QUEUEING SYSTEMS
VIA DELAY DIFFERENTIAL EQUATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sophia Novitzky

May 2020

QUEUEING SYSTEMS VIA DELAY DIFFERENTIAL EQUATIONS

Sophia Novitzky, Ph.D.

Cornell University 2020

Many service systems use internet or smartphone app technology to notify customers about their expected waiting times or queue lengths via delay announcements, allowing customers to decide which queue to join. 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. We model multiple-queue systems through delay differential equations, and study how the delay in information affects the dynamics of the queues.

When the delay is sufficiently large, the queues may oscillate indefinitely throughout time. We develop accurate approximations for the amplitude of these oscillations by implementing two numerical methods. The first technique is a classical analytic method that yields a closed-form approximation in terms of the model parameters. The second approximation method is novel, and it uses a statistical technique to deliver highly accurate approximations over a wider range of parameters.

The oscillations in queue lengths are generally undesirable both for the service providers and the customers. This motivates us to explore how the delay announcement can be used to limit the oscillations. We show that, in some cases, using information about queue's velocity (the rate at which the queue length is changing) in the delay announcement can eliminate oscilla-

tions 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.

When the delay in information is caused by customers traveling to the queues, the delay may vary from customer to customer. We propose a queueing model that treats the individual's delay not as a constant, but as a random variable drawn from a fixed distribution. This generalized model allows us to identify the properties of dynamics that are independent of the delay distribution, such as the existence and uniqueness of the equilibrium state. However, the stability of the equilibrium and the presence of oscillations depend on the delay distribution. We therefore give an overview of the system's stability region for different common delay distributions, and finally offer a numerical method of approximating the stability region when the distribution is unknown.

BIOGRAPHICAL SKETCH

Sophia thoroughly enjoyed her time at Cornell. Learning from the best, teaching the brightest, and suntanning by the waterfalls made it worthwhile. Falling in love, getting married, and having a baby made it unbeatable.

DEDICATION

To my father, Dr. Victor Novitzky, who taught me to find joy in scientific discovery. To my mother, Dr. Irina Novitzky, who taught me to never give up and get the job done.

ACKNOWLEDGEMENTS

I thank the Lord for blessing me with the presence of many wonderful people. I'd like to address the key people whose contributions to my success are invaluable.

Upon joining Cornell University, I was unprepared for the rigor of coursework and the self-discipline necessary to succeed in the academic research environment. It took encouragement from one of my professors, Jamol Pender, to get me through that first semester. Afterwards, he became my academic advisor. I drew inspiration from his work ethic, creativity in approaching research questions, and enthusiasm for everything related to queues. Jamol approached my extravagance with wisdom and empathy when three years later I told him that I need to move to Massachusetts to follow my love, and that I am expecting to become a mother. Jamol Pender's intellect capacity and human side make him an unparalleled advisor. This document could not exist without his guidance, patience, and mathematical talent.

This thesis also couldn't be possible without my co-advisor, Richard Rand, and his efforts to teach me about delay differential equations, perturbations methods, center manifolds, bifurcations, floquet theory, and other beauties in math. I equally appreciate him as a role model, who so clearly loves what he does in life - taking his grandchildren to a science museum, visiting Grand Canyon with his wife, playing musical instruments with her, and certainly being a professor. Richard's quote "I plan to die with chalk in my hands" summarizes how one can only hope to feel about their job!

I'd like to thank my "unofficial co-advisor" Dr. Elizabeth Wesson who worked with me throughout my first two research projects. Solely out of her generosity (and the convenience of having a shared office space), Elizabeth helped me with my Mathematica scripts and confusions over Lindstedt's method many more times than was reasonable.

I am grateful to professors Timothy Healey and Jim Dai for their mentorship and their active role as my Special Committee members. Over the five years of knowing Professor Healey, I consulted with him on many issues (while struggling with the material in his course on Partial Differential Equations, facing initial troubles of choosing a research area, as well as figuring out how to continue doing research after reallocating from Ithaca to Boston). He has always been patient, receptive, humorous, and extremely helpful. In the two years that I have known Professor Dai, he has been highly punctual, responsive, and flexible with his own very demanding schedule (despite a twelve hour difference with China!), making my job as a graduate student much easier.

Another person that has always been ready to take the load off my shoulders is the Center for Applied Mathematics administrative manager Erika Fowler-Decatur. She makes all CAM students aware that she genuinely cares about our individual success. Being both resourceful and highly attentive, she often anticipated my needs and provided help before I even knew to ask for it. Erika also happened to be my neighbor, and I will cherish the memories of our countless conversations, eating berries from her mulberry tree,

getting regular visits from Gryffindor, and taking Betty and Ashes for walks.

I was surprised by the number of faculty from the Mathematics department who always keep their doors open to CAM students. Many got to know me personally and gave me valuable advice throughout the years. Among them is Steve Strogatz, John Guckenheimer, and Alex Vladimirsky. I also thank my mentors outside of Cornell who helped me over the years - my undergraduate advisor Lizette Zietsman, physics teacher Ed Linz, pastor Reggie Tuck, and undergraduate research mentors Burt Tilley and Suzanne Weaks.

Unrelated to this thesis but most related to my time in graduate school, I thank Kevin O'Brien for resolutely changing the course of my life. He introduced me to the man I married, and two years later pulled my husband and me into Boston by setting up Ryan with an offer we could not refuse. Kevin's high integrity, impeccable taste in people, and strong leadership ability distinguish him even among my friends.

Finally, I acknowledge my family for all their support and unconditional love. My amazing husband Ryan, my parents Irina and Victor, and my parents-in-law Susan and Tom help me more than I could ever ask for. I thank them for having faith in me even when I'm at my worst. I am grateful to Julia and Kyle for taking me into their home and for being solid role models. And, I thank Emily for teaching me to not take myself too seriously.

:-)

Contents

1 Introduction

With access to more information, consumers are adjusting their behavior and expectations of the services they select. Customers are now actively seeking information about competing businesses prior to choosing their service provider. Figuring out the waiting time is often a trivial task for the customer, requiring nothing more than a phone call or a quick peek at a mobile phone application (app). Recently, services like the bike sharing networks, U-Haul truck rental locations, hospital emergency rooms, amusement parks like Disney World, and even restaurants, began to provide such information in the form of delay announcements to the inquiring potential customers. The phone application CycleFinder in Figure 1, for example, provides customers a map of the bike racks and the number of available bikes at each location, so that the bicyclists don't waste their time checking empty bike racks. Short waiting times can attract potential clients to a newly opened business, while extensive waiting lines can instead deter customers from joining the queue.

The availability of the waiting time information impacts the decision patterns of individual customers, and thus the dynamics of the queueing system as a whole. As a result, much research queueing theory is dedicated to quantifying the impact of delay announcements on the queue length process or the virtual waiting time process; see for example [23, 4, 16, 20, 5, 17, 26, 27, 1, 2, 43].

Accounting for the delay in information due to customers' travel time is

Figure 1: Bike-sharing network app.

what sets our work apart from the existing literature in queueing theory. We consider a multiple-queue system where each queue corresponds to a separate geographical location of the same service (such as competing restaurants in a neighborhood, for example). The customers can choose which queue to join giving a higher preference to the location with the shortest waiting time. After deciding which queue to join, the customer, such as Mr. Green in Figure 2, has to commute to the service location. Mr. Green's commute causes a time delay prior him securing a spot in line. In the meantime, other travelling customers may have joined the same queue, so the queue length as well as the waiting time may have changed during that travelling time. The waiting time information used by customers is therefore somewhat outdated and unreliable, causing ripple effects throughout the queueing system.

