an infinite server system that is always stable. Thus, a reduction of N to $N-1$ queues is always possible with this symmetry. However, when $N > 2$, it is unclear how to make this reduction to a one dimensional delay differential equation. The second reason that our work differs from previous work is that this work considers using the velocity of the queue length as information, which was not previous considered in any previous work.

The second paper, Pender et al. [31], also analyzed a similar model to the first paper; however, the second paper added the complexity of non-stationary arrival rates. This is a significant difference since non-stationary arrival rates are much more complicated than their stationary counterparts. Moreover, the essential insight from that paper was that non-stationary arrival will affect the dynamics of the queue, but does not impact the location of the Hopf bifurcation unless the frequency of the arrival rate function is twice that of the oscillations generated by the Hopf bifurcation itself. The paper Pender et al. [31] also proves how the Hopf bifurcation is shifted when the frequencies align properly. This paper does not consider non-stationary arrival rates and also incorporates velocity, which is quite different.

The third paper, Novitzky et al. [29], is also different from our current work in that the paper does not incorporate velocity into the announcement feature of the model. Moreover, the essential insight in that work is that we can develop a statistical method to compute the amplitude of the oscillations. We call this method the "Slope Function Method." The idea is that one can numerically integrate a few delay differential equations and compute the amplitude for those equations. Then we learn, exploiting nonlinear regression techniques, how the amplitude changes as a function of the model parameters. We show in the paper how the Slope Function Method accurately predicts the amplitude from the initial data.

There are a few papers that have recently explored the possibility of adding delays to fluid models in queueing systems. One paper that arises is the work of Lipshutz and Williams [26]. In this paper, the authors derive sufficient conditions for when oscillations will occur in reflected delay differential equations when they are present in the non-reflected system. One difference is that we do not consider reflected delay differential equations. Another difference is that our queueing model is inherently multi-dimensional, while the model in Lipshutz and Williams [26] is one-dimensional. Finally, we are interested in computing exactly where the Hopf bifurcation occurs, while Lipshutz and Williams [26] does not compute exactly when this bifurcation will occur in the reflected model. The second paper is by Raina and Wischik [33]. This particular paper combines concepts from queueing theory with delay differential equations and applies them to sizing router buffers in Internet infrastructure services. This paper analyzes the amplitude of oscillations that are a function of the delay and the model parameters. However, they do not compute closed-form expressions for the amplitude and only provide numerical examples in this regard. Our paper here provides explicit formulas for the amplitude in terms of the model parameters and also uses Lindstedt's method to second order, where Raina and Wischik [33] only does first order approximations.

The similar models that were previously considered by the authors in [29–31] have some important common themes. All of these papers are interested in the question of when Hopf bifurcations will occur as a function of model parameters. This is important to understand as it gives managers the ability to understand when oscillations will occur in their systems. In addition, the papers are also interested in knowing what the size and frequency of oscillations will be when they occur. This is important because it is helpful to know not only that oscillations will occur, but also how large they might be or how frequent they might be.

3 The queueing model

Customers arrive at a rate $\lambda > 0$ to a system of N queues, where they are given information about the waiting times at each queue based on the current queue length and the rate at which the queue is changing. Each customer chooses one of N queues to join, giving probabilistic preference to the shorter queue. In this work, we assume that the information that a customer receives is lagged by a quantity Δ. One way to view this delay in information is that the mechanism that provides the information to customer is delayed because it may take time to compute the information. Another perspective is that customers might experience a delay because they have to travel to a queue after committing to one. In this second setting, customers travel for $\Delta > 0$ time units to reach the queue of their choice. Our model assumes that the queueing dynamics for the departure process are identical to that of an infinite-server queue with service rate $\mu > 0$.

Infinite-server queues are quite common in the operations research literature [11–13,22,24,34] as they provide lower bounds on performance on multi-server queueing systems. A reasonable question a reader might ask is why we do not consider multi-server queues in this particular paper since customers in infinite-server queues do not have to wait for service. This is a fair question to ask; however, we have three valid reasons to limit this first analysis to the infinite-server assumption. The first reason that we ignore the multi-server case is that this work is the first work to the authors' knowledge to consider the velocity of queueing systems to construct delay announcements and, as one can see later, this case is already quite complicated.

The second reason stems from the fact that the Erlang-C fluid model basically behaves identically to that of an infinite-server queueing model when the staffing level is operated in the quality regime. More importantly, from a fluid model perspective, infinite-server queues and multi-server queues are identical when $\lambda < \mu C$, which is the same condition needed for stability of the queueing system in steady state. Finally, We want to emphasize that our infinite-server model, which we study in this paper, is vital to develop a better understanding of more complex models like the the Erlang-A with delayed information. The main reason is that, in equilibrium, for an Erlang-A model with C servers and an abandonment rate of β, the fluid limit either behaves like an infinite-server queue with service rate μ when $\lambda < \mu c$ or it behaves like an infinite-server queue with service rate β when $\lambda > \mu c$. Thus, the analysis of the infinite-server case is vital and **suffices** to conduct further analysis for more complicated models. Moreover, since the departure rate and abandonment rate are

non-differentiable functions, the critical delay analysis that we do later cannot be done for the Erlang-A model when the arrival rate is equal to the service rate times the number of servers, i.e., $\lambda/N = \mu C$. A deeper generalized stability analysis for non-smooth functions would have to be developed in the dynamical systems community, and we leave these more complex models for future study. Thus, our initial goal was to get an understanding of a simpler model first to get some intuition about other models. For a deeper understanding of why these relationships hold with the Erlang-A model, see, for example, recent work by Daw and Pender [8].

Another reason why we expect the same conclusions to hold in finite-server queueing systems is that we are studying fluid models. Because we are studying fluid models, we can expect that similar results to those derived in this paper would continue to hold in multi-server queues. More specifically, for an Erlang-A fluid model, we have the following nonlinear differential equation:

$$\dot{q}_i(t) = \lambda - \mu(q_i(t) \wedge C) - \beta(q_i(t) - C)^+. \tag{3.1}$$

First note that there is only one point that does not have a derivative in Eq. (3.1); and, that point is $q_i = C$. Outside of this point, Eq. (3.1) is differentiable and, by Daw and Pender [8], it can be shown that the point $q_i = C$ is achieved in steady state if and only if $\lambda = \mu C$. Otherwise, the steady-state queue length is strictly larger or smaller than the number of servers. Since our analysis in the sequel depends on the differentiability of the queue length equations, the same analysis should work as long as we are not at the point $q_i = C$.

Our infinite-server assumption implies that the departure rate for a queue is the service rate μ multiplied by the total number of customers in that queue, and it also implies that the departure rate is a linear function of the queue length process. Figure 2 shows the queueing system for N queues. The queue length for the ith queue is given by

$$\dot{q}_i(t) = \lambda p_i(q_1, \ldots, q_N, \dot{q}_1, \ldots, \dot{q}_N, \Delta) - \mu q_i(t), \quad \forall i \in \{1, \ldots, N\}, \tag{3.2}$$

where the function p_i represents the probability that a customer chooses the ith queue.

In most other work, most customers might receive information about the real-time queue length $q(t)$; however, we know real systems can experience transmission delays and customers might have to travel to join a queue. Thus, in the delayed context a customer may receive a delayed queue length $q(t - \Delta)$. From previous work, Novitzky et al. [29] and Pender et al. [31], we know that the delay in information may cause Hopf bifurcations and oscillations in the queue length processes. Since these Hopf bifurcations occur, it raises the question of whether we might add additional information to the delayed queue length to either make the oscillations smaller or make them disappear altogether. Thus, in this work, we propose that the manager of a service system gives customers a weighted sum of the delayed queue length and the delayed queue length velocity, i.e.,

$$\text{Information about the i}^{th}\text{ queue} = q_i(t - \Delta) + \delta \dot{q}_i(t - \Delta). \tag{3.3}$$

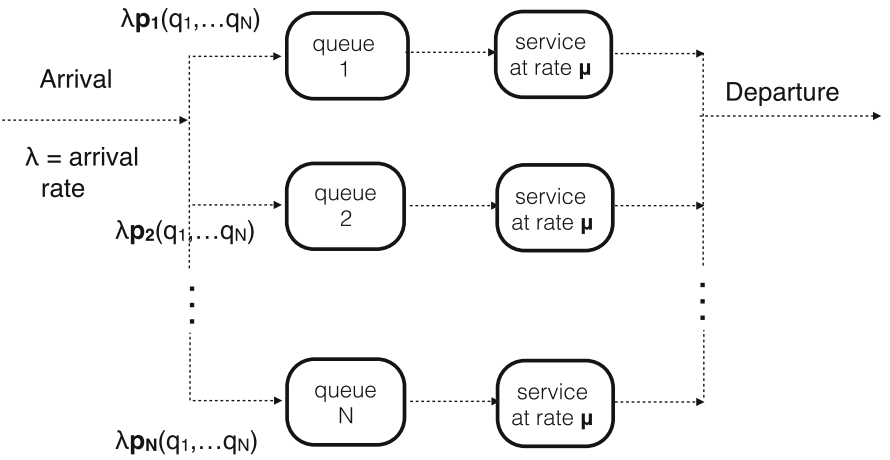

Fig. 2 Customers going through a N-queue service system

This type of delay announcement is really motivated by a Taylor expansion idea. We know from general Taylor expansions that if $q(t)$ is differentiable enough that we can write the queue length process at time t as

$$q(t) = q(t-\Delta) + \sum_{j=1}^{\infty} +\Delta^j q^{(j)}(t-\Delta), \tag{3.4}$$

where we define $q^{(j)}(t-\Delta)$ as the jth time derivative of the queue length process at the time $t-\Delta$. In this work, we only take the first-order expansion term and we provide this to the customer. The first-order term incorporates the velocity of the queue length over time, and as we will show this information can be helpful when trying to reduce the amplitude of oscillations, or in some cases can help us remove them altogether. Thus, the hope is that by giving the customer information from Eq. (3.1), we can reduce the delay in information by the quantity δ, since at first order we have that

$$q(t-(\Delta-\delta)) \approx q(t-\Delta) + \delta \cdot \dot{q}(t-\Delta). \tag{3.5}$$

From a practical perspective, it is important to understand how a manager would collect this velocity information. Even though our fluid model is continuous, in practical situations the queue length and its derivative are generally not continuous functions of time. Thus, our analysis is really intended for systems that are large scale and where fluid models are applicable. Moreover, in practice a manager might choose a time window, say of size ϵ. Then, in this time window a manager would calculate the number of arrivals at the ith queue as $A_i(t) - A_i(t-\epsilon)$ and the number of departures $D_i(t) - D_i(t-\epsilon)$. Using these quantities, one can approximate the time derivative at time t as

$$\dot{q}_i(t) \approx \frac{A_i(t) - A_i(t-\epsilon)}{\epsilon} - \frac{D_i(t) - D_i(t-\epsilon)}{\epsilon}. \tag{3.6}$$

It would be up to the manager to decide what value of ϵ makes the most sense for their calculations. When the number of arrivals is large and ϵ is small, then our approximate time derivative should be expected to be close to the one given by our neutral delay differential equation.

The information provided to the customer should help customers decide which queue to join and, as we will show later in the paper, our new delay announcement can do exactly that when implemented correctly and carefully. In our work, we assume that the probability of a customer choosing the ith queue is given by the Multinomial Logit Model (MNL), which takes in as an input our delay announcement information. The MNL model is commonly used to model customer choice behavior in a variety of fields such as operations research, economics, and applied psychology; see, for example, [19,27,36,38]. Using the MNL model, we have the following expression for the probability that a customer will choose the ith queue to join:

$$p_i(q_1, \ldots, q_N, \dot{q}_1, \ldots, \dot{q}_N, \Delta) = \frac{\exp\Big(-\theta\big(q_i(t-\Delta) + \delta\dot{q}_i(t-\Delta)\big)\Big)}{\sum_{j=1}^{N} \exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_j(t-\Delta)\big)\Big)}, \tag{3.7}$$

where $\theta > 0$ is a standard coefficient of the MNL, $\delta \geq 0$ is the weight of the information about queue's velocity, and $\Delta > 0$ is the delay in time due to customers traveling to service. Obviously, when $\delta = 0$ we revert back to the old model without velocity information and Δ, and the remaining model parameters will fully determine the local stability of the service system.

The parameter θ determines how strong the customer preference is for the shortest queue. For intuition, we will illustrate the MNL model on the simplest model where there are two queues, and the parameters δ, Δ are set to 0. Figure 3 shows the probability of a customer joining the 1st queue, as a function of θ and the difference in queue lengths $q_2 - q_1$,

$$p_1(q_1, q_2) = \frac{\exp(-\theta q_1)}{\exp(-\theta q_1) + \exp(-\theta q_2)} = \frac{1}{1 + \exp\big(-\theta(q_2 - q_1)\big)}. \tag{3.8}$$

When $\theta \to 0$, customers choose queues arbitrarily, giving no preference based on the queue length. Figure 3 indicates $\theta = 0$ by the yellow line, and $p_1 = p_2 = 0.5$ for any difference in queue lengths. When $\theta \to \infty$, customers always choose the shortest queue, even when the difference in lengths is marginal. This is marked by the black line in Fig. 3. However, for simplicity one can set $\theta = 1$, which is denoted by the red curve. In this case, when the queues are roughly of equal length, they will be joined with roughly the same probabilities, but once the difference in the queue lengths increases, the shorter queue will become more preferable.

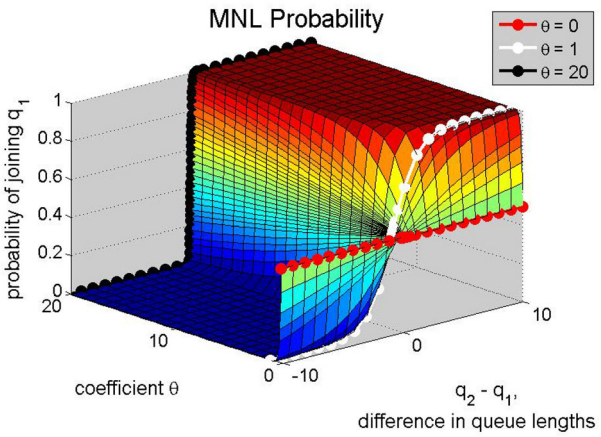

Fig. 3 MNL for two queues

Complete model The incorporation of the probabilities p_i into the queueing system provides a system of neutral delay differential equations (NDDE) for the queue lengths:

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\Big(-\theta\big(q_i(t-\Delta) + \delta\dot{q}_i(t-\Delta)\big)\Big)}{\sum_{j=1}^{N}\exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_i(t-\Delta)\big)\Big)} - \mu q_i(t), \quad \forall i \in \{1,\ldots,N\}, \tag{3.9}$$

with the initial conditions specified by nonnegative continuous functions f_i:

$$q_i(t) = f_i(t), \quad \dot{q}_i(t) = \dot{f}_i(t), \quad t \in [-\Delta, 0]. \tag{3.10}$$

3.1 Conditions for stability and Hopf bifurcations

In this section, we describe the behavior of the queues from Eq. (3.9). We begin by establishing the existence and uniqueness of the solution to the initial value problem (3.9) - (3.10). We note that there exists an extensive analysis of functional differential equations; see, for example, [6,17,26]. The existence and uniqueness of the solution for our specific model directly follows from Driver [10], as stated in the result below.

Theorem 3.1 *Let $f_i(t)$ from Eq. (3.10) be absolutely continuous on $t \in [-\Delta, 0]$, and $\dot{f}_i(t)$ be bounded for almost all $t \in [-\Delta, 0]$ for every $1 \le i \le N$. Then there exists a solution $q_i, \ldots, q_N$ for all $t > 0$ that satisfies Eqs. (3.9)–(3.10). Further, the solution is unique.*

Proof The existence of the solution is given by Theorem 1 of Driver [10]. The uniqueness of the solution follows from Theorem 2 of Driver [10], but we first need to ensure that the conditions of Theorem 2 are fulfilled. The theorem requires that the function

$$\lambda \cdot \frac{\exp\Big(-\theta\big(q_i(t-\Delta)+\delta\dot{q}_i(t-\Delta)\big)\Big)}{\sum_{j=1}^{N}\exp\Big(-\theta\big(q_j(t-\Delta)+\delta\dot{q}_i(t-\Delta)\big)\Big)}-\mu q_i(t),\quad \forall i\in\{1,\ldots,N\}, \tag{3.11}$$

satisfies the local Lipschitz condition with respect to $(q_1(t),\ldots,q_N(t))$ with Lipschitz constant L, where L is a continuous function of $(\dot{q}_1(t-\Delta),\ldots,\dot{q}_N(t-\Delta))$. Here $q_i(t)$ and $q_i(t-\Delta)$ are treated as different variables, so the local Lipschitz condition with respect to $(q_1(t),\ldots,q_N(t))$ is satisfied trivially with $L=2\mu$. Therefore, the solution to the system (3.9)–(3.10) is guaranteed to be unique. □

Theorem 3.2 *The unique equilibrium of $q_i(t)$ from Eq. (3.9) is given by*

$$\lim_{t\to\infty} q_i(t)=q_i^*=\frac{\lambda}{N\mu},\quad 1\le i\le N. \tag{3.12}$$

Proof See the Appendix for the proof. □

The stability of the equilibrium can be determined by the stability of the linearized system of equations [17,35]. Hence, we proceed by linearizing q_i about the equilibrium and finding the characteristic equation.