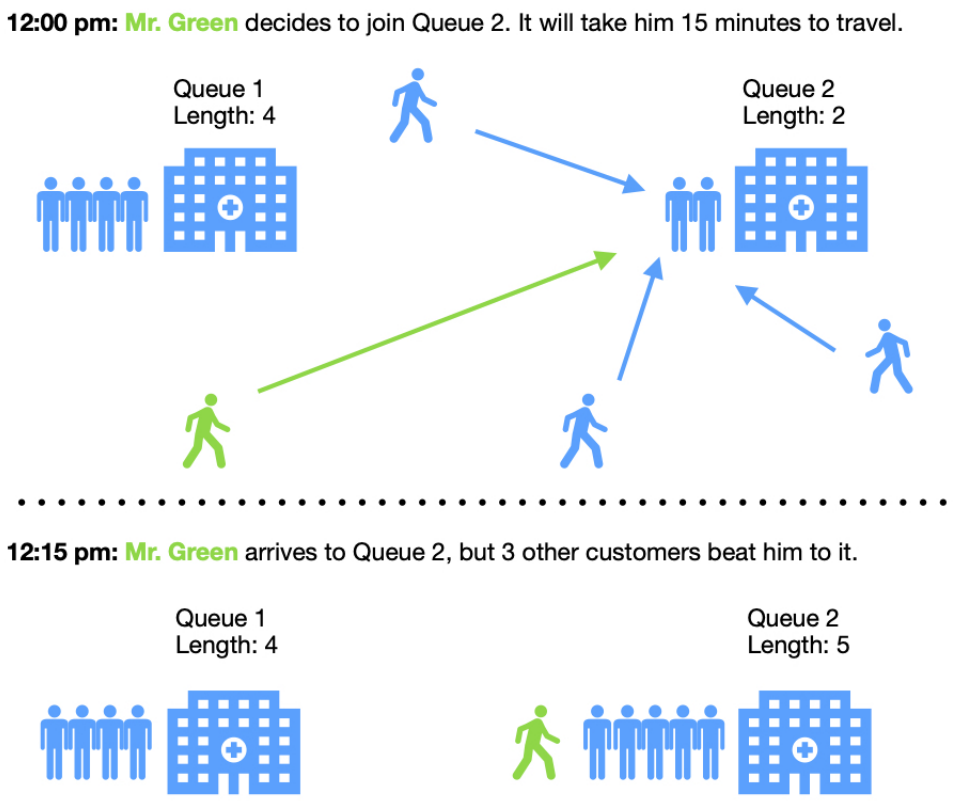

Figure 2: Mr. Green's attempt to join the shortest queue fails due to delay in information.

The delay in information introduces unique complexities to the dynamics that otherwise would be missed. For example, Pender et al. [33] shows that when all customers experience the same delay and this delay becomes sufficiently large, the queueing system will bifurcate and the queues will oscillate indefinitely. If in the same queueing system the delay is decreased or eliminated altogether, the queues will stop oscillating over time and converge to an equilibrium length.

Hence, accounting for the delay is important, but it is an open question

of how to do so accurately. In some physical settings, it may be sufficient to assume that all customers are delayed identically. For example, the delay may actually be caused by the service manager collecting and updating the delay announcement, which is then distributed to all customers simultaneously through a phone application. In many physical systems, however, each customer takes a slightly different time to arrive to the queue of their choice as seen in Figure 3, causing the delay in information to vary from customer to customer. This thesis proposes a way to model the queueing systems under different assumptions about customers' waiting times, and uncovers the implications of these assumptions.

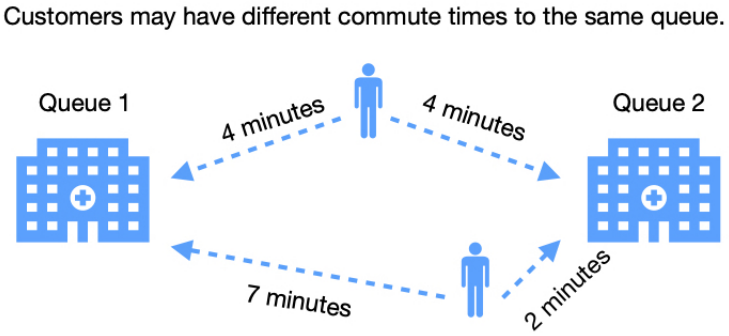

Figure 3: An individual's commute time can vary.

We model the queues through deterministic fluid models. This may seem as a counter-intuitive choice given that queues usually comprise of discrete units such as the number of people, jobs, or automobiles. However, large scale queueing systems can be well approximated by fluid models. These approximations are common in queueing literature [23, 5, 33], where the queues are modeled as stochastic processes and then shown to converge in some lim-

iting 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 park where the demand for service is large [35, 22, 5, 3, 32].

1.1 Thesis Overview

In this thesis we present fluid models of multiple-queue systems where customers are given information about the current length of each queue. Customers choose which queue to join, giving higher preference to the shorter queue. They then proceed to travel to the queue of their choice, causing a time lag between the point of receiving the delay announcement and joining the line.

Chapter 2 assumes that all customers experience the same delay, which we refer to as the *constant delay model*. We assume there are two queues, and characterize the queue lengths as a system of two delay-differential equations (DDEs). We analyze the asymptotic behavior, showing that the queues are locally stable when the delay is below a certain threshold. If the delay exceeds the threshold, a Hopf bifurcation may occur causing the queue lengths to oscillate indefinitely. The main objective of this chapter is to accurately approximate the size of these oscillations. We first use Lindstedt's method, a classic perturbation technique. Then we propose a novel method of approximation, which we call the *slope function method*. The accuracy of the two methods is compared based on the amplitude found from numerically

integrating the system of DDEs.

Chapter 3 follows the same structure as Chapter 2, except it considers a somewhat different queueing model where the assumption that all customers experience the same delay is dropped. Instead, the individual's travel time to the queue is modeled as a random variable drawn from a uniform distribution. The new queueing model that accounts for the variability in the customers' delay is equivalent to a moving average model, and the queue lengths are now modelled by a system of two functional differential equations (FDEs) that are more complicated than the DDEs from Chapter 2. We determine the conditions under which the queues oscillate, and then apply Lindstedt's and the slope function methods to approximate the amplitude of oscillations. Chapter 3 concludes by comparing the two methods and summarizes their respective strengths and weaknesses.

Chapter 4 returns to the queueing system with constant delay originally presented in Chapter 2, and generalizes it to N queues (instead of two queues). Since oscillations in queues are generally undesirable both for the service managers and the customers, we propose a novel type of delay announcement that aims to decrease the amplitude of oscillations, and in some cases prevent the oscillations altogether. Instead of reporting the current length of each queue to the customers, the service manager discloses a weighed sum of the queue length and the rate with which the queue is changing (the velocity). This information is often easily available to the service managers, and Chapter 4 rigorously argues why this new delay announcement

increases the overall stability of the queueing system.

Chapter 5 formulates a novel generalized N-queue system that treats the delay of an individual customer as a random variable from an arbitrary distribution. The queueing systems from Chapters 2 and 3 are special cases of this model. We are able to prove that the dynamics of the queueing system retain many properties regardless of the delay distribution. There exists a unique equilibrium state and it is guaranteed to be locally stable when a certain parameter relationship holds. Further, the equilibrium can become unstable only if a Hopf bifurcation occurs, so by finding the Hopf curve in the system's parameter space we can determine the stability region of the queueing system. The presence and the location of the Hopf curve in the parameter space, however, depends on the delay distribution, so we provide an overview of some common distributions and the resulting stability regions. Lastly, we discuss how to approximately determine the stability region for a queueing system where the delay distribution is unknown. We propose a numerical method that approximates the stability region relying only on information about the average delay and some central moments of the distribution. Central moments can have the advantage of being easy to estimate with data from service systems through sampling the travel times of the incoming customers.

Lastly, Chapter 6 summarizes the main results from each chapter, and concludes by discussing possible future directions of our research.