Proposition 3.3 *The characteristic equation of (3.9) is given by*

$$\Phi(R,\Delta)=-R-\frac{\lambda\theta}{N}\Big(e^{-R\Delta}+\delta Re^{-R\Delta}\Big)-\mu=0. \tag{3.13}$$

Proof We introduce the functions $u_i(t)$ that represent the deviation of $q_i(t)$ from the equilibrium:

$$u_i(t)=q_i(t)-q_i^*=q_i(t)-\frac{\lambda}{N\mu}. \tag{3.14}$$

Once the NDDE are linearized (first order Taylor expansion), $\dot{u}_i(t)$ can be approximated as

$$\begin{aligned}\dot{u}_i(t)\approx&-\frac{\lambda\theta}{N}\Big(u_i(t-\Delta)+\delta u_i'(t-\Delta)\Big)\\&+\frac{\lambda\theta}{N^2}\sum_{j=1}^{N}\Big(u_j(t-\Delta)+\delta u_j'(t-\Delta)\Big)-\mu u_i(t).\end{aligned} \tag{3.15}$$

In vector form, we have

$$\dot{\mathbf{u}}(t)=-\frac{\lambda\theta}{N}\cdot\big(\mathbf{u}(t-\Delta)+\delta\mathbf{u}'(t-\Delta)\big)+\frac{\lambda\theta}{N^2}A\big(\mathbf{u}(t-\Delta)+\delta\mathbf{u}'(t-\Delta)\big)-\mu\mathbf{u}(t), \tag{3.16}$$

where $A \in \mathbb{R}^{N\times N}$, and $A_{ij} = 1$ for $1 \le i, j \le N$. The matrix A can be diagonalized:

$$A = VDM, \quad \text{where} \quad V, D, M \in \mathbb{R}^{N\times N}, \tag{3.17}$$

$$VM = MV = I, \quad D_{ij} = 0 \quad \text{if} \quad i \neq j. \tag{3.18}$$

Since all rows of A are identical, A has only one eigenvalue. This implies that the diagonal matrix D has only one nonzero element, $D_{11} = N$. This property can be exploited with the introduction of a vector $\mathbf{w}(t)$:

$$\mathbf{u}(t) = V\mathbf{w}(t). \tag{3.19}$$

That is an acceptable form of definition because V is invertible. Equation (3.16) becomes

$$V\dot{\mathbf{w}}(t) = -\frac{\lambda\theta}{N} V\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) \tag{3.20}$$

$$+ \frac{\lambda\theta}{N^2} VDMV\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) - \mu V\mathbf{w}(t). \tag{3.21}$$

Pre-multiplying this equation by M yields the following simplification:

$$\dot{\mathbf{w}}(t) = -\frac{\lambda\theta}{N}\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) \tag{3.22}$$

$$+ \frac{\lambda\theta}{N^2} D\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) - \mu\mathbf{w}(t). \tag{3.23}$$

Writing out D explicitly reduces the system of N equations down to just two equations:

$$\dot{w}_1(t) = -\mu w_1(t), \tag{3.24}$$

$$\dot{w}_i(t) = -\frac{\lambda\theta}{N}\big(w_i(t-\Delta) + \delta\dot{w}_i(t-\Delta)\big) - \mu w_i(t), \qquad i \neq 1. \tag{3.25}$$

Equation (3.24) has a solution of the form $w_1(t) = ae^{-\mu t}$, so $w_1 \to 0$ over time. By assuming a solution of the form $w_i(t) = e^{Rt}$, the characteristic equation then follows from (3.25). □

The equilibrium is stable when all eigenvalues R of the characteristic equation have negative real parts. It is evident that any real root R must be negative. However, there are also infinitely many complex roots, and they depend on the delay Δ. When $\delta > \frac{N}{\lambda\theta}$, the equilibrium cannot be stable for any $\Delta > 0$ because there are infinitely many eigenvalues with positive real parts. We demonstrate this in the result below.

Proposition 3.4 *Suppose $\delta > \frac{N}{\lambda\theta}$. Then, for any $\Delta > 0$, there are infinitely many eigenvalues of the characteristic equation that have positive real parts.*

Proof Suppose $\delta > \frac{N}{\lambda\theta}$. We assume $R = a + ib$ with $a, b \in \mathbb{R}$. We can assume $b \geq 0$ without loss of generality. Plugging in R and separating the real and imaginary parts:

$$-(a+\mu)N = e^{-a\Delta}\lambda\theta\Big((1+a\delta)\cos(b\Delta) + b\delta\sin(b\Delta)\Big), \tag{3.26}$$

$$bN = e^{-a\Delta}\lambda\theta\Big(-b\delta\cos(b\Delta) + (1+a\delta)\sin(b\Delta)\Big). \tag{3.27}$$

We find the expressions for sine and cosine:

$$\cos(b\Delta) = -\frac{e^{a\Delta}\Big((a+\mu)(1+a\delta)N + b^2\delta N\Big)}{\lambda\theta\Big((a\delta+1)^2 + (b\delta)^2\Big)}, \tag{3.28}$$

$$\sin(b\Delta) = -\frac{e^{a\Delta}Nb(\delta\mu - 1)}{\lambda\theta\big((1+a\delta)^2 + (b\delta)^2\big)}. \tag{3.29}$$

The identity $\sin^2(b\Delta) + \cos^2(b\Delta) = 1$ gives an expression for b:

$$b = \sqrt{\frac{\lambda^2\theta^2(a\delta+1)^2 - e^{2a\Delta}N^2(a+\mu)^2}{e^{2a\Delta}N^2 - \delta^2\lambda^2\theta^2}}. \tag{3.30}$$

We will now show that there are infinitely many eigenvalues R, where $\mathrm{Re}[R] = a > 0$, by separately considering the cases when $\delta\mu > 1$, $\delta\mu < 1$, and $\delta\mu = 1$.

Case 1 $\delta\mu > 1$. We will construct an interval (a_1, a_2) with $0 < a_1 < a_2$, which contains infinitely many values $\mathrm{Re}[R] = a$ that together with b from Eq. (3.30) satisfy the characteristic equation. We will choose a to be such that both the numerator and the denominator of b are negative, therefore guaranteeing b to be real. This yields two inequalities:

$$\frac{\delta\theta\lambda}{N} > e^{a\Delta} > \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}. \tag{3.31}$$

Since $\frac{\delta\theta\lambda}{N} > 1$ by the assumption that $\delta > \frac{N}{\theta\lambda}$, then the inequality $\frac{\delta\theta\lambda}{N} > e^{a\Delta}$ holds for $a \in [0, a_2)$, where $a_2 = \frac{1}{\Delta}\ln(\frac{\delta\theta\lambda}{N}) > 0$. Further, as a increases, the exponent $e^{a\Delta}$ must inevitably outgrow $\frac{\theta\lambda(1+a\delta)}{N(a+\mu)}$, so there exists $a_1 \geq 0$ such the second part of the inequality from Eq. (3.31) holds for all $a \geq a_1$. Lastly, note that the condition $\frac{\delta\theta\lambda}{N} > \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}$ holds for all $a \geq 0$ because $\delta\mu > 1$, so we can choose a_1 to be less than a_2, i.e., $a_1 \in (0, a_2)$. This shows that there exists an interval (a_1, a_2) with $0 < a_1 < a_2$ where the inequalities from Eq. (3.31) hold, so by Eq. (3.30) we have $0 \neq b \in R$ for all $a \in (a_1, a_2)$.

If $b \in \mathbb{R}$ satisfies Eq. (3.29) for some value of a, then R is an eigenvalue of the characteristic equation. To show that there are infinitely many eigenvalues with real parts in (a_1, a_2), we consider the limit $a \to a_2^-$, when the denominator of b approaches zero and $b \to \infty$. In this limit, the right-hand side of Eq. (3.29) will

oscillate between -1 and 1 an infinite number of times, while the left-hand side of Eq. (3.29) will converge to 0. Hence, there are infinitely many solutions to Eq. (3.29) with $a \in (a_1, a_2)$, and so there are infinitely many eigenvalues with positive real parts.

Case 2 $\delta\mu < 1$. The argument here is analogous to Case 1, except to guarantee that b from Eq. (3.30) is real-valued, we will determine an interval in the range of a where the numerator and the denominator of b are *positive*. We get the condition

$$\frac{\delta\theta\lambda}{N} < e^{a\Delta} < \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}. \tag{3.32}$$

At $a = 0$, $\frac{\theta\lambda(1+a\delta)}{N(a+\mu)} = \frac{\theta\lambda\delta}{N\mu\delta} > \frac{\delta\theta\lambda}{N} > 1$, so there is an interval $[0, a_2)$ for a where $e^{a\Delta} < \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}$ holds. Further, $\frac{\delta\theta\lambda}{N} < \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}$ holds for all $a \geq 0$ because $\delta\mu < 1$, therefore $a_1 = \frac{1}{\Delta}\ln(\frac{\delta\theta\lambda}{N}) > 0$ must be smaller than a_2. Therefore for all $a \in (a_1, a_2)$, with $0 < a_1 < a_2$, $b \in \mathbb{R}$.

Just as in Case 1, when $a \to a_1^+$, $b \to \infty$ so the right-hand side of Eq. (3.29) will oscillate between 1 and -1 infinitely many times, while the left-hand side will converge to 0. Thus, there will be infinitely many eigenvalues that satisfy the characteristic equation (3.13).

Case 3 $\delta\mu = 1$. In this case, the expressions for sine and cosine simplify to

$$\cos(b\Delta) = -\frac{e^{a\Delta}N}{\lambda\theta\delta}, \quad \sin(b\Delta) = 0, \tag{3.33}$$

so $b = (2k-1)\pi/\Delta$ for $k = 1, 2, \ldots$, and $a = \frac{1}{\Delta}\ln(\frac{\lambda\theta\delta}{N})$. Since $\delta > \frac{N}{\theta\lambda}$, then $a > 0$, and the characteristic equation (3.13) has infinitely many eigenvalues with positive real parts. □

However, when the weight coefficient δ is sufficiently small, i.e., $\delta < \frac{N}{\lambda\theta}$, given a sufficiently small delay, the queues converge to a stable equilibrium. As the next result shows, the stability is due to all complex eigenvalues having negative real parts. This result is because when the delay Δ is equal to zero, the queues will converge to the equilibrium and do not oscillate since customers receive information in real-time.

Proposition 3.5 *Suppose $\delta < \frac{N}{\lambda\theta}$. When Δ is sufficiently small, all eigenvalues of the characteristic equation have negative real parts.*

Proof To reach contradiction, let us assume that for any $\Delta_0 > 0$ there exists some $\Delta \in (0, \Delta_0)$ and an eigenvalue $R = a + ib$ with $a \geq 0$ that satisfy the characteristic equation (3.13). We can assume $b \geq 0$ without loss of generality. Plugging in R and separating the real and imaginary parts:

$$-(a+\mu)N = e^{-a\Delta}\lambda\theta\Big((1+a\delta)\cos(b\Delta) + b\delta\sin(b\Delta)\Big), \tag{3.34}$$

$$bN = e^{-a\Delta}\lambda\theta\Big(-b\delta\cos(b\Delta) + (1+a\delta)\sin(b\Delta)\Big). \tag{3.35}$$

Solving for sine and cosine, we find

$$\cos(b\Delta) = -\frac{e^{a\Delta}\Big((a+\mu)(1+a\delta)N + b^2\delta N\Big)}{\lambda\theta\Big((a\delta+1)^2 + (b\delta)^2\Big)}, \tag{3.36}$$

$$\sin(b\Delta) = -\frac{e^{a\Delta}Nb(\delta\mu - 1)}{\lambda\theta\big((1+a\delta)^2 + (b\delta)^2\big)}. \tag{3.37}$$

The identity $\sin^2(b\Delta) + \cos^2(b\Delta) = 1$ gives an expression for b:

$$b = \sqrt{\frac{\lambda^2\theta^2(a\delta+1)^2 - e^{2a\Delta}N^2(a+\mu)^2}{e^{2a\Delta}N^2 - \delta^2\lambda^2\theta^2}}. \tag{3.38}$$

Since $e^{2a\Delta} \geq 1$ and $N > \delta\lambda\theta$ by assumption, the denominator of b is positive, so the numerator of b must be nonnegative. Therefore we get inequalities

$$1 \leq e^{a\Delta} \leq \frac{\lambda\theta(a\delta+1)}{N(a+\mu)}, \quad e^{a\Delta} > \frac{\delta\lambda\theta}{N}. \tag{3.39}$$

From the first inequality, we obtain an upper bound on a, $a \leq \frac{\lambda\theta - N\mu}{N - \lambda\theta\delta}$. If $\lambda\theta < N\mu$, then $a < 0$ so we reached a contradiction. If $\lambda\theta \geq N\mu$, then we use (3.38) and (3.39) to find an upper bound on b:

$$b \leq B = \sqrt{\frac{2N\lambda^2\theta^2(1-\delta\mu)^2}{N^2 - \delta^2\lambda^2\theta^2}}, \quad B > 0. \tag{3.40}$$

However, we note from the cosine equation (3.36) that $\cos(b\Delta) < 0$. Since b is nonnegative then $b\Delta > \frac{\pi}{2}$ so $b > \frac{\pi}{2\Delta}$ for any Δ. Choose $\Delta_0 = \frac{\pi}{4B}$. Then for any $\Delta < \Delta_0$ we get a contradiction with (3.40):

$$b > \frac{\pi}{2\Delta} > \frac{\pi}{2\Delta_0} > 2B. \tag{3.41}$$

Hence, when Δ_0 is sufficiently small, then for any $\Delta < \Delta_0$ the real part of any eigenvalue is negative. □

An interesting edge case, however, is when $\delta = \frac{N}{\lambda\theta}$. If the equality holds, three different behaviors may be observed. The equilibrium will be stable, regardless of the size of the delay, if $\delta\mu > 1$. However, the equilibrium will be unstable if $\delta\mu < 1$. Further, if $\delta\mu = 1$, then the behavior of the queues cannot be determined from the characteristic equation (3.13), as all eigenvalues will be purely imaginary. We justify these findings in the result below. However, before justifying these results, we should mention that a necessary condition for oscillations is that $\lambda\theta \geq N\mu$. This condition was derived in Novitzky et al. [29]. Thus, the results here about the velocity parameter also correspond to the conditions needed for oscillations in the case where $\delta = 0$.

Proposition 3.6 *Suppose* $\delta = \frac{N}{\lambda\theta}$. *If* $\delta\mu < 1$, *then for any* Δ *there exists at least one eigenvalue with positive real part. If* $\delta\mu > 1$, *then all eigenvalues have negative real parts. Further, if* $\delta\mu = 1$ *then all eigenvalues are purely imaginary.*

Proof As in Propositions 3.4–3.5, we express the eigenvalue as $R = a + ib$ and then separate the real and imaginary parts of the characteristic equation. The assumption $\delta = \frac{N}{\lambda\theta}$ simplifies the expressions to be

$$\sin(b\Delta) = -\frac{bNe^{a\Delta}(\mu N - \theta\lambda)}{N^2b^2 + (aN + \theta\lambda)^2}, \tag{3.42}$$

$$\cos(b\Delta) = -\frac{Ne^{a\Delta}\left(a^2N + a\theta\lambda + a\mu N + b^2N + \theta\lambda\mu\right)}{N^2b^2 + (aN + \theta\lambda)^2}. \tag{3.43}$$

We will address the three cases separately.

Case 1 $\delta\mu > 1$. To reach contradiction, suppose there exists an eigenvalue with a nonnegative real part, $a \geq 0$. The expression for b is given by

$$b = \frac{\sqrt{(aN + \theta\lambda)^2 - N^2e^{2a\Delta}(a + \mu)^2}}{\sqrt{N^2\left(e^{2a\Delta} - 1\right)}}, \tag{3.44}$$

where the denominator is positive, so the numerator must be nonnegative for b to be real. Therefore $aN + \theta\lambda - Ne^{a\Delta}(a + \mu) > 0$. However, the assumption $\delta\mu > 1$ is equivalent to $\lambda\theta < N\mu$, so we can show that

$$aN + \theta\lambda - Ne^{a\Delta}(a + \mu) \leq aN + \theta\lambda - N(a + \mu) = \theta\lambda - N\mu < 0, \tag{3.45}$$

and we reached a contradiction. Thus, if $\delta\mu > 1$ then any eigenvalue must have a negative real part.

Case 2 $\delta\mu < 1$. This condition is equivalent to $\lambda\theta < N\mu$. Again, b satisfies Eq. (3.44). As $a \to 0^+$, $b \to \infty$ so $\sin(b\Delta)$ oscillates between 1 and -1 infinitely quickly. Further, as $a \to 0^+$ the right-hand side of Eq. (3.42) goes to zero. Therefore, Eq. (3.42) will have infinitely many roots, while Eq. (3.43) will be satisfied at each root automatically since b is given by Eq. (3.44). Therefore $\delta\mu < 1$ implies that the characteristic equation will have infinitely many eigenvalues with positive real parts.

Case 3 $\delta\mu = 1$. This case is equivalent to the condition $\lambda\theta = N\mu$, which simplifies Eqs. (3.42)–(3.43) to be

$$\sin(b\Delta) = 0, \qquad \cos(b\Delta) = -e^{a\Delta}. \tag{3.46}$$

Hence, $b = (2k + 1)\pi/\Delta$, where $k = 0, 1, 2, \ldots$, and $1 = e^{a\Delta}$ so $a = 0$. Therefore the roots of the characteristic equation (3.13) are purely imaginary. □

Hence, the equilibrium is stable when $\delta = \frac{N}{\lambda\theta}$ and $\delta\mu > 1$, or when $\delta < \frac{N}{\lambda\theta}$ and the delay Δ is sufficiently small. Further, the only way for the equilibrium to become

unstable given that $\delta < \frac{N}{\lambda\theta}$ is if a pair of complex eigenvalues crosses from the negative real side of the complex plane into the positive real side. We will determine the threshold value of delay where the stability of the equilibrium may change by finding where the eigenvalues (if any) on the complex plane reach the imaginary axis.

Proposition 3.7 *The characteristic equation* (3.13) *has a pair of purely imaginary solutions* $R = \pm i\omega_{\text{cr}}$ *with* ω_{cr} *being real and positive, at each root* Δ_{cr}*, given that*

$$\omega_{\text{cr}} = \sqrt{\frac{\lambda^2\theta^2 - N^2\mu^2}{N^2 - \delta^2\lambda^2\theta^2}} \tag{3.47}$$

and Δ_{cr} *satisfies the transcendental equation*

$$\cos\left(\Delta_{\text{cr}}\sqrt{\frac{\lambda^2\theta^2 - N^2\mu^2}{N^2 - \delta^2\lambda^2\theta^2}}\right) = -\frac{\delta\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta\mu)}. \tag{3.48}$$

Proof Assume that R from the characteristic equation (3.13) is purely imaginary, $R = \pm i\omega_{\text{cr}}$. Plugging in R, the real and imaginary parts produce two equations:

$$\mu = -\frac{\lambda\theta}{N}\cos(\omega_{\text{cr}}\Delta_{\text{cr}}) - \frac{\lambda\theta}{N}\delta\omega_{\text{cr}}\sin(\omega_{\text{cr}}\Delta_{\text{cr}}), \tag{3.49}$$

$$\omega_{\text{cr}} = \frac{\lambda\theta}{N}\sin(\omega_{\text{cr}}\Delta_{\text{cr}}) - \frac{\lambda\theta}{N}\delta\omega_{\text{cr}}\cos(\omega_{\text{cr}}\Delta_{\text{cr}}). \tag{3.50}$$

We can solve for the values of the sine and cosine functions, i.e.,

$$\cos(\omega_{\text{cr}}\Delta_{\text{cr}}) = -\frac{N(\mu + \delta\omega_{\text{cr}}^2)}{\lambda\theta(1+\delta^2\omega_{\text{cr}}^2)}, \quad \sin(\omega_{\text{cr}}\Delta_{\text{cr}}) = \frac{N\omega_{\text{cr}}(1-\delta\mu)}{\lambda\theta(1+\delta^2\omega_{\text{cr}}^2)}, \tag{3.51}$$

and by the trigonometric identity $\sin^2(\omega_{\text{cr}}\Delta_{\text{cr}}) + \cos^2(\omega_{\text{cr}}\Delta_{\text{cr}}) = 1$, ω_{cr} is found. The cosine equation from (3.51) then gives the equation for Δ_{cr}. □

Proposition 3.7 provides the infinitely many critical delays Δ_{cr} as well as the necessary conditions on the other parameters ($\omega_{\text{cr}} \in \mathbb{R}, \omega_{\text{cr}} \neq 0$) for when Hopf bifurcations may occur. This information allows us to prove that a Hopf bifurcation occurs at every Δ_{cr}.

Theorem 3.8 *Suppose* ω_{cr} *from Eq.* (3.47) *is real and nonzero. Then a Hopf bifurcation occurs at* $\Delta = \Delta_{\text{cr}}$*, where* Δ_{cr} *is any positive root of*

$$\Delta_{\text{cr}}(\lambda, \mu, \theta, N, \delta) = \arccos\left(-\frac{\delta\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta\mu)}\right)\cdot\sqrt{\frac{N^2 - \delta^2\lambda^2\theta^2}{\lambda^2\theta^2 - N^2\mu^2}}. \tag{3.52}$$

Proof By Proposition 3.7, at each Δ_{cr} there is a pair of purely imaginary eigenvalues $R = i\omega_{\text{cr}}$, $\bar{R} = -i\omega_{\text{cr}}$. A Hopf bifurcation can only occur if $\frac{d}{d\Delta}\text{Re}[R(\Delta_{\text{cr}})] \neq 0$.

To verify this, we assume that $R(\Delta) = \alpha(\Delta) + i\omega(\Delta)$. The characteristic equation (3.13) is differentiated with respect to delay, and we find that at Δ_{cr}, where $\alpha = 0$ and $\omega = \omega_{\text{cr}}$, $\frac{d}{d\Delta}\text{Re}[R]$ is given by

$$\frac{d\alpha}{d\Delta} = \frac{(N^2 - \delta^2\lambda^2\theta^2)(1+\delta^2\omega^2)\omega^2}{\lambda^2\theta^2(1+\delta^2\omega^2)\big((\delta-\Delta)^2+\delta^2\Delta^2\omega^2\big) + N^2\big(1-2\delta\mu+2\Delta\mu+\delta^2\omega^2(2\Delta\mu-1)\big)}. \tag{3.53}$$

The assumption $\omega_{\text{cr}} > 0$ guarantees the numerator of $\frac{d\alpha}{d\Delta}(\Delta_{\text{cr}})$ to be nonzero. To show that the denominator D is nonzero as well, note that it is a quadratic function of Δ, with an absolute minimum at Δ^* such that

$$\frac{dD}{d\Delta}(\Delta^*) = 0 \implies \Delta^* = \frac{\delta\lambda^2\theta^2 - N^2\mu}{\lambda^2\theta^2(1+\delta^2\omega_{\text{cr}}^2)}. \tag{3.54}$$

Once Δ^* and $\omega = \omega_{\text{cr}}$ from Eqs. (3.54) and (3.47) are substituted into the denominator $D(\Delta)$ from Eq. (3.53), we find that the minimum of D with respect to Δ is positive:

$$D(\Delta) \geq D(\Delta^*) = \frac{(N^2-\delta^2\lambda^2\theta^2)(\lambda^2\theta^2-N^2\mu^2)}{\lambda^2\theta^2} = \frac{(N^2-\delta^2\lambda^2\theta^2)^2\omega_{\text{cr}}}{\lambda^2\theta^2} > 0. \tag{3.55}$$

Hence the denominator of $\frac{d\alpha}{d\Delta}(\Delta_{\text{cr}})$ is positive for any delay Δ, so

$$\frac{d\alpha}{d\Delta}(\Delta_{\text{cr}}) \neq 0. \tag{3.56}$$

In fact, if $\delta < \frac{N}{\lambda\theta}$ then $\frac{d\alpha}{d\Delta}(\Delta_{\text{cr}}) > 0$ so the eigenvalues always cross from left to right on the complex plane. If $\delta > \frac{N}{\lambda\theta}$ then $\frac{d\alpha}{d\Delta}(\Delta_{\text{cr}}) < 0$ so the eigenvalues always cross from right to left. At each root of Δ_{cr}, there is one purely imaginary pair of eigenvalues, but all other eigenvalues necessarily have a nonzero real part. Hence all roots $\Lambda_j \neq R, \bar{R}$ satisfy $\Lambda_j \neq mR, m\bar{R}$ for any integer m. Therefore all conditions of the infinite-dimensional version of the Hopf Theorem from Hale and Lunel [17] are satisfied, so a Hopf bifurcation occurs at every root Δ_{cr}. □

Theorem 3.8 provides an explicit expression for the critical delay Δ_{cr}, which implies that if the delay $\Delta > \Delta_{\text{cr}}$, then a Hopf bifurcation occurs and the queues begin to oscillate. Otherwise, the queues will not oscillate and will converge to the unique stable equilibrium. Since we have an explicit expression, we can observe many insights from the expression. The first insight is that the critical delay increases (becomes more stable) as the parameters μ, N, δ are increased, or λ or θ are decreased. This is insightful as this tells us that oscillations are more prevalent when the arrival rate is large and the sensitivity of the queue length is large. Increasing the number of queues has the same effect as decreasing the arrival rate, and this also is true of the service rate. Thus, for large scale systems, it is expected that the oscillations will be more prevalent in these systems than smaller systems. This is certainly in contrast

with much of the queueing literature, where having a large-scale system results in positive gains. Although this seems counter-intuitive at first, we can explain why this occurs. When the number of arrivals is large, more customers are following the wrong information, which generates a wild swing in one direction. However, when the number of customers is small, this wild oscillation does not occur since not as many people are following the wrong information.

In the proof of Theorem 3.8, for $\delta < \frac{N}{\lambda\theta}$ it is shown that any pair of complex eigenvalues which crosses the imaginary axis on the complex plane, necessarily crosses from left to right. The implication here is that once the real part of an eigenvalue becomes positive, it remains positive as the delay increases. This allows us to state conditions for local stability of the equilibrium.

Theorem 3.9 *When $\lambda\theta > N\mu$ and $\delta < \frac{N}{\lambda\theta}$, the equilibrium is locally stable for sufficiently small delay Δ. When either $\lambda\theta \leq N\mu$ and $\delta < \frac{N}{\lambda\theta}$, or $\lambda\theta < N\mu$ and $\delta = \frac{N}{\lambda\theta}$, the equilibrium is locally stable for all Δ.*

Proof If $\delta = \frac{N}{\lambda\theta}$ and $\lambda\theta < N\mu$ then, by Proposition 3.6, for any delay all eigenvalues of the characteristic equation have negative real parts; therefore, the equilibrium is locally stable.

If $\delta < \frac{N}{\lambda\theta}$, then by Proposition 3.5 there exists a sufficiently small Δ such that all eigenvalues of the characteristic equation have negative real parts. The only way for the equilibrium to become unstable is for an eigenvalue to reach the imaginary axis for some Δ. For that to happen, $\omega_{\text{cr}} = \sqrt{\frac{\lambda^2\theta^2 - N^2\mu^2}{N^2 - \delta^2\lambda^2\theta^2}} \in \mathbb{R}$, $\omega_{\text{cr}} \neq 0$ must hold. In the case when $\lambda\theta \leq N\mu$ and $\delta < \frac{N}{\lambda\theta}$, then either $\omega_{\text{cr}} \notin \mathbb{R}$ or $\omega_{\text{cr}} = 0$ so the eigenvalues have negative real parts for all Δ. Therefore, again the eigenvalues have negative real parts for all (finite) Δ. Finally, assume $\delta = \frac{N}{\lambda\theta}$ and $\lambda\theta < N\mu$. Then $\delta\mu = \frac{N\mu}{\lambda\theta} > 1$, so by Proposition 3.6 it follows that all eigenvalues have negative real parts. Therefore, the equilibrium is locally stable for any $\Delta > 0$. □

To summarize, the behavior of the queues from Eq. (3.9) can be categorized into two cases, when $\lambda\theta < N\mu$ and $\lambda\theta > N\mu$. In each case, two different types of behavior can be observed, depending on the size of the parameter δ. Hence, there can be four qualitatively different scenarios, as shown in Fig. 4. In the following discussion of the two cases, we will refer to this diagram and will explain it in detail.

Before explaining Fig. 4 in technical terms, we provide a bit of intuition about the diagram. In Part A of the figure, we have a region where the queues are never stable despite having the condition $\lambda\theta < N\mu$. This occurs because δ is too large and is outside of its stable region. Recall that a first-order Taylor expansion might work well in a small neighborhood of the point around which one is performing the expansion; however, it is not expected to work well significantly far from that point. The same is true here and in fact we lose stability when the velocity term is too large. In Part B, the queues are always stable. Not only is it the same that $\lambda\theta < N\mu$, it is also the case that δ is small enough, i.e., close to the point of expansion. In Part C, not only is the condition $\lambda\theta > N\mu$ true, but we also have that the velocity parameter δ is too large. This will definitely cause oscillations. Finally, in Part D, since the velocity parameter δ is small enough, we behave similarly to the no velocity case and we have oscillations if Δ is

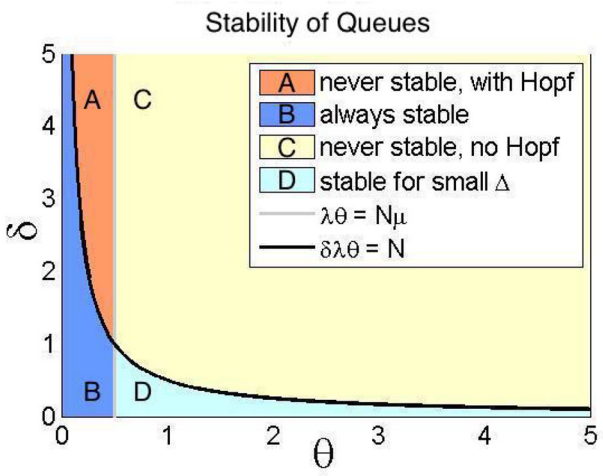

Fig. 4 The four stability cases

large enough as usual. From this partitions, we observe that the velocity parameter has a lot of power in determining the stability of the system. If it is too large in any setting, it can create unwanted oscillations. Thus, it requires a lot of care in choosing this velocity parameter so that the queues behave in the intended way.

Case 1: $\lambda\theta < N\mu$. This case is represented by the regions A and B that are to the left of the vertical line $\lambda\theta = N\mu$ from Fig. 4. When $\delta \leq \frac{N}{\lambda\theta}$, or region B, the queues approach a stable equilibrium for any delay Δ. Here all eigenvalues stay on the negative (real) side of the complex plane. As Δ increases, the complex eigenvalues approach the imaginary axis, but never reach it, as shown in Fig. 5. However, when $\delta > \frac{N}{\lambda\theta}$, which is region A of Fig. 4, the queues will never be stable, and will undergo infinitely many Hopf bifurcations as the delay increases. For sufficiently small delay Δ, the complex eigenvalues will be on the positive (real) side of the complex plane, and as Δ increases, the complex pairs will cross the imaginary axis from right to left, causing Hopf bifurcations to occur as shown in Fig. 6. Note, however, that queues will never gain stability because for any delay Δ there will be eigenvalues with positive real parts.

Case 2: $\lambda\theta > N\mu$. This case is represented by the regions C and D in Fig. 4. When $\delta < \frac{N}{\lambda\theta}$, or region D, the queues will approach a stable equilibrium for a sufficiently small delay Δ. All the eigenvalues will be on the negative (real) side of the complex plane. As the delay Δ increases, the complex pairs of eigenvalues will move towards the imaginary axis, crossing the axis eventually one by one from left to right as indicated in Fig. 7. During the crossing of each pair, a Hopf bifurcation occurs. When $\delta \geq \frac{N}{\lambda\theta}$, which is region C of the Fig. 4, the complex eigenvalues cannot reach the imaginary axis, and they all stay to the right side of the imaginary axis on the complex plane as show in Fig. 8, so there will never be a stable equilibrium.

Another aspect to point out is the dependence on the MNL parameter θ. When customers join the queues at random, or $\theta \to 0$, the parameters inevitably end up in

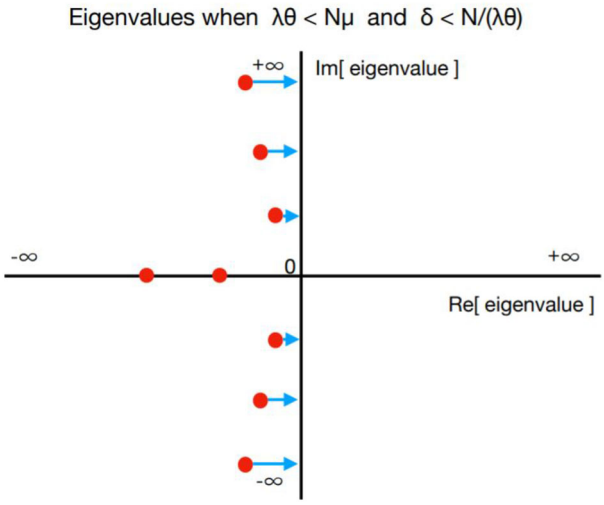

Fig. 5 Eigenvalues remain on the left side of the imaginary axis for all Δ

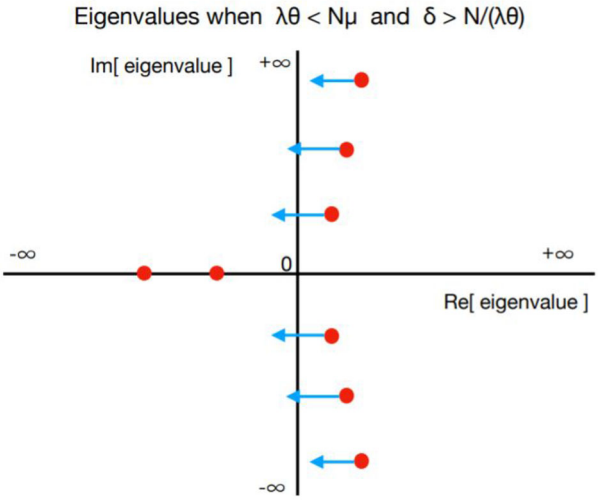

Fig. 6 Eigenvalues cross the imaginary axis from right to left as Δ increases

region B of Fig. 4, so the queues will stable for any delay. Alternatively, if customers always join the shortest queue, or $\theta \to \infty$, then for any $\delta > 0$ we inevitably end up in region C of Fig. 4, so the queues will always be unstable.

4 Achieving maximum stability

In physical settings, it is often important to preserve the stability of the queues. Stability evens out the individual waiting times of the customers, minimizing the negative

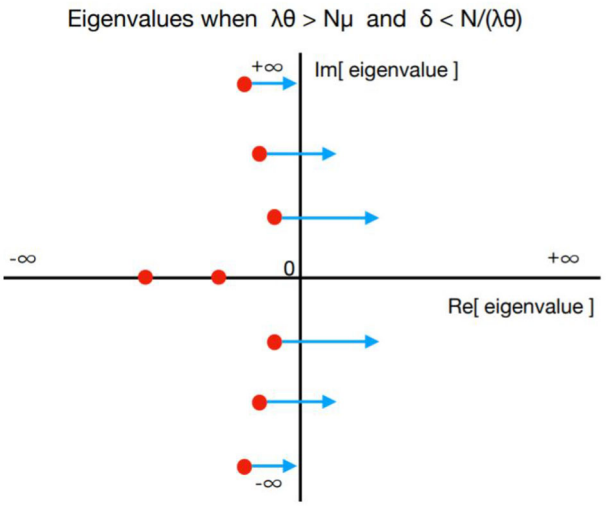

Fig. 7 Eigenvalues cross the imaginary axis from left to right as Δ increases

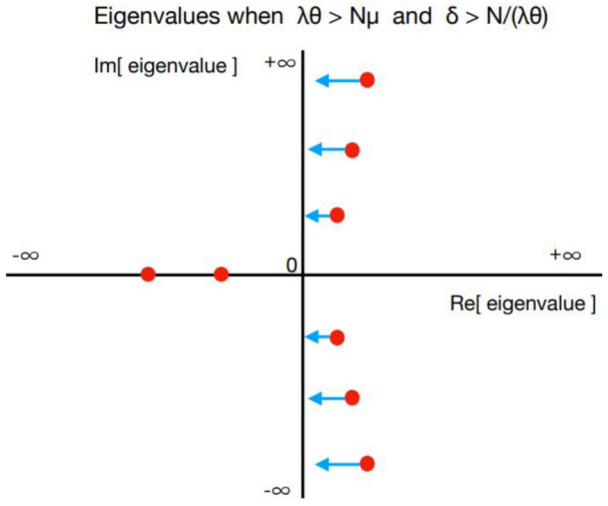

Fig. 8 Eigenvalues stay on the right side of imaginary axis for all Δ

experience. It is therefore useful to know when providing extra information helps to postpone the point of the bifurcation, and when the extra information makes the bifurcation happen sooner. For example, consider the numerical examples from Figs. 9, 10, 11 and 12, with two queues and fixed parameters $\lambda, \theta, \mu,$ and Δ. In Figs. 9 and 11, $\Delta < \Delta_{cr}$ so the queues converge to an equilibrium over time. However, in Figs. 10 and 12 we have $\Delta > \Delta_{cr}$, so the queues oscillate indefinitely. Although the delay Δ is the same, the change in behavior results from tweaking the parameter δ, which consequently regulates the bifurcation threshold Δ_{cr}.