2 Approximating Amplitude of Oscillations for Queueing System with Constant Delay

Internet and mobile services often provide waiting time or queue length information to customers. In the case of multiple lines, this information allows a customer to better decide which line to join. Unfortunately, there is usually a delay associated with waiting time information. Either the information itself is stale, or it takes time for the customers to travel to the service location after having received the information.

Recent work by Pender et al. [33] uses functional dynamical systems as limiting models for stochastic queueing systems. This work has shown that if information is delayed long enough, it results in unwanted oscillations may form in the queue lengths throughout time. In this chapter, we investigate the model further and establish that the oscillations are due to supercritical Hopf bifurcations. It is unknown how large the oscillations are when a Hopf bifurcation occurs. To answer this question, we implement two methods for approximating the amplitude of these oscillations. The first approximation is a classical perturbations technique called Lindstedt's method, and it yields a closed-form approximation in terms of the model parameters. The second method, which we call the *slope function method*, is a numerical technique we developed specifically to extend the range of parameters where the Lindstedt's approximation maintains accuracy.

Based on numerical results, we observe that the slope function method

successfully reduces the maximum error in approximation over a range of parameters by $60 - 75\%$ when compared to Lindstedt's method. In the context of queueing models, the accuracy of approximation matters because the amplitude of queue oscillations can provide valuable insights such as the average waiting time during busier hours, the longest waiting time a customer can experience, and the optimal moment for joining a queue that will guarantee the quickest service. Moreover, our slope function method is not restricted to queueing models and can be applied to any system where Hopf bifurcations are observed.

2.1 Chapter Outline

In this chapter considers a model originally presented in [33] and [34] as fluid limit of a stochastic queueing model. There are two queues and customers decide which queue to join based on information about the queue length that is delayed. In Subsection 2.3 we describe the qualitative behavior of the queues, stating the conditions for a unique stable equilibrium as well as the conditions for supercritical Hopf bifurcations. We focus on the behavior of the queues when the stable equilibrium transitions into a stable limit cycle, and approximate the amplitude of the resulting oscillations. Subsections 2.4 - 2.5 use Lindstedt's method, which is accurate on a limited range of parameters. To broaden this range, in Subsections 2.6 - 2.7 we implement the slope function method, which is a technique that uses known amplitude of a small subset of queues and extrapolates it for a larger set of parameters.

Overall, this method achieves higher accuracy than Lindstedt's method when compared over a range of model parameters.

2.2 The Queueing Model

In a model with two queues visualized by Figure 4, customers arrive at a rate $\lambda > 0$. Each customer is given a choice of joining either queue. The customer is told the length of each queue, and is likely to prefer the shorter queue. The probability p_i of a customer joining the i^{th} queue is given by the Multinomial Logit Model (MNL)

$$p_i(q(t), \Delta) = \frac{\exp\big(-q_i(t-\Delta)\big)}{\exp\big(-q_1(t-\Delta)\big) + \exp\big(-q_2(t-\Delta)\big)}, \tag{1}$$

where $q_i(t)$ is the length of i^{th} queue at time t. The MNL is commonly used to model customer choice in fields of operations research, economics, and applied psychology [41, 21, 31, 42]. The delay $\Delta > 0$ accounts for the customers' travel time to the service location, or for the time lag between when the service manager measures the queue length and discloses this information to customers. The model assumes an infinite-server queue, which is customary in operations research literature [15, 24, 38]. This assumption implies that the departure rate for a queue is the service rate $\mu > 0$ multiplied by the total number of customers in that queue. Therefore, the queue lengths can

be described by the following DDEs

$$\dot{q}_1(t) = \lambda \cdot \frac{\exp\big(-q_1(t-\Delta)\big)}{\exp\big(-q_1(t-\Delta)\big) + \exp\big(-q_2(t-\Delta)\big)} - \mu q_1(t) \tag{2}$$

$$\dot{q}_2(t) = \lambda \cdot \frac{\exp\big(-q_2(t-\Delta)\big)}{\exp\big(-q_1(t-\Delta)\big) + \exp\big(-q_2(t-\Delta)\big)} - \mu q_2(t) \tag{3}$$

for $t > 0$, with initial conditions specified by non-negative continuous functions f_1 and f_2

$$q_1(t) = f_1(t), \quad q_2(t) = f_2(t), \quad t \in [-\Delta, 0]. \tag{4}$$

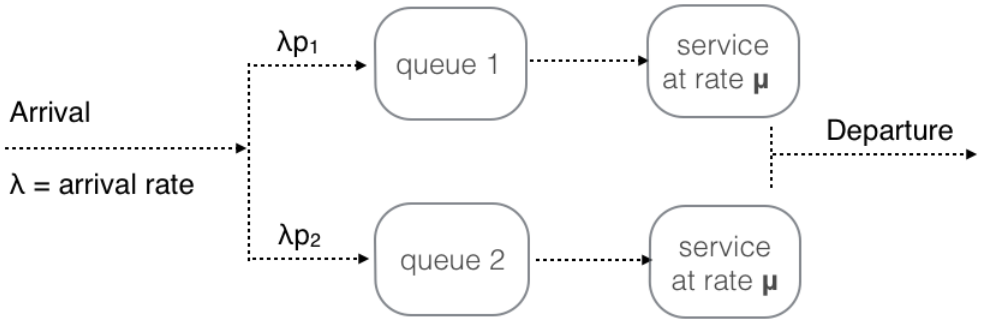

Figure 4: Customers going through a two-queue service system.

It is worth noting that Equations (2) - (3) can be uncoupled when the sum and the difference of q_1 and q_2 is taken. The system is then reduced to

the equations

$$\dot{v}_1(t) = \dot{q}_1(t) - \dot{q}_2(t) = \lambda \tanh\Big(-\frac{1}{2}v_1(t-\Delta)\Big) - \mu v_1(t), \tag{5}$$

$$\dot{v}_2(t) = \dot{q}_1(t) + \dot{q}_2(t) = \lambda - \mu v_2(t), \tag{6}$$

where $v_2(t)$ is solvable, and the equation for $v_1(t)$ is of a form commonly studied in the literature. Many papers, such as [25, 44, 46, 45, 9], prove properties for models similar to ours. In [46], the author uses asymptotic analysis to prove uniqueness and stability of the slowly oscillating periodic solutions that occur under certain parameter restrictions. The authors in [39] study the floquet multipliers. We complement these results by developing an approximation for the amplitude of the oscillations near the first bifurcation point.

2.3 Asymptotic Behavior

In this Section, we discuss the qualitative behavior of the queueing system given by Equations (2) - (3). We will begin by establishing the existence and uniqueness of the equilibrium.

Theorem 2.1. *For sufficiently small Δ, the unique equilibrium to the system of N equations*

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\big(-q_i(t-\Delta)\big)}{\sum_{j=1}^{N}\exp\big(-q_j(t-\Delta)\big)} - \mu q_i(t) \qquad \forall i = 1, 2, \cdots, N \tag{7}$$

is given by

$$q_i^* = \frac{\lambda}{N\mu} \quad \forall i = 1, 2, \cdots, N. \tag{8}$$

Proof. See the Appendix for the proof. □

Therefore the equilibrium lengths of the queues are

$$q_1^* = q_2^* = \frac{\lambda}{2\mu}. \tag{9}$$

Next, we consider the stability of the equilibrium, which can be determined by the stability of the linearized system of equations [18, 40]. Section 7.1 in the Appendix linearizes the system of equations (2) - (3) and separates the variables, reducing the system from two unknown functions to one:

$$\overset{\bullet}{\tilde{v}}_2(t) \;\; = -\tfrac{\lambda}{2} \cdot \tilde{v}_2(t - \Delta) - \mu \tilde{v}_2(t). \tag{10}$$

Assuming a solution of the form $\tilde{v}_2(t) = \exp(\Lambda t)$, the characteristic equation is found to be

$$\Phi(\Lambda, \Delta) = \Lambda + \frac{\lambda}{2} \exp(-\Lambda\Delta) + \mu = 0. \tag{11}$$

The equilibrium is stable whenever the real part of every eigenvalue Λ is negative. It is evident from the characteristic equation that any real root Λ must be negative. However, there are infinitely many complex roots. In the next result, we will show that for a sufficiently small delay, all complex eigenvalues have negative real parts.

Proposition 2.2. *For Equations* (2) - (3)*, as the delay approaches 0, i.e.* $\Delta \to 0^+$*, the real part of any complex eigenvalue approaches negative infinity.*

Proof. When $\Delta = 0$, the characteristic equation (11) has only one eigenvalue, namely $\Lambda = -\frac{\lambda}{2} - \mu$. When the delay is raised above 0, the characteristic equation becomes transcendental and an infinite sequence of roots is born. Since $\Phi(\Lambda, \Delta)$ is continuous with respect to both Λ and Δ, each eigenvalue Λ must be continuous with respect to Δ. Hence the real part of Λ must go to positive infinity or to negative infinity as the delay approaches 0. However, any root with positive real part is bounded as shown in the Appendix by Proposition 7.1, so the real part of any complex eigenvalue must go to negative infinity. □

By Proposition 2.2, all eigenvalues have negative real parts when Δ is small, so the equilibrium is stable until a pair of complex eigenvalues reaches the imaginary axis. To find when the equilibrium may become unstable, we assume $\Lambda = i\omega_{cr}$ with $\omega_{cr} > 0$, plug Λ into the characteristic equation (11), and separate the real and imaginary parts into two equations. We use the trigonometric identity $\cos^2(\omega\Delta) + \sin^2(\omega\Delta) = 1$ to find

$$\Delta_{cr}(\lambda, \mu) = \frac{2\arccos(-2\mu/\lambda)}{\sqrt{\lambda^2 - 4\mu^2}}, \quad \omega_{cr} = \sqrt{\frac{\lambda^2}{4} - \mu^2}. \tag{12}$$

For ω_{cr} to be real and nonzero the condition $\frac{\lambda^2}{4} - \mu^2 > 0$ must hold, so $\lambda > 2\mu$. If this condition is met, the equilibrium becomes unstable when Δ exceeds the **smallest positive root** of Δ_{cr} from Equation (12).

Theorem 2.3. *If $\lambda < 2\mu$, the equilibrium is locally stable for all $\Delta > 0$. If $\lambda > 2\mu$, the equilibrium is locally stable when Δ is less than the smallest positive root of Δ_{cr}.*

Proof. As discussed above, all eigenvalues of the characteristic equation are on the negative real side of the complex plane, unless $0 \neq \omega_{cr} \in \mathbb{R}$, and the delay reaches Δ_{cr}. □

Figures 5 - 6 show the behavior of the queues before and after the equilibrium becomes unstable. As suggested by Figure 6 and proved by Theorem 2.4, the conditions (12) specify where the Hopf bifurcations occur. We note that if $\lambda > 2\mu$, there will be infinitely many Hopf bifurcations as the delay grows, since the expression for Δ_{cr} has infinitely many roots.

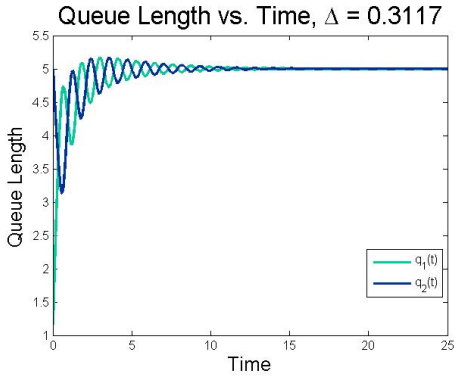

Figure 5: $\lambda = 10$, $\mu = 1$, $\Delta < \Delta_{cr}$.

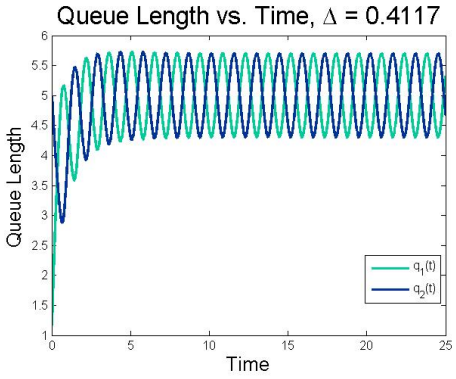

Figure 6: $\lambda = 10$, $\mu = 1$, $\Delta > \Delta_{cr}$.

Theorem 2.4. *If $\lambda > 2\mu$, a Hopf bifurcation occurs at $\Delta = \Delta_{cr}$, where Δ_{cr} is given by*

$$\Delta_{cr}(\lambda, \mu) = \frac{2\arccos(-2\mu/\lambda)}{\sqrt{\lambda^2 - 4\mu^2}}. \tag{13}$$

Proof. When $\Delta = \Delta_{cr}$, there is a a pair of purely imaginary eigenvalues Λ and $\bar{\Lambda}$. Further, Re $\Lambda'(\Delta_{cr}) > 0$. We show this by introducing $\Lambda = \alpha(\Delta) + i\omega(\Delta)$ into the characteristic equation (11), separating the real and imaginary parts into two equations, and implicitly differentiating with respect to delay. We find $\frac{d\omega}{d\Delta}(\Delta_{cr})$ to be

$$\frac{d\omega}{d\Delta}(\Delta_{cr}) = \frac{\frac{\lambda}{2}e^{-\alpha\Delta}\Big(\cos(\omega\Delta)\omega - \sin(\omega\Delta)(\alpha'\Delta + \alpha)\Big)}{1 - \frac{\lambda\Delta}{2}\cos(\omega\Delta)e^{-\alpha\Delta}} = -\frac{\omega_{cr}(\alpha'\Delta_{cr} + \mu)}{1 + \mu\Delta_{cr}}, \quad (14)$$

This result is used to determine Re $\Lambda'(\Delta_{cr}) = \frac{d\alpha}{d\Delta}(\Delta_{cr})$:

$$\alpha' - \frac{\lambda}{2}e^{-\alpha\Delta}(\alpha'\Delta + \alpha)\cos(\omega\Delta) - \frac{\lambda}{2}e^{-\alpha\Delta}\sin(\omega\Delta)(\omega'\Delta + \omega) = 0, \quad (15)$$

$$\frac{d\alpha}{d\Delta}(\Delta_{cr}) = \frac{\omega_{cr}^2}{(1 + \mu\Delta_{cr})^2 + \omega_{cr}^2\Delta_{cr}^2} > 0 \quad \forall\Delta_{cr} > 0. \quad (16)$$

where we use that at Δ_{cr}, $\alpha = 0$, $\omega = \omega_{cr}$, $\sin(\Delta_{cr}\omega_{cr}) = \frac{2\omega_{cr}}{\lambda}$, and $\cos(\Delta_{cr}\omega_{cr}) = -\frac{2\mu}{\lambda}$.

At each root of Δ_{cr} there is one purely imaginary pair of eigenvalues, but all other eigenvalues necessarily have a nonzero real part. This implies that all roots $\Lambda_j \neq \Lambda, \bar{\Lambda}$ satisfy $\Lambda_j \neq m\Lambda$ for any integer m. Hence, all conditions of the infinite-dimensional version of the Hopf theorem from [18] are satisfied, so a Hopf bifurcation occurs at every root of Δ_{cr}. □

Once the equilibrium becomes unstable, a limit cycle emerges. The following theorem shows that the resulting limit cycle is stable.

Theorem 2.5. *The Hopf bifurcations given by Theorem 2.4 are supercritical, i.e. each Hopf produces a stable limit cycle in its center manifold.*

Proof. One way to establish stability of limit cycles is by the method of *slow flow*, or the method of multiple scales. This method has previously been applied to systems of DDEs [10, 6, 29]. Another standard way to determine the stability of limit cycles is by showing that the floquet exponent has negative real part, as outlined in Hassard et al. [19]. In this theorem, we follow the first approach (the method of slow flow), but for the interested reader we include the floquet exponent method in the Appendix 7.1. We note that the results of the two methods agree.