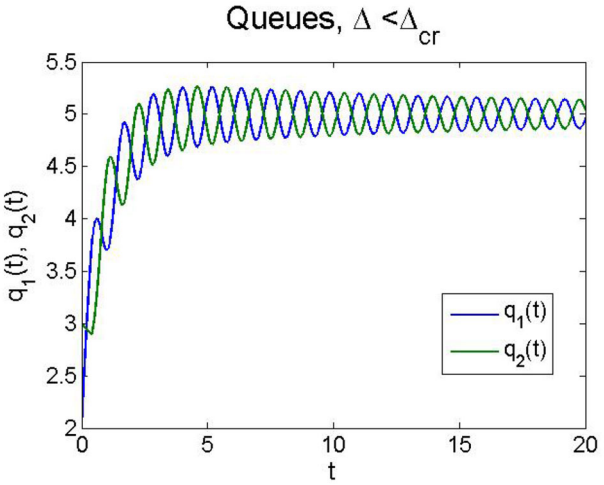

Fig. 9 Queues before Hopf bifurcation; $\delta = 0.08$, $\lambda = 10$, $\mu = 1$, $\theta = 1$

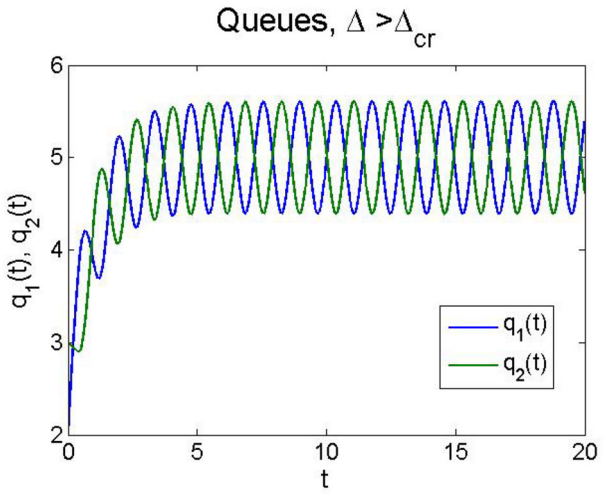

Fig. 10 Queues after Hopf bifurcation; $\delta = 0$, $\lambda = 10$, $\mu = 1$, $\theta = 1$

In this section, we will consider the scenario $\lambda\theta > N\mu$, where the equilibrium of the queues can become unstable. We will study how the bifurcation threshold Δ_{cr} changes depending on the weight of the velocity information δ. Our next result shows that the threshold Δ_{cr} is a concave function of δ.

Proposition 4.1 *Suppose* $\lambda\theta > N\mu$. *Then the function* $\Delta_{\text{cr}}(\delta)$ *is concave for all* $\delta \in [0, \frac{N}{\lambda\theta})$.

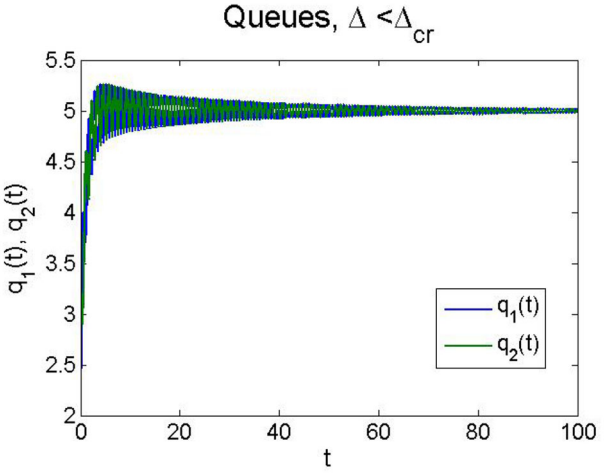

Fig. 11 Queues before Hopf bifurcation; $\delta = 0.08$, $\lambda = 10$, $\mu = 1$, $\theta = 1$

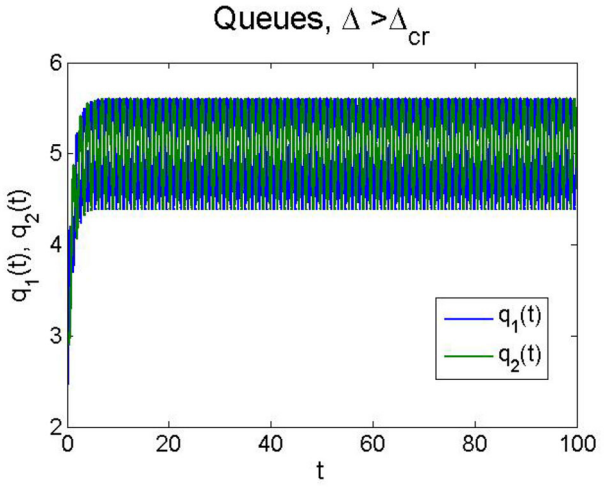

Fig. 12 Queues after Hopf bifurcation; $\delta = 0$, $\lambda = 10$, $\mu = 1$, $\theta = 1$

Proof The critical delay $\Delta_{\rm cr}$ is given by Eq. (3.52). It is clear that the second derivative $\frac{d^2\Delta_{\rm cr}}{d\delta^2}$ is negative for all $\delta \in [0, \frac{N}{\lambda\theta})$:

$$\frac{d^2\Delta_{\rm cr}}{d\delta^2} = -\frac{1}{C_3}\cdot\left(C_1 + C_2 \arccos\left(-\frac{\delta\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+N\delta\mu)}\right)\right), \quad \text{where} \tag{4.1}$$

$$C_1 = (N^2 - \delta^2\lambda^2\theta^2)(\lambda^2\theta^2 - N^2\mu^2)(\delta\lambda^2\theta^2 + N^2\mu) > 0, \tag{4.2}$$

$$C_2 = N^2\lambda^2\theta^2(1+\delta\mu)^2\sqrt{(N^2 - \delta^2\lambda^2\theta^2)(\lambda^2\theta^2 - N^2\mu^2)} > 0, \tag{4.3}$$

$$C_3 = N\lambda\theta(N^2 - \delta^2\lambda^2\theta^2)^{\frac{3}{2}}\sqrt{\lambda^2\theta^2 - N^2\mu^2}(1+\delta\mu)^3\sqrt{1 - \frac{(\delta\lambda^2\theta^2 + N^2\mu)^2}{(N\lambda\theta + N\delta\lambda\theta\mu)^2}} > 0, \tag{4.4}$$

$$\arccos\left(-\frac{\delta\lambda^2\theta^2+N^2\mu}{N\lambda\theta(1+\delta\mu)}\right)=\Delta_{\mathrm{cr}}\cdot\sqrt{\frac{\lambda^2\theta^2-N^2\mu^2}{N^2-\delta^2\lambda^2\theta^2}}>0. \tag{4.5}$$

□

Proposition 4.1 allows us to show that there exists a specific size of the weight δ that makes the queueing system optimally stable. We call this size of the weight $\delta_{\max}$, and it is such that $\delta=\delta_{\max}$ maximizes the threshold Δ_{cr}. In Proposition 4.2, we give an equation that determines $\delta_{\max}$ and provide closed-form expressions for an upper and a lower bound of $\delta_{\max}$.

Proposition 4.2 *Suppose $\lambda\theta>N\mu$. There exists a unique $\delta_{\max}\geq 0$ that maximizes a given root Δ_{cr} for fixed parameters λ,μ,N,θ. It is given by the solution of*

$$\frac{\sqrt{N^2-\delta_{\max}^2\lambda^2\theta^2}}{1+\delta_{\max}\mu}=\frac{\delta_{\max}\lambda^2\theta^2}{\sqrt{\lambda^2\theta^2-N^2\mu^2}}\cdot\arccos\left(-\frac{\delta_{\max}\lambda^2\theta^2+N^2\mu}{N\lambda\theta(1+\delta_{\max}\mu)}\right). \tag{4.6}$$

Furthermore, $\delta_{\max}$ is bounded by $\delta_1<\delta_{\max}<\delta_2$, where

$$\delta_1=\frac{-\left(\Delta_0+\frac{N}{\lambda\theta}\right)\lambda\theta+\sqrt{\lambda^2\theta^2\left(\Delta_0+\frac{N}{\lambda\theta}\right)^2+4N^2\left(\Delta_0+\frac{N}{\lambda\theta}\right)\mu+4N^2}}{2\lambda\theta\left(1+\left(\Delta_0+\frac{N}{\lambda\theta}\right)\mu\right)}, \tag{4.7}$$

$$\delta_2=\frac{-\Delta_0\lambda\theta+\sqrt{\lambda^2\theta^2\Delta_0^2+4N^2\Delta_0\mu+4N^2}}{2\lambda\theta(1+\Delta_0\mu)}, \tag{4.8}$$

$$\Delta_0=\arccos\left(-\frac{N\mu}{\lambda\theta}\right)\cdot\sqrt{\frac{N^2}{\lambda^2\theta^2-N^2\mu^2}}. \tag{4.9}$$

Proof We can treat Δ_{cr} as a function of δ. Implicit differentiation of (3.48) gives the rate with which Δ_{cr} changes:

$$\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta)=\frac{N^2-\delta\lambda^2\theta^2\left(\delta+\Delta_{\mathrm{cr}}(\delta)+\delta\mu\Delta_{\mathrm{cr}}(\delta)\right)}{(N^2-\delta^2\lambda^2\theta^2)(1+\delta\mu)}=\frac{1}{1+\delta\mu}-\frac{\delta\lambda^2\theta^2\Delta_{\mathrm{cr}}(\delta)}{N^2-\delta^2\lambda^2\theta^2} \tag{4.10}$$

$$=\frac{1}{1+\delta\mu}-\frac{\delta\lambda^2\theta^2}{N^2-\delta^2\lambda^2\theta^2}\cdot\arccos\left(-\frac{\delta\lambda^2\theta^2+N^2\mu}{N\lambda\theta(1+\delta\mu)}\right)\cdot\sqrt{\frac{N^2-\delta^2\lambda^2\theta^2}{\lambda^2\theta^2-N^2\mu^2}}. \tag{4.11}$$

By Proposition 4.1, $\Delta_{\mathrm{cr}}(\delta)$ is concave on the interval $[0,\frac{N}{\lambda\theta})$. Further, it can be shown that $\frac{d}{d\delta}\Delta_{\mathrm{cr}}(0)=1>0$ and $\lim_{\delta\to\frac{N}{\lambda\theta}}\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta)=-\infty<0$, so there is a point $\delta_{\max}$ where $\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta_{\max})=0$. Therefore $\Delta_{\mathrm{cr}}(\delta)$ reaches its absolute maximum at $\delta_{\max}\in(0,\frac{N}{\lambda\theta})$. For intuition, we plot $\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta)$ in Fig. 13.

The value $\delta_{\max}$ can be found numerically by solving $\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta_{\max})=0$ from Eq. (4.11), alternatively written as (4.6). It remains to find closed-form expressions

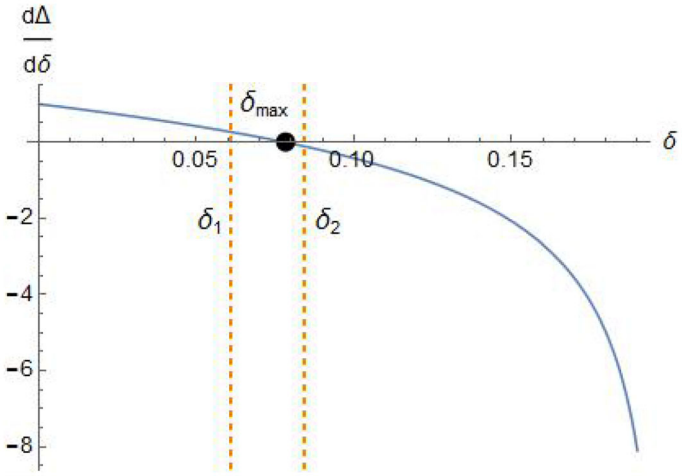

Fig. 13 $\delta_{\max}$ and its bounds $\delta_1 < \delta_{\max} < \delta_2$

for the bounds on $\delta_{\max}$. By Eq. (4.10), we can express $\delta_{\max}$ as

$$\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta_{\max}) = \frac{1}{1+\delta_{\max}\mu} - \frac{\delta_{\max}\lambda^2\theta^2\Delta_{\mathrm{cr}}(\delta_{\max})}{N^2-\delta_{\max}^2\lambda^2\theta^2} = 0, \tag{4.12}$$

$$\frac{1}{1+\delta_{\max}\mu} - \frac{\delta_{\max}\lambda^2\theta^2\Delta_0}{N^2-\delta_{\max}^2\lambda^2\theta^2} > 0, \tag{4.13}$$

where $\Delta_0 = \Delta_{\mathrm{cr}}(0) < \Delta_{\mathrm{cr}}(\delta_{\max})$. When solved for $\delta_{\max}$, the inequality (4.13) produces an upper bound condition $\delta_{\max} < \delta_2$ given by Eq. (4.8).

To find the lower bound, we note that $\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta)$ is monotonically decreasing. Thus, $\frac{d}{d\delta}\Delta_{\mathrm{cr}}(\delta) < \frac{d}{d\delta}\Delta_{\mathrm{cr}}(0) = 1$ for all $\delta \in (0, \frac{N}{\lambda\theta})$, and $\Delta_{\mathrm{cr}}(\delta) \le \delta + \Delta_{\mathrm{cr}}(0) < \frac{N}{\lambda\theta} + \Delta_0$. Therefore, by Eq. (4.12), we get

$$\frac{1}{1+\delta_{\max}\mu} - \frac{\delta_{\max}\lambda^2\theta^2\left(\Delta_0+\frac{N}{\lambda\theta}\right)}{N^2-\delta_{\max}^2\lambda^2\theta^2} < 0, \tag{4.14}$$

which produces the bound $\delta_{\max} > \delta_1$ from Eq. (4.7) when solved for $\delta_{\max}$. □

Figures 14 and 15 show Δ_{cr} as a function of λ and δ. For each arrival rate λ, the maximum Δ_{cr} is attained for some δ between the two curves δ_1 and δ_2. Similarly, Figs. 16 and 17 show Δ_{cr} as a function of μ and δ with the two curves δ_1 and δ_2. As seen in Figs. 14, 15, 16 and 17, the bounds on $\delta_{\max}$ are tight.

Besides knowing at which value δ the maximal bifurcation threshold Δ_{cr} may occur, it is also important to know how large that threshold actually is. In the next result, we develop bounds for the maximum Δ_{cr} that can be attained for fixed parameters λ, N, μ, and θ.

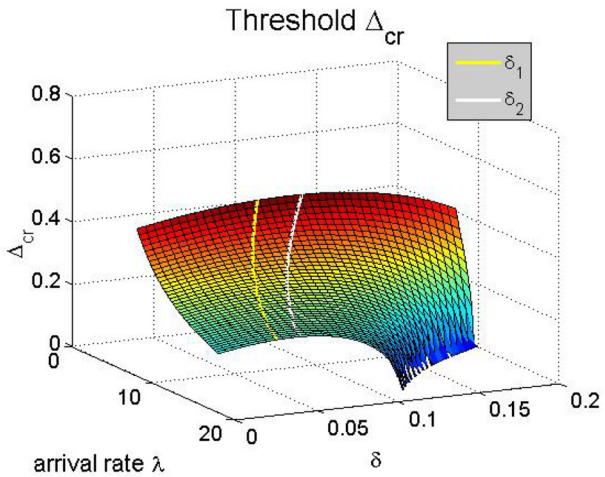

Fig. 14 For each λ, the maximum Δ_{cr} is achieved when $\delta \in (\delta_1, \delta_2)$; $\mu = 1$, $\theta = 1$

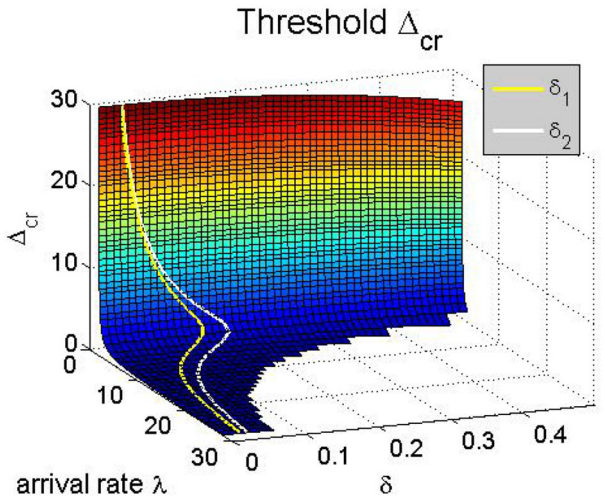

Fig. 15 For each λ, the maximum Δ_{cr} is achieved when $\delta \in (\delta_1, \delta_2)$; $\mu = 1$, $\theta = 1$

Proposition 4.3 *The maximum value of a root* Δ_{cr} *for fixed parameters* λ, μ, *and* N *is attained at* $\delta_{\max}$ *and is bounded by* $\Delta_1 < \Delta_{\text{cr}}(\delta_{\max}) < \Delta_2$, *where*

$$\Delta_1 = \max[\Delta_{\text{cr}}(\delta_1), \Delta_{\text{cr}}(\delta_2)], \quad \Delta_2 = \min[\Delta_{2a}, \Delta_{2b}], \tag{4.15}$$

$$\Delta_{2a} = \Delta_{\text{cr}}(\delta_1) + (\delta_2 - \delta_1) \cdot \frac{d}{d\delta}\Delta_{\text{cr}}(\delta_1), \quad \Delta_{2b} = \Delta_{\text{cr}}(\delta_2) - (\delta_2 - \delta_1) \cdot \frac{d}{d\delta}\Delta_{\text{cr}}(\delta_2). \tag{4.16}$$

Proof By Proposition 4.2, $\Delta_{\text{cr}}(\delta)$ attains its maximum at $\delta = \delta_{\max}$. Hence the lower bound $\Delta_1 < \Delta_{\text{cr}}(\delta_{\max})$ trivially follows, since $\delta_{\max} \neq \delta_1, \delta_2$. To find an upper bound,

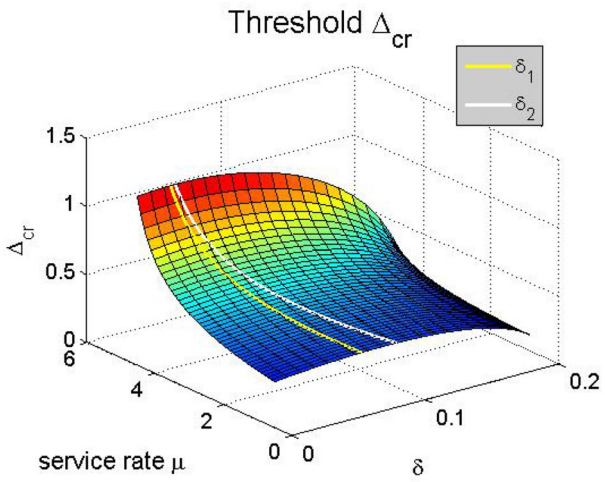

Fig. 16 For each μ, the maximum $\Delta_{\rm cr}$ is achieved when $\delta \in (\delta_1, \delta_2)$; $\lambda = 10$, $\theta = 1$