We consider the third order polynomial expansions of q_1 and q_2 about the equilibrium. The resulting equations can be uncoupled, with the function of our interest given by

$$\dot{\tilde{v}}_2(t) = \lambda\left(-\frac{\tilde{v}_2(t-\Delta)}{2} + \frac{\tilde{v}_2^3(t-\Delta)}{24}\right) - \mu\tilde{v}_2(t). \tag{17}$$

For details, see the Appendix 7.1. We set $\tilde{v}_2(t) = \sqrt{\epsilon}x(t)$ in order to prepare the DDE for perturbation treatment, and replace the independent variable t by two new time variables $\xi = \omega t$ (stretched time) and $\eta = \epsilon t$ (slow time). The delay and frequency are expanded about the critical Hopf values, $\Delta = \Delta_{cr} + \epsilon\alpha, \quad \omega = \omega_{cr} + \epsilon\beta$, so $\dot{x}$ becomes

$$\dot{x} = \frac{dx}{dt} = \frac{\partial x}{\partial \xi}\frac{d\xi}{dt} + \frac{\partial x}{\partial \eta}\frac{d\eta}{dt} = \frac{\partial x}{\partial \xi}\cdot(\omega_{cr} + \epsilon\beta) + \frac{\partial x}{\partial \eta}\cdot\epsilon. \tag{18}$$

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

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

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

$$\frac{dx}{dt} = \omega_{cr}\frac{\partial x_0}{\partial\xi} + \epsilon\beta\frac{\partial x_0}{\partial\xi} + \epsilon\frac{\partial x_0}{\partial\eta} + \epsilon\omega_{cr}\frac{\partial x_1}{\partial\xi}. \tag{19}$$

After the proposed transformations are carried out, the DDE (17) can be separated into two equations by collecting the terms with like powers of ϵ,

$$\omega_{cr}\frac{\partial x_0}{\partial\xi} + \frac{\lambda}{2}\tilde{x}_0 + \mu x_0 = 0, \tag{20}$$

$$\omega_{cr}\frac{\partial x_1}{\partial\xi} + \frac{\lambda}{2}\tilde{x}_1 + \mu x_1 = -\beta x_{0\xi} - x_{0\eta} + \frac{\lambda}{2}\big(\beta\Delta_{cr} + \alpha\omega_{cr}\big)\cdot\tilde{x}_{0\xi} + \frac{\lambda}{24}\tilde{x}_0^3. \tag{21}$$

Equation (20) shows that x_0 can be written as $x_0(t) = A(\eta)\cos(\xi) + B(\eta)\sin(\xi)$. Eliminating the secular terms $\sin(\xi)$ and $\cos(\xi)$ in Equation (21), we get two equations that involve $\frac{d}{d\eta}A(\eta)$ and $\frac{d}{d\eta}B(\eta)$, and we remove the delay terms by using Equation (20). We switch into polar coordinates by introducing $R(\eta) = \sqrt{A(\eta)^2 + B(\eta)^2}$, and we find $\frac{dR}{d\eta}$:

$$\frac{dR}{d\eta} = -\frac{R\Big((\Delta_{cr}\lambda^2 + 4\mu)R^2 - 16\alpha(\lambda^2 - 4\mu^2)\Big)}{16(4 + \Delta_{cr}^2\lambda^2 + 8\Delta_{cr}\mu)}. \tag{22}$$

Since $R \geq 0$ by definition, the two equilibrium points are $R_1 = 0$, which is unstable, and $R_2 = \sqrt{\frac{16\alpha(\lambda^2 - 4\mu^2)}{(\Delta_{cr}\lambda^2 + 4\mu)}}$, which is stable. Thus the limit cycle born

when Δ exceeds any root of Δ_{cr} is locally stable in its center manifold. $\square$

To summarize, the queues converge to an equilibrium regardless of the delay when $\lambda < 2\mu$. However, when $\lambda > 2\mu$, infinitely many pairs of complex eigenvalues will (one by one) cross the imaginary axis from negative to positive real half of the complex plane as the delay increases. Each point of the delay where a pair of eigenvalues reaches the imaginary axis results in a supercritical Hopf bifurcation, and is denoted by the critical delay Δ_{cr}. Figure 7 displays the curves along which the Hopf bifurcations occur, as a function of the arrival rate λ. For any λ, the queues become unstable when the delay exceeds the first Hopf curve, at which point a stable limit cycle is established. We will now approximate the amplitude of the limit cycle near the bifurcation point via Lindstedt's method.

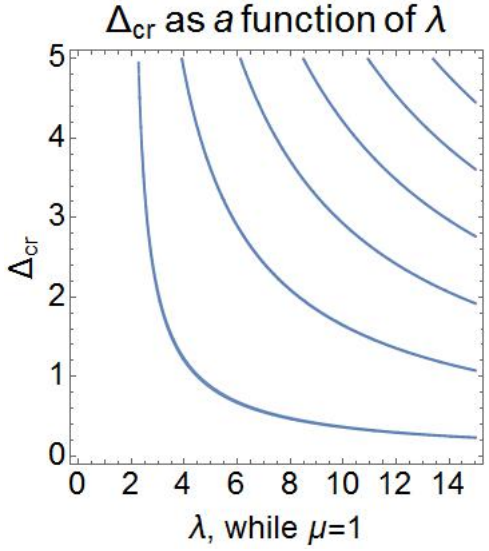

Figure 7: The Hopf curves for $\mu = 1$.

2.4 Main Steps of Lindstedt's Method

Lindstedt's method was originally formulated for finite-dimensional differential equations, but has been later extended to delay differential equations. Work such as [14] and [36] apply Lindstedt's method for equations with delays. We synthesize the main steps into four essential parts, which provide clarity to the reader who might be unfamiliar with asymptotic techniques. We outline a complete methodology for replicating our results for other types of models.

1. The third order Taylor expansions of the DDEs (2) - (3) can be uncoupled, yielding $\tilde{v}_2$ from Equation (17) as our function of interest. We stretch the time t and scale the function $\tilde{v}_2$:

$$\tau = \omega t, \quad \tilde{v}_2(t) = \sqrt{\epsilon} v(t). \tag{23}$$

2. We approximate the unknown function $v(t)$, the delay Δ, and the oscillation frequency ω by performing asymptotic expansions in ϵ:

$$v(t) = v_0(t) + \epsilon v_1(t) + ..., \quad \Delta = \Delta_0 + \epsilon \Delta_1 + ..., \quad \omega = \omega_0 + \epsilon \omega_1 + ... \tag{24}$$

3. After the expansions from Equation (24) are constructed, the resulting equation can be separated by the terms with like powers of ϵ (ϵ^0 and

ϵ^1):

$$\mu v_0(\tau) + \tfrac{\lambda}{2} v_0(\tau - \Delta_0\omega_0) + \omega_0 v_0'(\tau) = 0, \tag{25}$$

$$\mu v_1(\tau) + \tfrac{\lambda}{2} v_1(\tau - \Delta_0\omega_0) + \omega_0 v_1'(\tau) \tag{26}$$

$$+\omega_1 v_0'(\tau) - \tfrac{1}{24}\lambda v_0^3(\tau - \Delta_0\omega_0) - \tfrac{1}{2}\lambda(\Delta_1\omega_0 + \Delta_0\omega_1) v_0'(\tau - \Delta_0\omega_0) = 0.$$

Equation (25) is satisfied by the solution $v_0(\tau) = A_v \sin(\tau)$, which is expected since v_0 describes the queue behavior at the Hopf bifurcation where a limit cycle is born. It can be verified by substitution of $\Delta_0 = \Delta_{cr}$ and $\omega_0 = \omega_{cr}$. Furthermore, the equation for $v_1(\tau)$ contains a homogeneous and a non-homogeneous part. The homogeneous part $v_1^H(\tau)$ satisfies an equation which is identical to the Equation (25), so any linear combination of $\sin(\tau)$ and $\cos(\tau)$ will satisfy the equation for $v_1^H(\tau)$. To avoid secular terms in the non-homogeneous solution, the coefficients of $\sin(\tau)$ and $\cos(\tau)$ resulting from v_0 in Equation (26) must vanish. This gives two equations with two unknowns, A_v and ω_1.

4. The resulting equations can be solved for A_v and ω_1. Substituting in $\Delta_0 = \Delta_{cr}$ and $\omega_0 = \omega_{cr}$, the results are

$$\omega_1 = -\frac{(\Delta - \Delta_{cr})\lambda^2(\lambda^2 - 4\mu^2)^{3/2}}{4\Big(2\lambda^2\mu - 8\mu^3 + \lambda^2\sqrt{\lambda^2 - 4\mu^2}\arccos(-\frac{2\mu}{\lambda})\Big)}, \tag{27}$$

$$A_v(\Delta) = \sqrt{\Delta - \Delta_{cr}} \cdot \sqrt{\frac{8(\lambda^2 - 4\mu^2)^2}{2\lambda^2\mu - 8\mu^3 + \lambda^2\sqrt{\lambda^2 - 4\mu^2}\arccos(-\frac{2\mu}{\lambda})}}. \tag{28}$$

Amplitude of the Queues

The function $\tilde{v}_2$ from Equation (17) attains a steady state amplitude approximately given by A_v. A change of variables reveals the amplitude of q_1 and q_2, showing that the steady state of queues up to a phase shift is given by

$$q_1(t) \to \frac{\lambda}{2\mu} + \frac{1}{2}A_v \sin(\omega t), \qquad q_2(t) \to \frac{\lambda}{2\mu} - \frac{1}{2}A_v \sin(\omega t), \tag{29}$$

where ω is the frequency of oscillations and the amplitude is $\frac{1}{2}A_v$.

2.5 Numerical Results of Lindstedt's Method

Although Figures 8 - 9 demonstrate that the amplitude approximation from Equation (29) matches the behavior of the queues quite well, they do not reveal whether the approximation remains equally accurate when the model parameters vary. In this Section, we determine under what conditions the approximation of the steady state amplitude is accurate. We consider the queue lengths to be determined with sufficient accuracy by numerical integration of Equations (2) - (3) using MATLAB's 'dde23' function, and use numerical integration to assess the validity of the approximation.

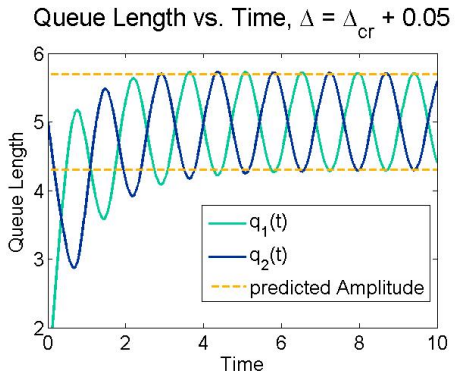

Figure 8: $\lambda = 10$, $\mu = 1$.

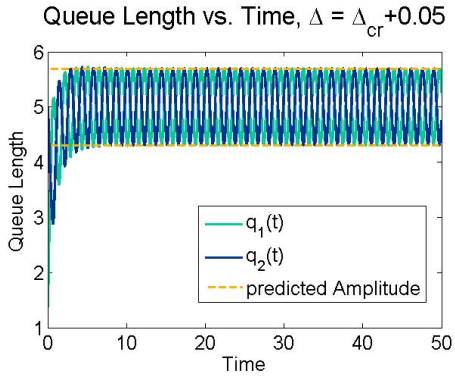

Figure 9: $\lambda = 10$, $\mu = 1$.

Lindstedt's method perturbs the system about Δ_{cr}, so the approximated amplitude must approach the true amplitude as $\Delta \to \Delta_{cr}$. This is consistent with our numerical results, and is evident from Figures 10 - 11. The two plots compare the numerically found amplitude with Lindstedt's amplitude while treating each as a function of delay for parameters $(\lambda, \mu) = (10, 1)$ for the ranges $\Delta \in [\Delta_{cr}, \Delta_{cr} + 0.2]$ and $\Delta \in [\Delta_{cr}, \Delta_{cr} + 1]$, respectively. In both cases the approximation is highly accurate when $\tau = \Delta - \Delta_{cr} \to 0$. However, Lindstedt's method cannot provide theoretical guarantees as the gap between Δ and Δ_{cr} increases, and as seen from Figures 10 - 11 the approximation loses accuracy.

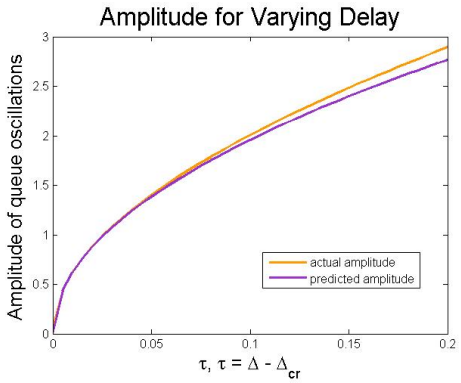

Figure 10: $\lambda = 10$, $\mu = 1$.

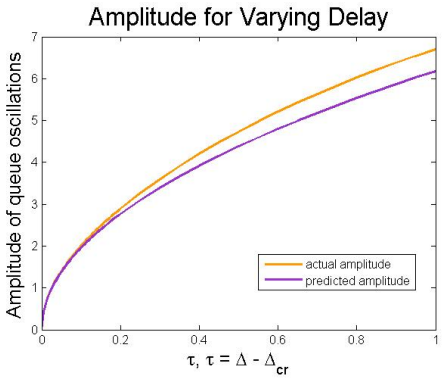

Figure 11: $\lambda = 10$, $\mu = 1$.

The method's performance is also affected by the choice of parameters λ and μ. Lindstedt's method works better for smaller λ, as shown by the surface plot in Figure 12 of the absolute error of Lindstedt's approximation across a range of λ and Δ. Based on the plot, the error of approximation monotonically increases with respect to both λ and Δ. While the absolute error in Figure 12 is constructed for $\mu = 1$, the same holds for other choices of μ. The performance of Lindstedt's method depends on μ in a similar fashion. The accuracy of the method improves when μ increases, and the error is monotone with respect to both μ and Δ. This trend is exemplified by the surface plot in Figure 13, which shows the absolute error of Lindstedt's method as a function of μ and Δ, for $\lambda = 10$.

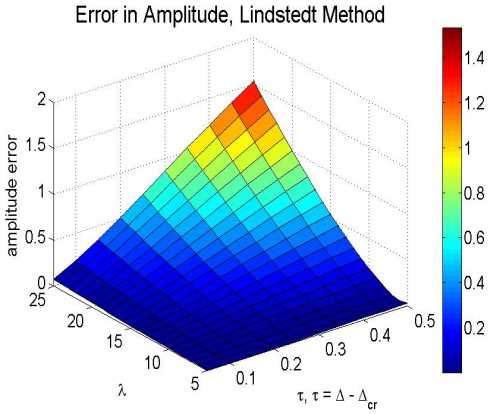

Figure 12: Absolute error, varying λ.

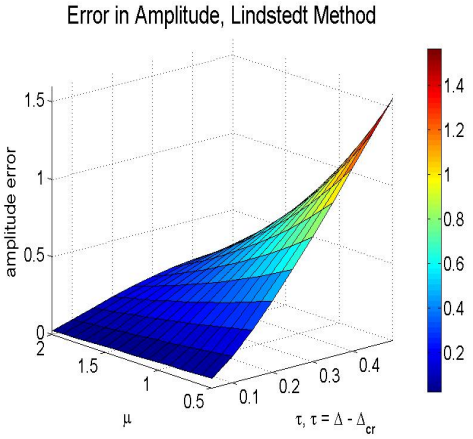

Figure 13: Absolute error, varying μ.