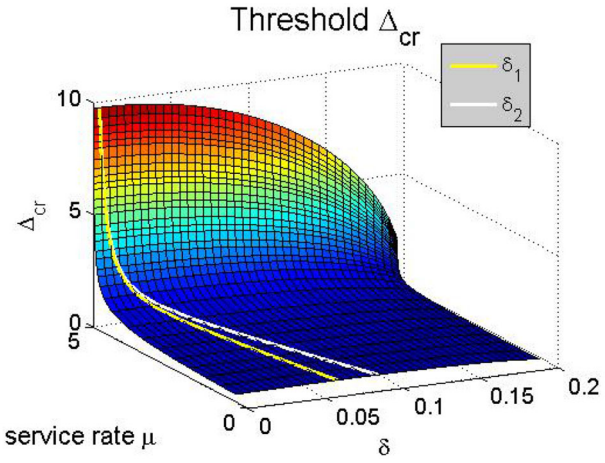

Fig. 17 For each μ, the maximum $\Delta_{\rm cr}$ is achieved when $\delta \in (\delta_1, \delta_2)$; $\lambda = 10$, $\theta = 1$

note that $\frac{d}{d\delta}\Delta_{\rm cr}(\delta)$ is a monotonically decreasing function, so $\frac{d}{d\delta}\Delta_{\rm cr}(\delta_1) > \frac{d}{d\delta}\Delta_{\rm cr}(\delta)$ for all $\delta > \delta_1$, and also that $\frac{d}{d\delta}\Delta_{\rm cr}(\delta_1) > 0$ since $\Delta_{\rm cr}(\delta)$ increases while $\delta < \delta_{\max}$. Hence

$$\Delta_{\rm cr}(\delta_{\max}) = \Delta_{\rm cr}(\delta_1) + \int_{\delta_1}^{\delta_{\max}} \frac{d}{d\delta}\Delta_{\rm cr}(\delta)d\delta < \Delta_{\rm cr}(\delta_1) + \int_{\delta_1}^{\delta_{\max}} \frac{d}{d\delta}\Delta_{\rm cr}(\delta_1)d\delta \tag{4.17}$$

$$= \Delta_{\rm cr}(\delta_1) + (\delta_{\max} - \delta_1)\frac{d}{d\delta}\Delta_{\rm cr}(\delta_1) < \Delta_{\rm cr}(\delta_1) + (\delta_2 - \delta_1)\frac{d}{d\delta}\Delta_{\rm cr}(\delta_1) = \Delta_{2a}. \tag{4.18}$$

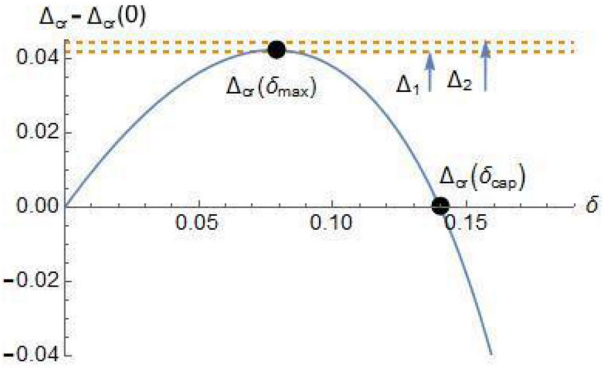

Fig. 18 $\delta_{\max}$ maximizes $\Delta_{\rm cr}$

In addition, it is known that $\frac{d}{d\delta}\Delta_{\rm cr}(\delta) < 0$ when $\delta > \delta_{\max}$, so

$$\Delta_{\rm cr}(\delta_{\max}) = \Delta_{\rm cr}(\delta_2) - \int_{\delta_{\max}}^{\delta_2} \frac{d}{d\delta}\Delta_{\rm cr}(\delta)d\delta < \Delta_{\rm cr}(\delta_2) - \int_{\delta_{\max}}^{\delta_2} \frac{d}{d\delta}\Delta_{\rm cr}(\delta_2)d\delta \tag{4.19}$$

$$= \Delta_{\rm cr}(\delta_2) - (\delta_2 - \delta_{\max})\frac{d}{d\delta}\Delta_{\rm cr}(\delta_2) < \Delta_{\rm cr}(\delta_2) - (\delta_2 - \delta_1)\frac{d}{d\delta}\Delta_{\rm cr}(\delta_2) = \Delta_{2b}. \tag{4.20}$$

Therefore $\Delta_{\rm cr}(\delta_{\max}) < \min[\Delta_{2a}, \Delta_{2b}] = \Delta_2$, as desired. □

Figure 18 illustrates $\Delta_{\rm cr}(\delta) - \Delta_{\rm cr}(0)$ as a function of δ, with the maximum attained at $\delta_{\max}$ and the bounds on the maximum given by Δ_1 and Δ_2. Further, it is evident from Fig. 18 that there is a threshold value, which we call $\delta_{\rm cap}$, that places a cap on the potential utility of the velocity information. When δ is less than $\delta_{\rm cap}$, the queueing system becomes more stable from the velocity information because $\Delta_{\rm cr}(\delta) > \Delta_{\rm cr}(0)$. However, when δ exceeds $\delta_{\rm cap}$, the queues become more unstable in the sense that $\Delta_{\rm cr}(\delta) < \Delta_{\rm cr}(0)$. The result below provides an equation for $\delta_{\rm cap}$.

Proposition 4.4 *Suppose $\lambda\theta > N\mu$. There exists a unique $\delta_{\rm cap} > 0$ such that $\Delta_{\rm cr}(\delta) > \Delta_{\rm cr}(0)$ for all $\delta < \delta_{\rm cap}$, and $\Delta_{\rm cr}(\delta) < \Delta_{\rm cr}(0)$ for all $\delta > \delta_{\rm cap}$. It is given by the solution to*

$$\arccos\left(-\frac{N\mu}{\lambda\theta}\right)\sqrt{\frac{N^2}{\lambda^2\theta^2 - N^2\mu^2}} = \arccos\left(-\frac{\delta_{\rm cap}\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta_{\rm cap}\mu)}\right)\sqrt{\frac{N^2 - \delta_{\rm cap}^2\lambda^2\theta^2}{\lambda^2\theta^2 - N^2\mu^2}}. \tag{4.21}$$

Proof As previously shown, $\Delta_{\rm cr}(\delta)$ is monotonically increasing on $\delta \in [0, \delta_{\max})$ and monotonically decreasing on $\delta \in (\delta_{\max}, \frac{N}{\lambda\theta})$. Further, $\lim_{\delta \to \frac{N}{\lambda\theta}} \Delta_{\rm cr}(\delta) = 0 < \Delta_{\rm cr}(0)$ since $\Delta_{\rm cr}(0) > 0$ by assumption, so there exists exactly one point $\delta_{\rm cap}$ on the interval $(\delta_{\max}, \frac{N}{\lambda\theta})$ where $\Delta_{\rm cr}(\delta_{\rm cap}) = \Delta_{\rm cr}(0)$, and it also follows that $\Delta_{\rm cr}(\delta_{\rm cap}) > \Delta_{\rm cr}(0)$ for all $\delta < \delta_{\rm cap}$ and $\Delta_{\rm cr}(\delta_{\rm cap}) < \Delta_{\rm cr}(0)$ for all $\delta > \delta_{\rm cap}$. By substituting the expression for $\Delta_{\rm cr}$ from (3.52) into $\Delta_{\rm cr}(0) - \Delta_{\rm cr}(\delta_{\rm cap}) = 0$ we get Eq. (4.21). □

To summarize, when $\lambda\theta > N\mu$, the queues are stable when the delay is less than $\Delta_{\rm cr}$. We can therefore provide the most stability for the queues by choosing δ that maximizes $\Delta_{\rm cr}$, i.e., $\delta_{\max}$. Proposition 4.2 proves the existence of $\delta_{\max}$, gives an equation describing $\delta_{\max}$ and provides closed-form expressions for bounds δ_1 and δ_2 such that $\delta_1 < \delta_{\max} < \delta_2$. Proposition 4.3 also provides bounds Δ_1 and Δ_2 for the maximum value that $\Delta_{\rm cr}$ can take as a function of δ, so $\Delta_1 < \Delta_{\rm cr}(\delta_{\max}) < \Delta_2$. Lastly, we show that even if $\delta \neq \delta_{\max}$, it is still beneficial to include the velocity information as long as $\delta < \delta_{\rm cap}$. When δ exceeds $\delta_{\rm cap}$, however, $\Delta_{\rm cr}(\delta)$ becomes less than $\Delta_{\rm cr}(0)$, so the queues are less likely to be stable than if the velocity information was omitted altogether. Proposition 4.4 proves the existence of $\delta_{\rm cap}$ and provides an equation for it.

5 Impact of velocity information on the amplitude

Now that we have a good understanding of how the velocity information impacts the critical delay, we address a more practical question: What is the impact of the velocity on the amplitude of the oscillations? This question is important because it reveals how much the queues will oscillate when they are not in equilibrium. Moreover, it can provide an estimate of how much throughput is lost because of the oscillations (amusement park capacity) or even provide valuable estimates of how much fuel or energy is lost in transportation settings.

Although our previous analysis holds for an arbitrary number of queues, in the sequel we will demonstrate how δ affects the amplitude dynamics of the queues in the case of a two-queue network. One reason for this restriction is that we must move beyond linearization techniques. In fact, we must use third-order Taylor expansions to obtain information about the amplitude; see, for example, [29]. Thus, many of the matrix techniques we exploited for linearizing the NDDE in Section 2 cannot be used in the context of tensors for the third-order Taylor expansion. Thus, for the case of two dimensions, we have the following system of equations:

$$\dot{q}_1(t) = \lambda \cdot \frac{\exp\Big(-\theta\big(q_1(t-\Delta) + \delta\dot{q}_1(t-\Delta)\big)\Big)}{\sum_{j=1}^{2}\exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_j(t-\Delta)\big)\Big)} - \mu q_1(t), \tag{5.1}$$

$$\dot{q}_2(t) = \lambda \cdot \frac{\exp\Big(-\theta\big(q_2(t-\Delta) + \delta\dot{q}_2(t-\Delta)\big)\Big)}{\sum_{j=1}^{2}\exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_j(t-\Delta)\big)\Big)} - \mu q_2(t), \tag{5.2}$$

where as usual $\Delta, \lambda, \mu, \theta > 0$ and $\delta \geq 0$. Similarly to Sect. 4, we will consider the scenario with $\lambda\theta > 2\mu$, where even for a small δ the equilibrium of queues loses stability for sufficiently large delay. Our first result shows that the Hopf bifurcations that occur at each root $\Delta_{\rm cr}$ are supercritical.

Theorem 5.1 *Suppose $\omega_{\rm cr} \in \mathbb{R}$ and $\omega_{\rm cr} \neq 0$. The NDDE system (5.1)–(5.2) undergoes a supercritical Hopf bifurcation at each root $\Delta_{\rm cr}$. If $\delta\mu < 1$ then the limit cycle is born when $\Delta \leq \Delta_{\rm cr}$. If $\delta\mu > 1$ then the limit cycle is born when $\Delta \geq \Delta_{\rm cr}$.*

Proof We will use the method of *slow flow*, or the Method of Multiple Scales, to determine the stability of the Hopf bifurcations given by Theorem 3.8. This method is often applied to systems of delay differential equations (DDEs) [5,7,25]. We note, however, that the stability of the limit cycles can also be determined by showing that the floquet exponent has negative real part, as outlined in Hassard et al. [18].

The first step in the method of slow flow is to consider the perturbation of q_1 and q_2 from the equilibrium point $q_1^* = q_2^* = \frac{\lambda}{2\mu}$, and to approximate the resulting derivatives by third-order Taylor expansion. The two resulting DDEs can be uncoupled when their sum and their difference are taken:

$$w_1(t) = q_1(t) + q_2(t), \quad w_2(t) = q_1(t) - q_2(t), \tag{5.3}$$

$$\dot{w}_1(t) = -\mu w_1(t), \tag{5.4}$$

$$\dot{w}_2(t) = -\mu w_2(t) - \frac{\lambda\theta}{2}(w_2(t-\Delta) + \delta\dot{w}_2(t-\Delta)), \tag{5.5}$$

$$+ \frac{\lambda\theta^3}{24}(w_2(t-\Delta) + \delta\dot{w}_2(t-\Delta))^3 + O(w_2^4). \tag{5.6}$$

The function $w_1(t) = Ce^{-\mu t}$ decays to 0, while the function $w_2(t)$ has a Hopf bifurcation at Δ_{cr} where the periodic solutions are born.

We set $w_2(t) = \sqrt{\epsilon}x(t)$ in order to prepare the NDDE for perturbation treatment:

$$\dot{x}(t) = -\mu x(t) - \frac{\lambda\theta}{2}(x(t-\Delta) + \delta\dot{x}(t-\Delta)) + \frac{\sqrt{\epsilon}\lambda\theta^3}{24}(x(t-\Delta) + \delta\dot{x}(t-\Delta))^3. \tag{5.7}$$

We replace the independent variable t by two new time variables $\xi = \omega t$ (stretched time) and $\eta = \epsilon t$ (slow time). Then we expand Δ and ω about the critical Hopf values:

$$\Delta = \Delta_{\text{cr}} + \epsilon\alpha, \quad \omega = \omega_{\text{cr}} + \epsilon\beta. \tag{5.8}$$

The time derivative $\dot{x}$ becomes

$$\dot{x} = \frac{dx}{dt} = \frac{\partial x}{\partial \xi}\frac{d\xi}{dt} + \frac{\partial x}{\partial \eta}\frac{d\eta}{dt} = \frac{\partial x}{\partial \xi}\cdot(\omega_{\text{cr}} + \epsilon\beta) + \frac{\partial x}{\partial \eta}\cdot\epsilon. \tag{5.9}$$

The expression for $x(t-\Delta)$ may be simplified by Taylor expansion for small ϵ:

$$x(t-\Delta) = x(\xi - \omega\Delta, \eta - \epsilon\Delta) \tag{5.10}$$

$$= x(\xi - (\omega_{\text{cr}} + \epsilon\beta)(\Delta_{\text{cr}} + \epsilon\alpha), \eta - \epsilon(\Delta_{\text{cr}} + \epsilon\alpha)) + O(\epsilon^2) \tag{5.11}$$

$$= \tilde{x} - \epsilon(\omega_{\text{cr}}\alpha + \Delta_{\text{cr}}\beta)\cdot\frac{\partial\tilde{x}}{\partial\xi} - \epsilon\Delta_{\text{cr}}\frac{\partial\tilde{x}}{\partial\eta} + O(\epsilon^2), \tag{5.12}$$

where $x(\xi - \omega_{cr}\Delta_{cr}, \eta) = \tilde{x}$. The function x is represented as $x = x_0 + \epsilon x_1 + \ldots$, and we get

$$\frac{dx}{dt} = \omega_{cr}\frac{\partial x_0}{\partial \xi} + \epsilon\beta\frac{\partial x_0}{\partial \xi} + \epsilon\frac{\partial x_0}{\partial \eta} + \epsilon\omega_{cr}\frac{\partial x_1}{\partial \xi}. \tag{5.13}$$

After these substitutions are made into (5.7), the resulting equation can be separated by the powers of ϵ into two equations. For the ϵ^0 terms, we get an equation for x_0 without any terms involving x_1, namely $L(x_0) = 0$, where

$$L(x_0) = \mu x_0 + \frac{\lambda}{2}\tilde{x}_0 + \omega_{cr}\frac{\partial x_0}{\partial \xi} + \frac{\delta\lambda\omega_{cr}}{2}\frac{\partial \tilde{x}_0}{\partial \xi} = 0, \tag{5.14}$$

which is satisfied with a solution of the form

$$x_0(t) = A(\eta)\cos(\xi) + B(\eta)\sin(\xi). \tag{5.15}$$

The equation resulting from ϵ^1 terms is $L(x_1) + M(x_0) = 0$. Since $L(x_1) = 0$ is satisfied by a solution of the form (5.15), then the terms from $M(x_0)$ involving $\cos(\xi)$ and $\sin(\xi)$ are resonant. To eliminate the resonant terms, their coefficients must be 0, which gives two equations for $A(\eta)$ and $B(\eta)$. Switching into polar coordinates, we define $R = \sqrt{A^2 + B^2}$, and find

$$\frac{dR}{d\eta} = -\frac{R\left(c_1 R^2 - c_2\right)}{c_3}, \quad \text{where} \tag{5.16}$$

$$\begin{aligned} c_1 &= (\mu^2 + \omega_{cr}^2)(\mu^2 + \omega_{cr}^2 + \delta^2\omega_{cr}^2\mu^2 + \delta^2\omega_{cr}^4)\Delta_{cr} \\ &\quad + (\mu^2 + \omega_{cr}^2)(\mu - \delta\mu^2 + \delta^2\omega_{cr}^2\mu - \delta\omega_{cr}^2), \end{aligned} \tag{5.17}$$

$$c_2 = 4\alpha\lambda^2\omega_{cr}^2(1 - \delta^2\mu^2), \tag{5.18}$$

$$\begin{aligned} c_3 &= \Delta_{cr}^2 \cdot 4\lambda^2(\mu^2 + \omega_{cr}^2 + \delta^2\mu^2\omega^2 + \delta^2\omega_{cr}^2) \\ &\quad + \Delta_{cr} \cdot 8\lambda^2(\mu - \delta\mu^2 - \delta\omega_{cr}^2 + \delta^2\mu\omega_{cr}^2) + 4\lambda^2(1 - \delta\mu)^2. \end{aligned} \tag{5.19}$$