The observation that Lindstedt's method works differently for varying values of λ, μ, and Δ leads to two points. The first point is that even though the parameters depend on the physical circumstances and cannot be easily manipulated, it is beneficial to know when to expect a larger error in approximation. The second point is that the limitations of Lindstedt's method motivate us to develop a different numerical technique with the objective of decreasing the maximum error over a larger set of parameter values. Specifically, we would like to eliminate the peaks of error observed in Figures 12 - 13 when λ is large or μ is small, and therefore obtain a more accurate approximation of the amplitude. With this in mind, we introduce the *slope function method*.

2.6 The Slope Function Method

The theory of Hopf bifurcation together with numerical examples highlight that the amplitude is approximately proportional to the square root of the difference of the actual delay and the critical delay, i.e.

$$\text{Amplitude} \approx C(\lambda, \mu) \cdot \sqrt{\Delta - \Delta_{cr}}, \tag{30}$$

where the $C(\lambda, \mu)$ does not depend on Δ. We call $C(\lambda, \mu)$ the **slope function** as it characterizes the slope of the amplitude as a function of system's parameters. In this section, we propose a **statistical** way to fit the slope function, which turns out to approximate the amplitude in some cases better than Lindstedt's method.

The Slope Function Algorithm

1. For a fixed pair of parameters λ_1 and μ_1, we find the amplitude $A(\tau)$ via numerical integration for a finite number of points $\tau = \Delta - \Delta_{cr} := 0, d, 2d, ..., (K-1)d$, where $d > 0$ and $K \in \mathbb{N}$. Then $C(\lambda_1, \mu_1)$ is defined to be such coefficient C that for $A_p(\tau) = C\sqrt{\tau}$, the error $A_p(\tau) - A(\tau)$ is minimized in the least squares sense.

   The sum of squared errors for the K points of delay is given by the function $F(C) = \sum_{j=0}^{K-1} \left(C\sqrt{jd} - A(jd) \right)^2$, which by definition reaches

its minimum at C. Hence

$$\frac{dF(C)}{dC} = \sum_{j=0}^{K-1} 2\sqrt{jd}\Big(C\sqrt{jd} - A(jd)\Big) = 0. \tag{31}$$

The closed-form solution for C is found to be

$$C = \frac{\sum_{j=0}^{K-1} \sqrt{jd}A(jd)}{\sum_{j=0}^{K-1} jd}. \tag{32}$$

This gives us the value of the slope function at (λ_1, μ_1). To see how this approximation compares to the Lindstedt's method, consider Figures 14 and 15, which show the amplitude as a function of delay for $\lambda = 10$ and $\lambda = 20$, respectively. The slope function offers a relatively good approximation for the fixed λ and μ, and it is left to determine the function for the other values of λ and μ.

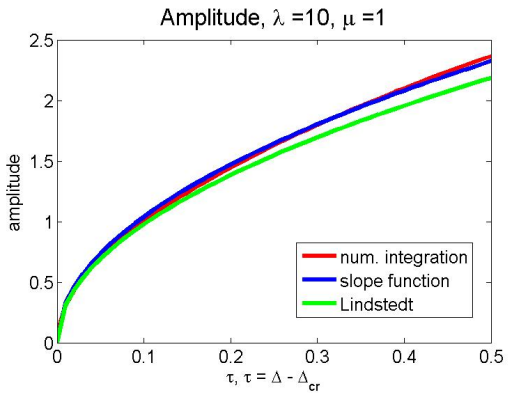

Figure 14: Amplitude approximation comparison.

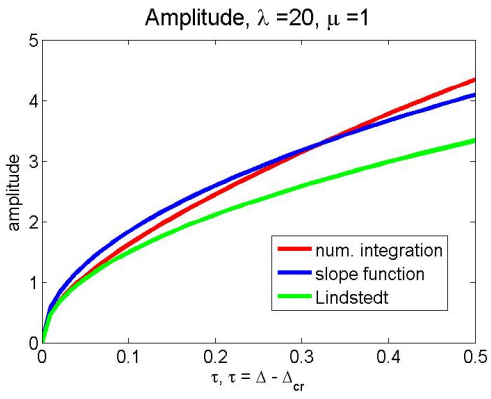

Figure 15: Amplitude approximation comparison.

2. We extrapolate to find the slope function at arbitrary λ and μ based on the function's values computed for a few points. We assume that $C(\lambda, \mu)$ is a separable function,

$$C(\lambda, \mu) = \Lambda(\lambda) M(\mu), \tag{33}$$

and then approximate the functions Λ and M by first degree polynomials

$$\Lambda(\lambda) \approx l_0 + l_1 \lambda, \qquad M(\mu) \approx m_0 + m_1 \mu, \qquad l_0, l_1, m_0, m_1 \in \mathbb{R}. \tag{34}$$

We cannot prove that C is a separable function because it depends on the unknown function A, the "true" amplitude, which is not necessarily separable. However, the separability assumption is a reasonable approximation based on our observations from numerical examples. Further, $C(\lambda, \mu)$ from Equation (32), as seen from experimental data, indeed is very close to a linear function of λ when μ is constant, and it is close to linear as a function of μ while λ is constant. This approximately linear behavior with respect to λ and μ is demonstrated in Figures 16 - 17, respectively, where the blue line in each plot represents the values of $C(\lambda, \mu)$ computed according to Equation (32).

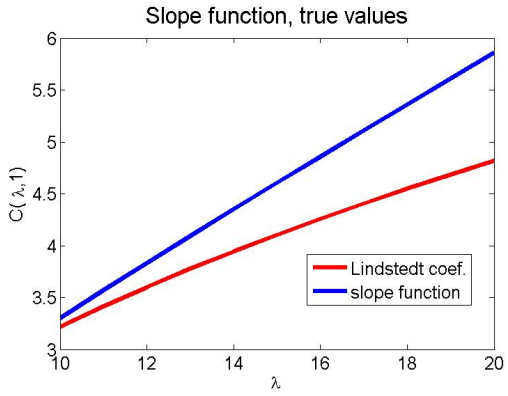

Figure 16: C is approximately linear in λ.

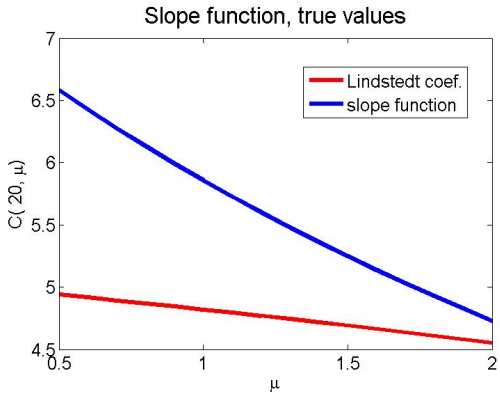

Figure 17: C is approximately linear in μ.