In order to find the equilibrium points of R and to discuss their stability, we need to show that the coefficients c_1, c_2, and c_3 are positive. Notice that c_3 is a quadratic function of Δ_{cr} with the minimum located at Δ_{cr}^* such that $\frac{d}{d\Delta_{cr}}c_3(\Delta_{cr}^*) = 0$, hence

$$\Delta_{cr}^* = \frac{\delta}{1 + \delta^2\omega_{cr}^2} - \frac{\mu}{\mu^2 + \omega_{cr}^2}, \tag{5.20}$$

$$c_3 = c_3(\Delta_{cr}) \geq c_3(\Delta_{cr}^*) = \frac{4\lambda^2\omega_{cr}^2(1 - \delta^2\mu^2)^2}{(\mu^2 + \omega_{cr}^2)(1 + \delta^2\omega_{cr}^2)} > 0, \tag{5.21}$$

therefore the denominator of $\frac{dR}{d\eta}$, c_3, is always positive. Also, we can show c_1 to be positive. We first note that at the Hopf, Eq. (3.51) must be satisfied so $\cos(\omega_{cr}\Delta_{cr}) < 0$, which implies that $\omega_{cr}\Delta_{cr} > \frac{\pi}{2}$ and

$$\Delta_{\text{cr}} > \frac{\pi}{2\omega_{\text{cr}}}. \tag{5.22}$$

Next, we note that c_1 is an increasing linear function of Δ_{cr}, so c_1 must be positive for any $\Delta_{\text{cr}} > \Delta^*_{\text{cr}}$, where $c_1(\Delta^*_{\text{cr}}) = 0$. This Δ^*_{cr} is found to be

$$\Delta^*_{\text{cr}} = \frac{\delta}{1+\delta^2\omega_{\text{cr}}^2} - \frac{\mu}{\mu^2+\omega_{\text{cr}}^2}. \tag{5.23}$$

Using the inequality in (5.22), we can show by contradiction that Δ_{cr} is always greater than Δ^*_{cr}. Suppose that, for some parameters, we have $\Delta^*_{\text{cr}} > \frac{\pi}{2\omega_{\text{cr}}}$. From the equation (3.47), this implies that

$$\frac{\pi}{2\omega_{\text{cr}}} < \Delta^*_{\text{cr}} = \frac{\delta}{1+\delta^2\omega_{\text{cr}}^2} - \frac{\mu}{\mu^2+\omega_{\text{cr}}^2} < \frac{\delta}{1+\delta^2\omega_{\text{cr}}^2}, \tag{5.24}$$

$$\frac{\pi}{2}(1+\delta^2\omega_{\text{cr}}^2) < \delta\omega_{\text{cr}}, \tag{5.25}$$

$$\frac{2\pi(1-\delta^2\mu^2)}{4-\delta^2\lambda^2\theta^2} < \delta\sqrt{\frac{\lambda^2\theta^2-4\mu^2}{4-\delta^2\lambda^2\theta^2}}, \tag{5.26}$$

$$4\pi^2\cdot\frac{(1-\delta^2\mu^2)^2}{(4-\delta^2\lambda^2\theta^2)^2} < \delta^2\cdot\frac{\lambda^2\theta^2-4\mu^2}{4-\delta^2\lambda^2\theta^2}, \tag{5.27}$$

$$4\pi^2\left(1-\delta^2\mu^2\right)^2 < \delta^2(\lambda^2\theta^2-4\mu^2)(4-\delta^2\lambda^2\theta^2). \tag{5.28}$$

Set $\bar{\delta} = \delta^2$. The inequality can be written as

$$f(\bar{\delta}) = \left(\lambda^4\theta^4 - 4\lambda^2\theta^2\mu^2 + 4\pi^2\mu^4\right)\bar{\delta}^2 - 4\left(\lambda^2\theta^2 + 2\pi^2\mu^2 - 4\mu^2\right)\bar{\delta} + 4\pi^2 < 0. \tag{5.29}$$

Notice that the coefficient of $\bar{\delta}^2$ is always positive. It can be shown by finding μ^2 that minimizes the coefficient, $\mu^2 = \frac{\lambda^2\theta^2}{2\pi^2}$, and then finding the minimum value of that coefficient, which is $\lambda^4\theta^4\left(1-\frac{1}{\pi^2}\right)$ so it is clearly positive. This means that $f(\bar{\delta})$ is a convex function, with a minimum at $\bar{\delta}^*$:

$$\bar{\delta}^* = \frac{2(\lambda^2\theta^2+2\pi^2\mu^2-4\mu^2)}{\lambda^4\theta^4-4\lambda^2\theta^2\mu^2+4\pi^2\mu^4}, \tag{5.30}$$

$$f(\bar{\delta}) \geq f(\bar{\delta}^*) = \frac{4(\pi^2-1)(\lambda^2\theta^2-4\mu^2)^2}{\lambda^4\theta^4-4\lambda^2\theta^2\mu^2+4\pi^2\mu^4} > 0, \tag{5.31}$$

where the denominator is the same as the coefficient of $\bar{\delta}^2$ from Eq. (5.29), so it must be positive. The inequalities (5.29) and (5.31) contradict each other, and so $\Delta^*_{\text{cr}} \leq \frac{\pi}{2\omega_{\text{cr}}}$ for all parameters. Hence by Eq. (5.22), $\Delta_{\text{cr}} > \Delta^*_{\text{cr}}$, which implies that c_1 must be positive.

Since c_1 is positive, the only way for R from (5.16) to have a nonzero equilibrium point is for c_2 to be also positive. This produces the conditions on the direction of the Hopf

$$\delta\mu < 1 \implies \alpha > 0 \tag{5.32}$$
$$\delta\mu > 1 \implies \alpha < 0. \tag{5.33}$$

Recall that α represents the perturbation from Δ_{cr}. So when $\delta\mu < 1$, the limit cycle is born when Δ exceeds Δ_{cr}. If $\delta\mu > 1$, then the limit cycle is born when Δ becomes less than Δ_{cr}. In either case, the equilibrium points of $R(\eta)$ are given by

$$R_0 = 0, \quad R_1 = \sqrt{\frac{c_2}{c_1}} > 0. \tag{5.34}$$

Since $c_1, c_2, c_3 > 0$, R_0 is unstable and R_1 is stable. In its explicit form,

$$R_1 = \sqrt{\frac{4\alpha(\lambda^2\theta^2 - 4\mu^2)(4 - \delta^2\lambda^2\theta^2)^2}{\theta^2(1-\delta^2\mu^2)(16\mu + \lambda^2\theta^2(4\Delta_{\text{cr}} - 4\delta + \delta^3\lambda^2\theta^2 - 4\delta^2\mu - 4\delta^2\Delta_{\text{cr}}\mu^2))}}, \tag{5.35}$$

which represents the amplitude of the limit cycle near the Hopf. Since R_1 is stable, then the Hopf bifurcation is supercritical. □

Theorem 5.1 establishes that as Δ increases, the equilibrium becomes unstable and a stable limit cycle is born. We also observe from Eq. (5.35) that the amplitude of the oscillations depends heavily on the model parameters. From taking simple derivatives, one can observe that the amplitude is increasing as a function of λ and θ and is decreasing as a function δ, μ, and N.

5.1 Throughput lost by oscillations

Now that we have an expression for the amplitude of the oscillations, one can estimate how much throughput is lost in this context. There are two reasons why the throughput analysis is quite simple in this setting. First, the oscillations are given by sine functions, which have a lot of symmetry. Second, we only have our amplitude calculations for the two queue case, which has the property that if a customer doesn't go to the first queue, the customer must go to the second one. Thus, there is a nice symmetry for the two queue case around the equilibrium. We define lost throughput by capacity below the equilibrium that is not being used and the other queue is above the equilibrium. Thus, in the two queue case this is equal to

$$\text{Throughput Lost} = \frac{\frac{R_1}{2}\int_0^{\pi/\omega_{\text{cr}}} \sin(\omega_{\text{cr}}t)\mathrm{d}t}{\frac{\lambda}{N\mu}\cdot\frac{\pi}{\omega_{\text{cr}}}} \tag{5.36}$$

$$= \frac{\frac{R_1}{2} \cdot \frac{2}{\omega_{\text{cr}}}}{\frac{\lambda}{N\mu} \cdot \frac{\pi}{\omega_{\text{cr}}}} \tag{5.37}$$

$$= \frac{R_1}{\frac{\lambda}{N\mu} \cdot \pi} \tag{5.38}$$

$$= \underbrace{\left(\frac{2}{\pi}\right)}_{\text{sine to rectangle ratio}} \cdot \underbrace{\left(\frac{R_1/2}{\frac{\lambda}{N\mu}}\right)}_{\text{amplitude to equilibrium ratio}}. \tag{5.39}$$

This can be interpreted as the ratio of the sine function to a rectangle on one full period multiplied by the ratio of the amplitude of the oscillation divided by the equilibrium level. In an extreme case where the amplitude was the entire size of the equilibrium value, then clearly the loss is $2/\pi$.

5.2 First-order approximation of amplitude

We would like to choose the weight coefficient δ in a way that minimizes the amplitude of the oscillation in queues. To do this, we first need to know what the amplitude of the oscillations is as a function of the system parameters. In the following result, we use a perturbation method to approximate the amplitude of oscillations around the bifurcation point.

Proposition 5.2 *The amplitude of the oscillations of the queues near the first Hopf can be approximated by* $\frac{R_1}{2}$*, where* R_1 *is given by Eq.* (5.35)*.*

Proof The radius of the limit cycle from (5.35) approximates the amplitude of the oscillations of $w_2(t)$ from (5.5). By the change in variables given in Eq. (5.3), as $t \to \infty$, the behavior of the queues up to a phase shift is

$$q_1 = \frac{1}{2}(w_1 + w_2) \to \frac{1}{2} R_1 \sin(\omega \Delta t), \tag{5.40}$$

$$q_2 = \frac{1}{2}(w_1 - w_2) \to -\frac{1}{2} R_1 \sin(\omega \Delta t). \tag{5.41}$$

Thus, the amplitude of oscillations of queues is $\frac{R_1}{2}$. □

Therefore, when Δ exceeds Δ_{cr}, the amplitude of oscillations can be approximated to first order by

$$\text{amplitude} \approx \sqrt{\frac{(\Delta - \Delta_{\text{cr}})(\lambda^2\theta^2 - 4\mu^2)(4 - \delta^2\lambda^2\theta^2)^2}{\theta^2(1 - \delta^2\mu^2)(16\mu + \lambda^2\theta^2(4\Delta_{\text{cr}} - 4\delta + \delta^3\lambda^2\theta^2 - 4\delta^2\mu - 4\delta^2\Delta_{\text{cr}}\mu^2))}} \tag{5.42}$$

when $\Delta - \Delta_{\text{cr}}$ is small.

The approximation is accurate when δ is substantially smaller than the ratio $\frac{2}{\lambda\theta}$. For example, in Fig. 19 the queues oscillate throughout time, and the two horizontal lines

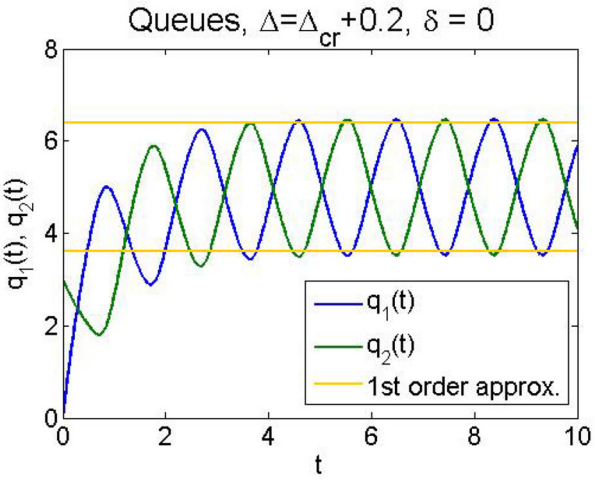

Fig. 19 Amplitude approximation; $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{\text{cr}} + 0.2$, $\delta = 0$

provide a good approximation of the amplitude of oscillations based on Eq. (5.42). However, the approximation becomes inaccurate when δ approaches $\frac{2}{\lambda\theta}$. As demonstrated in Fig. 20, when $\delta = 0.195$ and $\frac{2}{\lambda\theta} = 0.2$, the approximated amplitude is only about a half of what the actual amplitude is. The discrepancy is observed in Figs. 21, 22, 23 and 24 as well. The surface plot in Fig. 21 shows the true amplitude based on numerical integration as a function of the delay Δ and the coefficient δ, while the surface plot in Fig. 22 shows the amplitude's first-order approximation. Furthermore, the surface plot in Fig. 23 shows the error of first-order approximation, where the error increases with δ. Finally, Fig. 24 provides intuition for why the approximation fails as δ approaches $\frac{2}{\lambda\theta}$. Figure 24 presents a plot comparing the amplitude and its approximation as functions of delay while $\delta = 0.19$ is close to the threshold $\frac{2}{\lambda\theta} = 0.2$. The approximation is proportional to $\sqrt{\Delta - \Delta_{\text{cr}}}$, while the true amplitude appears to be a linear function of $(\Delta - \Delta_{\text{cr}})$ (even though it is not exactly linear).

Since we are interested in using the analytical expression of the amplitude approximation to determine the coefficient δ that minimizes the amplitude for a given delay, it is important for the approximation to be accurate. As seen from Fig. 22, for a fixed delay, say $\Delta = 0.5$, the point of the approximated minimum amplitude (at $\delta \approx 0.2$) does not agree with the true minimum amplitude (at $\delta \approx 0.11$). Hence, the first-order approximation of amplitude is insufficient for our purposes, and we must derive the second order approximation.

5.3 Second-order approximation of amplitude

The first-order approximation, as seen from Eq. (5.42), is of the form

$$\text{Amplitude} \approx c_0(\Delta - \Delta_{\text{cr}})^{0.5}, \tag{5.43}$$

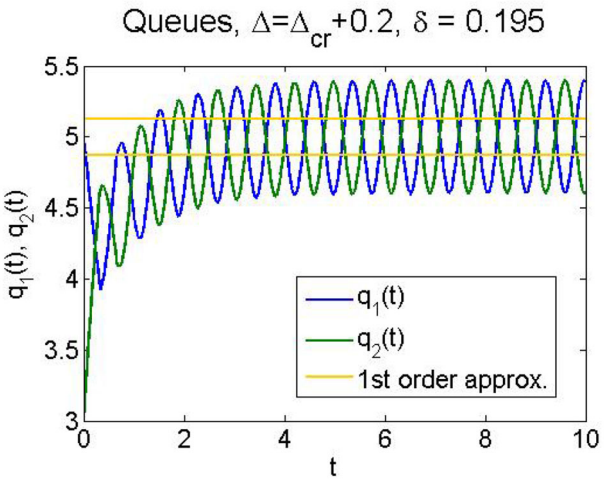

Fig. 20 Amplitude approximation, $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{\text{cr}} + 0.2$, $\delta = 0.195$

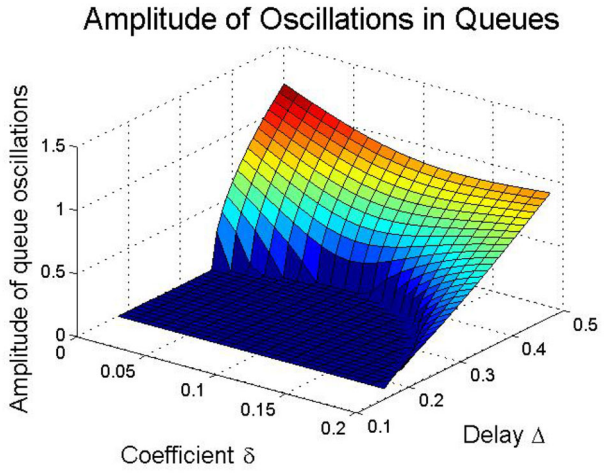

Fig. 21 Amplitude of oscillations; $\theta = 1$, $\lambda = 10$, $\mu = 1$

where c_0 is a factor determined by the system parameters and is independent of delay. The second-order approximation takes the form

$$\text{Amplitude} \approx c_0(\Delta - \Delta_{\text{cr}})^{0.5} + c_1(\Delta - \Delta_{\text{cr}})^{1.5}, \tag{5.44}$$

where c_1 is also independent of the delay. The full expression for c_1 is long and messy, so we omit it from this section. However, the reader can refer to Appendix 7.2 for the expressions as well as a discussion on how the second-order approximation is obtained. As shown in Figs. 25 and 26, the second-order approximation performs just as well as the first-order approximation when δ is significantly smaller than $\frac{2}{\lambda\theta}$, but

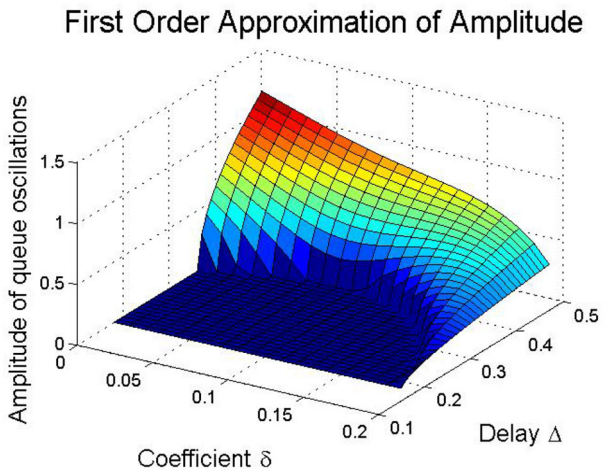

Fig. 22 First-order approximation; $\theta = 1, \lambda = 10, \mu = 1$

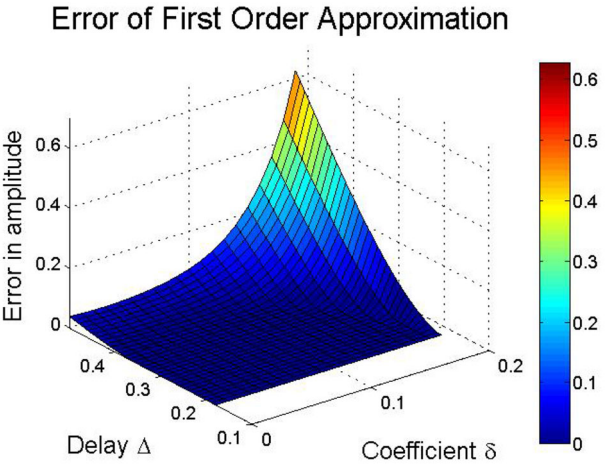

Fig. 23 Error of approximation; $\theta = 1, \lambda = 10, \mu = 1$

is much more accurate when δ approaches $\frac{2}{\lambda\theta}$. Figures 27 and 28 confirm that this trend holds throughout the parameter space in δ and delay Δ. Figure 29 compares the true amplitude with the two approximations when $\delta = 0.1$. The next plot in Fig. 30 draws the same comparison but when $\delta = 0.19$ is closer to its upper limit $\frac{2}{\lambda\theta} = 0.2$. It is evident from the two plots that the second-order approximation is significantly more accurate than the first-order approximation, especially as $\delta \to \frac{2}{\lambda\theta}$. Figures 31 and 32 illustrate the same point more systematically, by comparing the errors of first- and second-order approximations. These surface plots reveal that the higher-order approximation decreases the maximum error by a factor of 10.

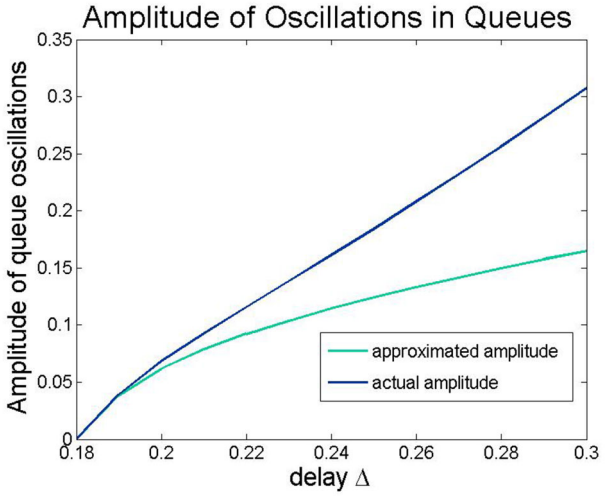

Fig. 24 First-order approximation when $\delta = 0.19$, $\theta = 1$, $\lambda = 10$, $\mu = 1$

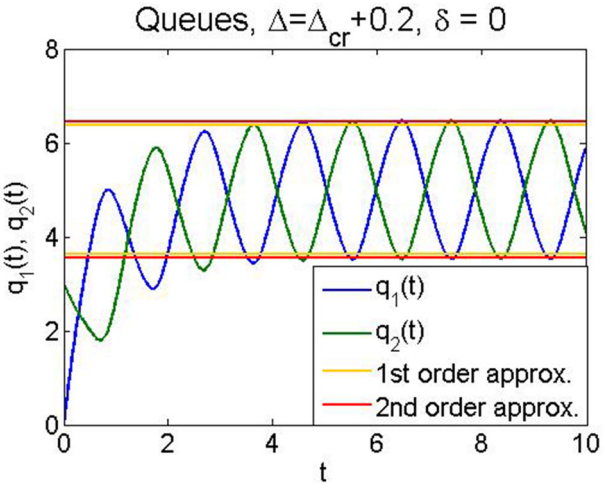

Fig. 25 Amplitude approximation; $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{\text{cr}} + 0.2$, $\delta = 0$

5.4 Minimizing the amplitude of oscillations

Since the second-order approximation is sufficiently accurate, we proceed by using the analytical formula of the second-order approximation to determine the coefficient δ that minimizes the amplitude of oscillations. Figure 33 shows the numerically computed amplitude, together with its minimum for each delay according to the second-order approximation. The minimum of the amplitude as a function of δ is found numerically in MATLAB. It is evident that the approximated minimum closely corresponds to where the true minimum is.

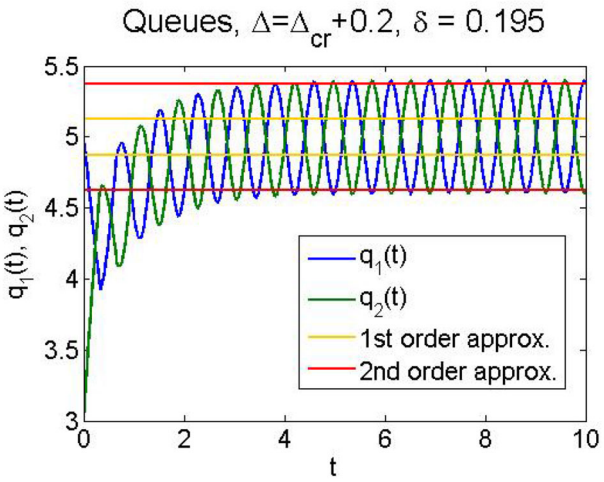

Fig. 26 Amplitude approximation; $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{\text{cr}} + 0.2$, $\delta = 0.195$

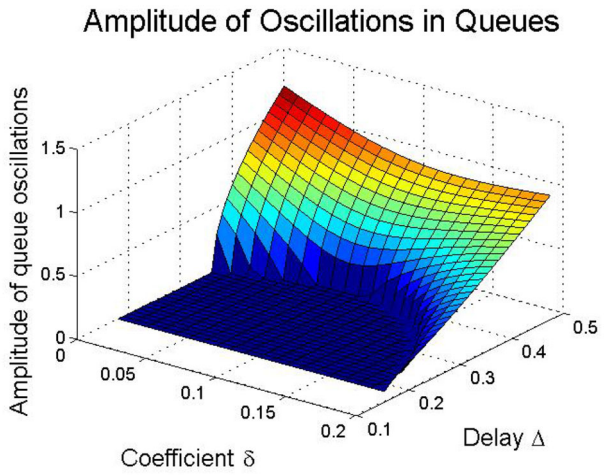

Fig. 27 Amplitude of oscillations; $\theta = 1$, $\lambda = 10$, $\mu = 1$

Figure 33 shows that the velocity information indeed affects the amplitude of oscillations, and the amplitude can be reduced with a proper choice of the coefficient δ. Figure 33 also reveals an important finding: The value $\delta_{\max}$ for the coefficient δ that maximizes Δ_{cr} is not the same as δ_{amp} that minimizes the amplitude of oscillations. Specifically, $\delta_{\max}$ is independent of the delay Δ, while δ_{amp} is a function of the delay. The one point where the two values are guaranteed to be equal each other, $\delta_{\max} = \delta_{\text{amp}}$, is when δ_{amp} is computed for the delay equal to the maximum possible Δ_{cr}, i.e., $\Delta = \Delta_{\text{cr}}(\delta_{\max})$. Therefore, one should use $\delta_{\max}$ as the weight coefficient as long as the delay is less than the bifurcation threshold Δ_{cr} evaluated at $\delta_{\max}$, but when the delay exceeds Δ_{cr} one should use δ_{amp} for the weight coefficient instead. Thus,

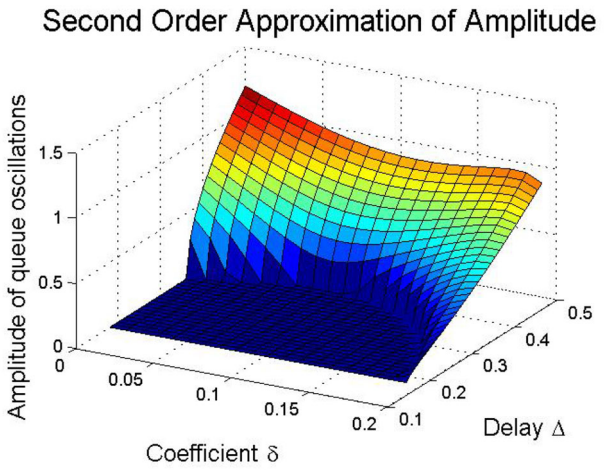

Fig. 28 First-order approximation; $\theta = 1, \lambda = 10, \mu = 1$

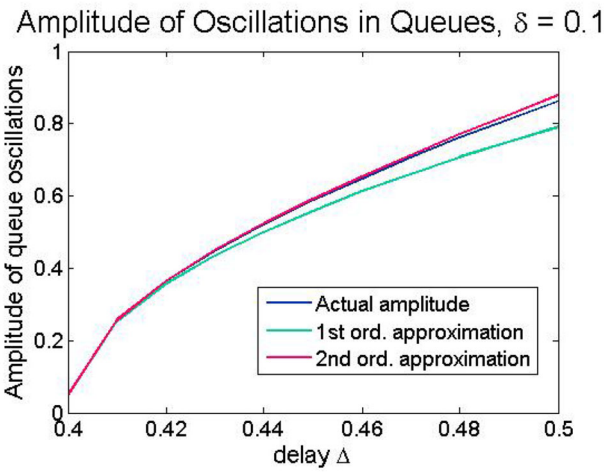

Fig. 29 Comparison when $\delta = 0.10$; $\theta = 1, \lambda = 10, \mu = 1$

a service manager can choose the value of δ either to increase the general stability of the whole system, $\delta = \delta_{\max}$, or they can choose to minimize the amplitude of the oscillations at their current level of service, $\delta = \delta_{\text{amp}}$.

6 Conclusion

This paper answers important questions with regards to businesses incorporating the queue length velocity into the information that is provided to customers via delay announcements. We consider the information passed to the customers about each of N queues to be a linear combination of the current queue length and the rate at which

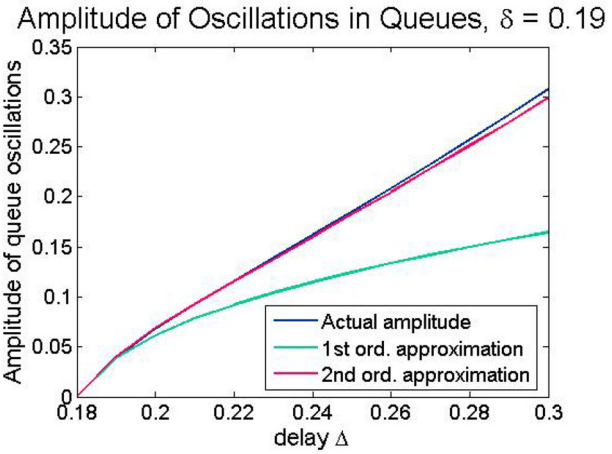

Fig. 30 Comparison when $\delta = 0.19$; $\theta = 1$, $\lambda = 10$, $\mu = 1$

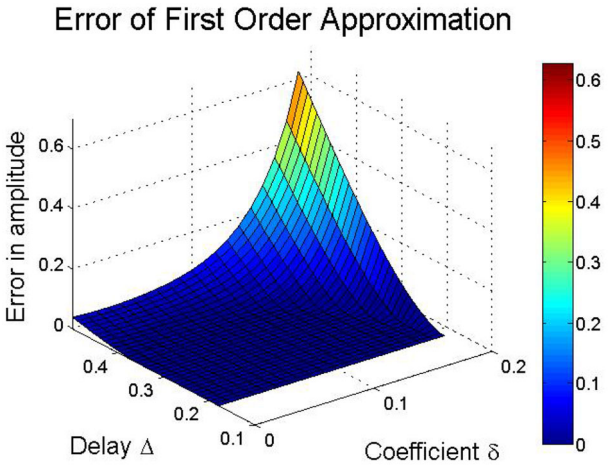

Fig. 31 First-order error; $\theta = 1$, $\lambda = 10$, $\mu = 1$

that queue is moving, or the queue velocity:

$$\text{Delay announcement about } i^{th} \text{ queue} = q_i(t-\Delta) + \delta \dot{q}_i(t-\Delta), \tag{6.1}$$

with the delay Δ being the time of customers traveling to the selected queue.

The most evident finding is that the coefficient δ that weighs the queue velocity information should always be less than the ratio $\frac{N}{\lambda\theta}$. Maintaining this limit guarantees that, at best, the queues will be locally stable for any delay in information. At worst, the queues will be stable when the delay Δ is sufficiently small, eventually undergoing a Hopf bifurcation at $\Delta = \Delta_{\text{cr}}$ and becoming unstable. Alternatively, if $\delta > \frac{N}{\lambda\theta}$, then

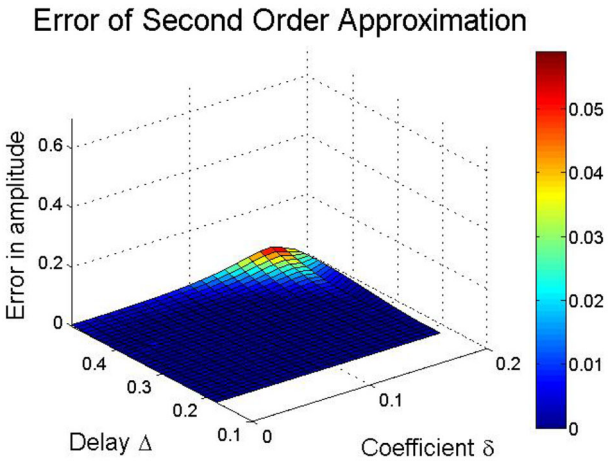

Fig. 32 Second-order error; $\theta = 1$, $\lambda = 10$, $\mu = 1$

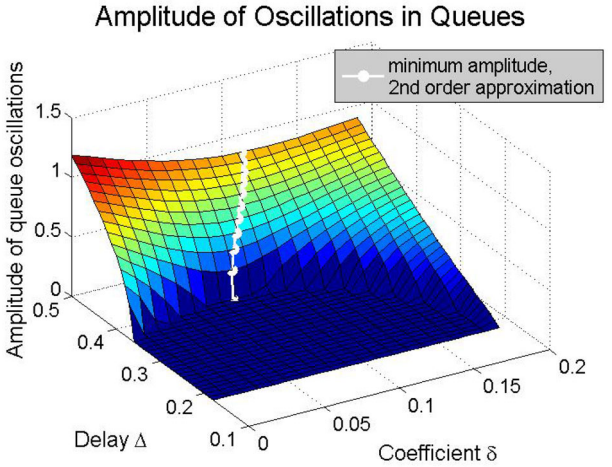

Fig. 33 For any delay, the amplitude can be minimized as a function of δ

the queues will never be stable, even when the delay in information is infinitesimally small. The reader can refer to Fig. 4 for more details.

Even when the condition $\delta < \frac{N}{\lambda\theta}$ is met, significant improvements can still be made by choosing δ optimally. In the case when queues become unstable as the delay in information increases (so when $\lambda\theta > N\mu$), the weight δ can shift the delay threshold Δ_{cr} at which the queues become unstable. In fact, there is exists a "cap" on the weight, δ_{cap}, such that it is safe and beneficial to include the queue velocity whenever $\delta \leq \delta_{\text{cap}}$, meaning that the queues will remain stable under greater delay than if the velocity information was omitted. Further, if the threshold δ_{cap} is exceeded, then queue velocity information will be harmful to the system. In this case, the queues will lose their stability for a smaller delay Δ than if the queue velocity was omitted altogether. An

edge case that exemplifies the usefulness of this discovery is as follows: If we take $\delta \to \frac{N}{\lambda\theta}$, at which point it is clear that $\delta > \delta_{\text{cap}}$, then the queues bifurcate almost immediately because $\Delta_{\text{cr}} \to 0$ even though the same queues would have remained stable under a much larger delay if δ was set to 0. Hence, it is important to keep δ smaller than δ_{cap}.

We also showed that there exists an optimal value for δ, called $\delta_{\max}$, that gives the most stability to the queues. For $\delta = \delta_{\max}$, queues will be stable for greater delay than is possible given any other choice of δ. We provide an equation from which $\delta_{\max}$ can be found numerically, as well as closed-form expressions for upper and lower bounds on $\delta_{\max}$. Choosing δ within those bounds is a safe choice for service managers.

This leads to a natural assessment of the limitations of providing the queue velocity information. The threshold Δ_{cr} where the queues lose stability can be arbitrarily close to 0 when δ is chosen poorly, but even the best choice of δ can only help so much. We provide a formula from which the maximum attainable Δ_{cr} can be computed. Further, we give expressions on the bounds for that optimal Δ_{cr} because they don't rely on $\delta_{\max}$ and hence they may be easier to evaluate. This means that while including δ can always improve the queue dynamics to some degree, there is a limit on how much impact δ may have.

The presence of the queue velocity information can also affect the amplitude with which the queues oscillate after losing stability. From numerical integration of the queues such as in Fig. 21, it is clear that incorporating the queue velocity information can decrease the amplitude of the oscillations, which is beneficial from the managerial perspective. Using a perturbation technique, we derive an analytic expression that approximates the amplitude of oscillations very accurately. Based on the analytic expression, for any delay we can determine the coefficient δ that will minimize the amplitude of the oscillations. We note this coefficient as a function of delay and is not necessarily equal to the coefficient $\delta_{\max}$ that maximizes the delay threshold.

In the future, it would be interesting to extend this model to include terms with higher-order derivatives. Under the assumption that service managers can measure the information about the queues $(q_i^{(2)}, q_i^{(3)}, \dots)$, it would be natural to incorporate these data into the information that is provided to the customers:

$$\text{Delay announcement about } i^{th} \text{ queue} = q_i(t-\Delta) + \sum_{n=1}^{K} \delta_n q_i^{(n)}(t-\Delta), \quad K \in \mathbb{N}. \tag{6.2}$$

The equations describing such a queueing system will no longer be neutral and may be more complicated. However, such a queueing system may answer new questions. One question of significance is to determine the minimum sufficient number of higher-order derivatives (K) that should be included in order to guarantee that the queues will be stable for a given delay.

Acknowledgements Funding was provided by Division of Civil, Mechanical and Manufacturing Innovation (Grant No. 1751975).

7 Appendix

7.1 Uniqueness and existence of the equilibrium

Proof of Theorem 3.2: To check that $q_i(t) = \frac{\lambda}{N\mu}$ is an equilibrium, plug into Eq. (3.9) to get

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp(-\frac{\lambda\theta}{N\mu} - 0)}{\sum_{j=1}^{N} \exp(-\frac{\lambda\theta}{N\mu} - 0)} - \mu \cdot \frac{\lambda}{N\mu} = \frac{\lambda}{N} - \frac{\lambda}{N} = 0. \tag{7.1}$$

To show uniqueness, we will argue by contradiction. Suppose there is another equilibrium given by $\bar{q}_i$, $1 \le i \le N$, and for some i we have $q_i^* \neq \bar{q}_i$. The following condition must hold:

$$0 = \sum_{i=1}^{N} \dot{\bar{q}}_i(t) = \lambda \cdot \frac{\sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i(t-\Delta)\big)}{\sum_{j=1}^{N} \exp(-\theta \bar{q}_j(t-\Delta))} - \mu \sum_{i=1}^{N} \bar{q}_i(t), \quad \sum_{i=1}^{N} \bar{q}_i(t) = \frac{\lambda}{\mu}. \tag{7.2}$$

Hence, the mean of $\bar{q}_i$ is $\frac{\lambda}{N\mu}$ and, since $\bar{q}_i$ cannot all be equal to each other, there must exist some $\bar{q}_s$ that is smaller than the mean, and some $\bar{q}_g$ that is greater than the mean:

$$\bar{q}_s = \frac{\lambda}{\mu N} - \gamma, \qquad \bar{q}_b = \frac{\lambda}{N\mu} + \epsilon, \qquad \gamma, \epsilon > 0. \tag{7.3}$$

This leads to a contradiction:

$$\dot{\bar{q}}_s(t) = \lambda \frac{\exp\big(-\theta \bar{q}_s\big)}{\sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i\big)} - \mu \bar{q}_s = 0 \tag{7.4}$$

$$\implies \sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i\big) = \frac{\lambda}{\mu} \cdot \frac{\exp\big(-\frac{\theta\lambda}{N\mu} + \theta\gamma\big)}{(\frac{\lambda}{N\mu} - \gamma)}, \tag{7.5}$$

$$\dot{\bar{q}}_g(t) = \lambda \frac{\exp\big(-\theta \bar{q}_g\big)}{\sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i\big)} - \mu \bar{q}_g(t) \tag{7.6}$$

$$= \lambda \frac{\exp\big(-\frac{\theta\lambda}{N\mu} - \theta\epsilon\big)}{\frac{\lambda}{\mu} \cdot \frac{\exp(-\frac{\theta\lambda}{N\mu} + \theta\gamma)}{(\frac{\lambda}{N\mu} - \gamma)}} - \mu\Big(\frac{\lambda}{N\mu} + \epsilon\Big) \tag{7.7}$$

$$= -\frac{\lambda}{N}\big(1 - e^{-\theta(\epsilon+\gamma)}\big) - \mu\big(\epsilon + \gamma e^{-\theta(\epsilon+\gamma)}\big) < 0. \tag{7.8}$$

Since $\dot{\bar{q}}_g(t) \neq 0$, then $\bar{q}_i(t)$ is not an equilibrium, and the equilibrium (3.12) is unique.

7.2 Approximation to amplitude of oscillations in queues

To see how the velocity information affects the behavior of the queues after a Hopf bifurcation occurs, we need to develop approximations for the amplitude of oscillations. In Sect. 5, we find a first-order approximation to the amplitude but observe that it is not sufficiently accurate. Hence, we require a second-order approximation. The steps to determine the second-order approximation are outlined below.

This process is very closely related to the steps taken in Theorem 5.1. We begin with Eq. (5.7), and expand the time $\tau = \omega t$. Then expand our functions of interest in ϵ to the second order:

$$\begin{aligned} x(\tau) &= x_0(\tau) + \epsilon x_1(\tau) + +\epsilon^2 x_2(\tau), \\ \Delta &= \Delta_0 + \epsilon\Delta_1 + \epsilon^2\Delta_2, \quad \omega = \omega_0 + \epsilon\omega_1 + \epsilon^2\omega_2, \end{aligned}$$

where Δ_0 and ω_0 are the delay and frequency at bifurcation, so $\Delta_0 = \Delta_{\text{cr}}$ and ω_{cr}. By collecting all the terms with the like powers of ϵ into separate equations, we get equations from which we can solve for x_0 and x_1. From the equation for ϵ^0, we find that $x_0(\tau) = A\cos(\tau)$ is a solution. Next, we use the equation for ϵ^1 terms to solve for A, which has the expression given by Eq. (5.35). We can now find x_1 that has a solution of the form $x_1(\tau) = a_1\sin(\tau) + a_2\cos(\tau) + a_3\sin(3\tau) + a_4\cos(3\tau)$. The coefficients a_3 and a_4 are determined from the equation for ϵ^1 terms. We impose the initial condition $x'(0) = 0$ to ensure that the maximum amplitude is at 0, which implies $a_1 = -3a_3$. Lastly, we determine a_2 by eliminating the secular terms from the equation for ϵ^2 terms. Therefore, the second-order approximation of the amplitude of oscillations can be deduced from

$$\begin{aligned} x(\tau) &\approx x_0(\tau) + \epsilon x_1(\tau) && (7.9) \\ &= A\cos(\tau) + \epsilon\big(a_1\sin(\tau) + a_2\cos(\tau) + a_3\sin(3\tau) + a_4\cos(3\tau)\big), && (7.10) \end{aligned}$$

where the coefficients are given below:

$$\begin{aligned} A &= \sqrt{\frac{4\Delta_1(\lambda^2\theta^2 - 4\mu^2)(4 - \delta^2\lambda^2\theta^2)^2}{\theta^2(1-\delta^2\mu^2)(16\mu + \lambda^2\theta^2(4\Delta_0 - 4\delta + \delta^3\lambda^2\theta^2 - 4\delta^2\mu - 4\delta^2\Delta_0\mu^2))}}, \\ \omega_1 &= \frac{4\Delta_1\theta^2\lambda^2\left(\delta^2\mu^2 - 1\right)\sqrt{\theta^2\lambda^2 - 4\mu^2}}{\sqrt{4 - \delta^2\theta^2\lambda^2}\left(\theta^2\lambda^2\left(\delta\left(\delta^2\theta^2\lambda^2 - 4\delta\mu(\Delta 0\mu + 1) - 4\right) + 4\Delta 0\right) + 16\mu\right)}, \\ a_1 &= -3a_3 = -\Big(2A^3\theta^2\omega_0^3\left(\theta^2\lambda^2\mu\left(\delta^2\omega_0^2 + 1\right)^3 - 4\delta^3\left(\mu^2 + \omega_0^2\right)^3\right)\Big) \\ &\quad /\Big(\theta^4\lambda^4\left(\delta^2\omega_0^2 + 1\right)^3\left(\mu^2 + 9\omega_0^2\right) + 16\left(9\delta^2\omega_0^2 + 1\right)\left(\mu^2 + \omega_0^2\right)^3 \\ &\quad +8\theta^2\lambda^2\big(-9\delta^4\omega_0^8 - 6\mu^2\omega_0^2\left(\delta^2\mu^2 + 1\right) + 2\delta^2\omega_0^6(\delta\mu(9\delta\mu - 32) + 9) \\ &\quad +3\omega_0^4\left(\delta^4\mu^4 - 12\delta^2\mu^2 + 1\right) - \mu^4\big)\Big), \end{aligned}$$

$$
\begin{aligned}
a_4 = &-\frac{1}{12}\Big(A^3\theta^2\left(\theta^2\lambda^2\left(\delta^2\omega_0^2+1\right)^3\left(\mu^4+6\mu^2\omega_0^2-3\omega_0^4\right)\right.\\
&\left.+4\left(3\delta^4\omega_0^4-6\delta^2\omega_0^2-1\right)\left(\mu^2+\omega_0^2\right)^3\right)\Big)\\
&/\Big(\theta^4\lambda^4\left(\delta^2\omega_0^2+1\right)^3\left(\mu^2+9\omega_0^2\right)+16\left(9\delta^2\omega_0^2+1\right)\left(\mu^2+\omega_0^2\right)^3\\
&+8\theta^2\lambda^2\big(-9\delta^4\omega_0^8-6\mu^2\omega_0^2\left(\delta^2\mu^2+1\right)+2\delta^2\omega_0^6(\delta\mu(9\delta\mu-32)+9)\\
&+3\omega_0^4\left(\delta^4\mu^4-12\delta^2\mu^2+1\right)-\mu^4\big)\Big),\\
a_2 = &\frac{1}{12}\Big(A^5\theta^4\left(\delta^2\omega_0^2+1\right)^2\left(\delta^2\mu\omega_0^2+\mu^2\left(\delta\left(\delta\Delta_0\omega_0^2,-1\right)+\Delta_0\right)\right.\\
&\left.+\omega_0^2\left(\delta\left(\delta\Delta_0\omega_0^2-1\right)+\Delta_0\right)+\mu\right)\\
&-12A^3\theta^2\omega_0\big(\omega_1\left(\delta\left(3\delta^3\Delta_0\omega_0^4+\delta\omega_0^2(\delta(\delta\mu(2\Delta_0\mu+3)-3)\right.\right.\\
&\left.+4\Delta_0)+\delta\mu(-2\delta\mu+2\Delta_0\mu+3)-1)+\Delta_0\right)\\
&\left.-\Delta_1\omega_0\left(\delta^2\mu^2-1\right)\left(\delta^2\omega_0^2+1\right)\right)+12A^2\theta^2\big(a_1\omega_0\left(\delta^2\mu^2-1\right)\left(\delta^2\omega_0^2+1\right)\\
&+a_3\omega_0\left(\delta^2\left(\omega_0^2(\delta\mu(-3\delta\mu+8\Delta_0\mu+8)-5)+8\delta\Delta_0\omega_0^4+\mu^2\right)-1\right)\\
&+a_4(3\delta^4\Delta_0\omega_0^6+3\delta^2\omega_0^4(\delta(\delta\mu(\Delta_0\mu+1)-1)-2\Delta_0)\\
&+\omega_0^2(\delta(\delta\mu(5\delta\mu-6\Delta_0\mu-6)+1)-\Delta_0)\\
&+\mu(\delta\mu-\Delta_0\mu-1))\big)-96A\big(2\Delta_1\omega_0\omega_1(\mu\left(\mu\left(2\delta^2\right.\right.\\
&\left.\left.-2\delta\Delta_0+\Delta_0^2\right)-\delta+\Delta_0\right)+\delta^2\Delta_0^2\omega_0^4\\
&+\Delta_0\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-2)+\Delta_0)-1)+\Delta_0\omega_1^2(\delta^2\Delta_0^2\omega_0^4\\
&+\Delta_0\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-3)+\Delta_0)\\
&+\mu(2\delta-\Delta_0)(\delta\mu-\Delta_0\mu-1))+\Delta_1^2\omega_0^2(\delta^2\Delta_0\omega_0^4\\
&+\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-1)+\Delta_0)\\
&+\mu(-\delta\mu+\Delta_0\mu+1))\big)-192a_1\big(\omega_1(\delta^2\Delta_0^2\omega_0^4\\
&+\Delta_0\omega_0^2(\delta(\delta\mu(\Delta_0\mu+2)-2)+\Delta_0)\\
&+(-\delta\mu+\Delta_0\mu+1)^2)+\Delta_1\omega_0\left(\delta^2\Delta_0\omega_0^4+\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)\right.\right.\\
&\left.\left.-1)+\Delta_0)+\mu(-\delta\mu+\Delta_0\mu+1))\right)\Big)\\
&/\Big(3A^2\theta^2\left(\delta^2\omega_0^2+1\right)\left(\delta^2\Delta_0\omega_0^4+\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)\right.\\
&\left.-1)+\Delta_0)+\mu(-\delta\mu+\Delta_0\mu+1))\right.\\
&\left.+16\Delta_1\omega_0^2\left(\delta^2\mu^2-1\right)\right).
\end{aligned}
$$

To reproduce our numerical results from Sects. 5.3–5.4, set $\epsilon = 1$ and $\Delta_1 = \frac{1}{\epsilon}(\Delta - \Delta_0)$, with Δ_0 given by Eq. (3.52). Note that in the equations above there is no presence of Δ_2, because we have set $\Delta_2 = 0$. There is no equation that determines Δ_2 and Δ_1 uniquely, and the only restriction is that $\Delta = \Delta_0 + \epsilon\Delta_1 + \epsilon^2\Delta_2$. Prior to choosing Δ_2 to be 0, we experimented numerically with different combinations of Δ_1 and Δ_2 and determined that the pair $\Delta_1 = \frac{1}{\epsilon}(\Delta - \Delta_0)$ and $\Delta_2 = 0$ results in nearly the most accurate approximation.

References

1. Abboud, K., Zhuang, W.: Modeling and analysis for emergency messaging delay in vehicular ad hoc networks. In: GLOBECOM 2009—2009 IEEE Global Telecommunications Conference, pp. 1–6 (2009). https://doi.org/10.1109/GLOCOM.2009.5425839
2. Armbruster, H., Ringhofer, C., Jo, T.C.: Continuous models for production flows. In: Proceedings of the American Control Conference, vol. 5, pp. 4589–4594 (2004). ISBN 0780383354. https://doi.org/10.1109/ACC.2004.182675
3. Armony, M., Maglaras, C.: On customer contact centers with a call-back option: customer decisions, routing rules, and system design. Oper. Res. **52**(2), 271–292 (2004)
4. Armony, M., Shimkin, N., Whitt, W.: The impact of delay announcements in many-server queues with abandonment. Oper. Res. **57**(1), 66–81 (2009)
5. Belhaq, M., Sah, S.I.: Fast parametrically excited van der Pol oscillator with time delay state feedback. Int. J. Non-Linear Mech. **43**(2), 124–130 (2008)
6. Bellena, A., Guglielmi, N.: Solving neutral delay differential equations with state-dependent delays. J. Comput. Appl. Math. **229**(2), 350–362 (2009)
7. Das, S.L., Chatterjee, A.: Multiple scales without center manifold reductions for delay differential equations near Hopf bifurcations. Nonlinear Dyn. **30**, 323–335 (2002)
8. Daw, A., Pender, J.: New perspectives on the Erlang-A queue. Adv. Appl. Probab. **51**(1), 268–299 (2019)
9. Dong, J., Yom-Tov, E., Yom-Tov, G.B.: The impact of delay announcements on hospital network coordination and waiting times. Manag. Sci. **65**(5), 1969–1994 (2019)
10. Driver, R.D.: Existence and continuous dependence of solutions of a neutral functional-differential equation. Arch. Ration. Mech. Anal. **19**(2), 149–166 (1965). https://doi.org/10.1007/BF00282279
11. Eick, S.G., Massey, W.A., Whitt, W.: M_t/G/∞ queues with sinusoidal arrival rates. Manag. Sci. **39**(2), 241–252 (1993)
12. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the M_t/G/∞ queue. Oper. Res. **41**(4), 731–742 (1993)
13. Fralix, B.H., Adan, I.J.B.F.: An infinite-server queue influenced by a semi-Markovian environment. Queueing Syst. **61**(1), 65–84 (2009)
14. Freund, D., Henderson, S., Shmoys, D.: Minimizing multimodular functions and allocating capacity in bike-sharing systems (2016). arXiv preprint arXiv:1611.09304
15. Guo, P., Zipkin, P.: Analysis and comparison of queues with different levels of delay information. Manag. Sci. **53**(6), 962–970 (2007)
16. Guo, P., Zipkin, P.: The impacts of customers' delay-risk sensitivities on a queue with balking. Probab. Eng. Inf. Sci. **23**(3), 409–432 (2009)
17. Hale, J., Lunel, V.: Introduction to Functional Differential Equations. Springer Science, Berlin (1993)
18. Hassard, B.D., Kazarinoff, N.D., Wan, Y.H.: Theory and Applications of Hopf Bifurcation. Cambridge University Press, Cambridge (1981)
19. Hausman, J., McFadden, D.: Specification tests for the multinomial logit model. Econometrica **52**(5), 1219–1240 (1984)
20. Helbing, D.: Improved fluid-dynamic model for vehicular traffic. Phys. Rev. E **51**, 3164–3169 (1995). https://doi.org/10.1103/PhysRevE.51.3164

21. Ibrahim, R., Armony, M., Bassamboo, A.: Does the past predict the future? The case of delay announcements in service systems. Manag. Sci. **63**(6), 1657–2048 (2017)
22. Iglehart, D.L.: Limiting diffusion approximations for the many server queue and the repairman problem. J. Appl. Probab. **2**(2), 429–441 (1965)
23. Jouini, O., Aksin, Z., Dallery, Y.: Call centers with delay information: models and insights. Manuf. Serv. Oper. Manag. **13**(4), 534–548 (2011)
24. Ko, Y.M., Pender, J.: Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes. Stochastic Models **34**(2), 186–206 (2018)
25. Lazarus, L., Davidow, M., Rand, R.: Periodically forced delay limit cycle oscillator. Int. J. Non-Linear Mech. **94**, 216–222 (2017)
26. Lipshutz, D., Williams, R.J.: Existence, uniqueness, and stability of slowly oscillating periodic solutions for delay differential equations with nonnegativity constraints. SIAM J. Math. Anal. **47**(6), 4467–4535 (2015)
27. McFadden, D.: Modelling the choice of residential location. Cowles Foundation Discussion Papers 477, Cowles Foundation for Research in Economics, Yale University (1977). https://EconPapers.repec.org/RePEc:cwl:cwldpp:477
28. Nirenberg, S., Daw, A., Pender, J.: The impact of queue length rounding and delayed app information on Disney world queues. In: Proceedings of the 2018 Winter Simulation Conference. Winter Simulation Conference (2018)
29. Novitzky, S., Pender, J., Rand, R.H., Wesson, E.: Nonlinear dynamics in queueing theory: Determining size of oscillations in queues with delay. SIAM J. Appl. Dyn. Syst. **18**, 279–311 (2018)
30. Pender, J., Rand, R.H., Wesson, E.: Queues with choice via delay differential equations. Int. J. Bifurc. Chaos **27**(4), 1730016 (2017)
31. Pender, J., Rand, R.H., Wesson, E.: An asymptotic analysis of queues with delayed information and time varying arrival rates. Nonlinear Dyn. **91**, 2411–2427 (2018)
32. Perkins, J., Kumar, P.: Optimal control of pull manufacturing systems. IEEE Trans. Autom. Control **40**(12), 2040–2051 (1995)
33. Raina, G., Wischik, D.: Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. In: Next Generation Internet Networks, 2005. IEEE (2005)
34. Resnick, S., Samorodnitsky, G.: Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. Queueing Syst. **33**(1–3), 43–71 (1999)
35. Smith, H.: An Introduction to Delay Differential Equations with Applications to the Life Sciences. Springer Science, Berlin (2011)
36. So, Y., Kuhfeld, W.: Multinomial logit models. In: SUGI 20 Conference Proceedings (1995)
37. Tao, S., Pender, J.: A stochastic analysis of bike sharing systems. arXiv preprint arXiv:1708.08052 (2017)
38. Train, K.: Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge (2009)

Affiliations

Sophia Novitzky[1] · Jamol Pender[2] · Richard H. Rand[3] · Elizabeth Wesson[4]

✉ Jamol Pender
jjp274@cornell.edu

Sophia Novitzky
sn574@cornell.edu

Richard H. Rand
rand@math.cornell.edu

Elizabeth Wesson
enw27@cornell.edu

[1] Center for Applied Mathematics, Cornell University, 657 Rhodes Hall, Ithaca, NY 14853, USA

[2] School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, Ithaca, NY 14853, USA

[3] Department of Mathematics, Sibley School of Mechanical and Aerospace Engineering, Cornell University, 535 Malott Hall, Ithaca, NY 14853, USA

[4] Center for Applied Mathematics, Cornell University, Rhodes Hall 657, Ithaca, NY 14853, USA