3. We reduce the number of coefficients by a change of variables $a_1 = l_1 m_1$, $l_0 = a_2 l_1$, and $m_0 = a_3 m_1$. Equation (33) then becomes

$$C(\lambda, \mu) = a_1(a_2 + \lambda)(a_3 + \mu). \tag{35}$$

Determining three unknown coefficients requires three data points $C(\lambda_1, \mu_1)$, $C(\lambda_2, \mu_1)$, and $C(\lambda_1, \mu_2)$ that are evaluated based on Equation (32) from Step 1 of the algorithm. Then Equation (35) allows us to solve for a_1, a_2, and a_3:

$$\frac{C(\lambda_1, \mu_1)}{C(\lambda_2, \mu_1)} = \frac{a_2 + \lambda_1}{a_2 + \lambda_2}, \quad \frac{C(\lambda_1, \mu_1)}{C(\lambda_1, \mu_2)} = \frac{a_3 + \mu_1}{a_3 + \mu_2}, \quad a_1 = \frac{C(\lambda_1, \mu_2)}{(a_2 + \lambda_1)(a_3 + \mu_2)}. \tag{36}$$

Therefore the coefficients of interest are

$$a_1 = \frac{C(\lambda_1, \mu_2)}{(a_2 + \lambda_1)(a_3 + \mu_2)}, \quad a_2 = \frac{\lambda_1 - x_1\lambda_2}{x_1 - 1}, \quad a_3 = \frac{\mu_1 - x_2\mu_2}{x_2 - 1}, \tag{37}$$

$$\text{where} \quad x_2 = \frac{C(\lambda_1, \mu_1)}{C(\lambda_1, \mu_2)}, \quad x_1 = \frac{C(\lambda_1, \mu_1)}{C(\lambda_2, \mu_1)}. \tag{38}$$

Remark. By this algorithm, the amplitude of the queues is estimated to be

$$\text{Amplitude} \approx a_1(a_2 + \lambda)(a_3 + \mu)\sqrt{\Delta - \Delta_{cr}}, \tag{39}$$

where the coefficients a_1, a_2, and a_3 are given by Equations (37) - (38). The specific values of these coefficients will slightly vary depending on the choice of parameters λ_1, λ_2, μ_1, and μ_2 because the linearity assumption of Equations (34) is only an approximation of the true behavior as shown in Figures 16 - 17. For optimal results one should choose the data points $C(\lambda_1, \mu_1)$, $C(\lambda_2, \mu_1)$, and $C(\lambda_1, \mu_2)$ around the range of λ and μ that one is interested in.

2.7 Numerical Results for the Slope Function Method

We will now numerically compare the performance of the slope function method to Lindstedt's method. Figures 18 and 19 show the absolute error of the amplitude for varying λ and Δ resulting from the slope function and Lindstedt's method, respectively. Note that overall the slope function

results in a smaller error for a wide range of λ and Δ, with a maximum error of 0.4 compared with a maximum error of 1.5 in Lindstedt's approximation. However, unlike Lindstedt's technique the slope function does not guarantee to be accurate when Δ approaches Δ_{cr}. Thus, it is advantageous to use the slope function for predicting the amplitude when the delay is sufficiently greater than the critical value, while Lindstedt's method is preferable when the delay is close to the threshold. A similar observation holds in the case when λ is constant and μ varies. Surface plots in Figures 20 and 21 show that the slope function has a maximum error of less than a third of the error seen in Lindstedt's method, being outperformed mainly when the delay approaches the critical value.

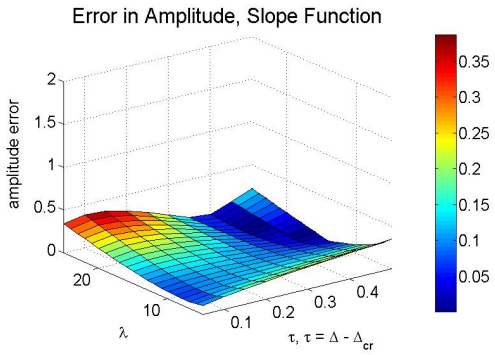

Figure 18: Absolute error from the slope function, with $\mu = 1$.

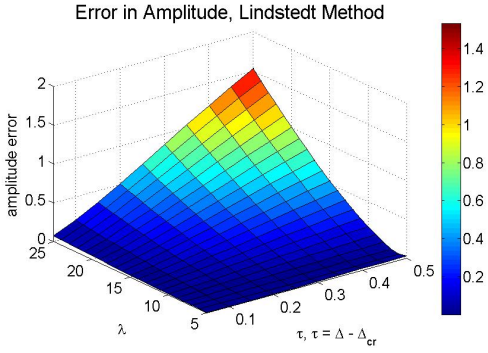

Figure 19: Absolute error from Lindtsedt's method, with $\mu = 1$.

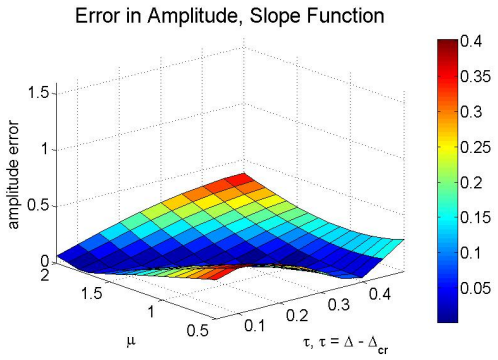

Figure 20: Absolute error from the slope function, with $\lambda = 20$.

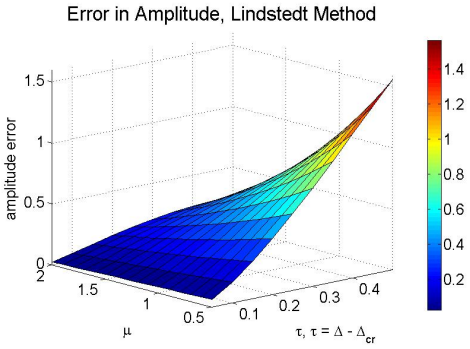

Figure 21: Absolute error from Lindtsedt's method, with $\lambda = 20$.

Since neither numerical method comes with an analytic expression for an error bound, a comparison of numeric results provides a valuable insight and gives intuition about the performance of the two methods. However, our insight from numerics is of course limited because we do not have guarantees that the numerical trends observed for this one queueing model can be extended to other models. In the next section, we introduce a different queueing model not only to study the model's behavior (which is interesting in of itself given the model's relevance to applications), but also to verify that the numerical trends of the method performance are consistent with the trends we observed so far.

3 Approximating Amplitude of Oscillations for Queueing System with Uniformly Distributed Delay

In physical settings, services often require the customers to travel in order to join a queue. This is true for bike sharing networks, hospital emergency rooms, and even restaurants - unless you are ordering a food delivery to your home! For such queueing systems, the assumption that each customer will take exactly the same time to travel is unrealistic. In fact, each customer may be departing from a slightly different location and therefore taking a slightly different time to arrive to the queue of their choice. To account for the variability in customers' travel time to the queue, we assume that an individual's travel time can be modeled as a random variable drawn from a uniform distribution.

The mathematical analysis of this modified queueing system presents new challenges. The queue lengths can no longer be modeled by delay-differential equations and instead have to be modeled by a system of functional differential equations. The Hopf bifurcation as observed in the constant delay model from Chapter 2 still occurs, but the Hopf curves are no longer monotonically decreasing functions of the arrival rate parameter λ. In fact, we can observe isola center bifurcations.

The Hopf bifurcation causes oscillations to form in queue lengths, and

we approximate the amplitude of these oscillations with the two numerical methods from Chapter 2, and compare their results to those of numerical integration.

3.1 Chapter Outline

The structure of Chapter 3 is identical to Chapter 2, where we present the mathematical model, describe the qualitative behavior of queue lengths, and later approximate the amplitude of oscillations via Lindstedt's method and the slope function method. Finally, we compare the performance of the two techniques, and conclude by highlighting the strengths and weaknesses of each method.

3.2 The Queueing Model

In this section, we present a queueing model similar to the constant delay model from Section 2, except here, customers have varying travel times to the queues. The individual's travel time is a random variable drawn from a uniform distribution. The fluid model accounts for the variability in customers' commute by uniformly distributing the delay over the last $\Delta > 0$ time units, or the **moving average**. Figure 4 still accurately represents the overall system: the customers appear at a rate λ, join one of the two queues with probabilities p_1 and p_2, and get service at a rate μ with an infinite number of servers. Customers join the queues according to the Multinomial Logit Model, giving higher preference to the queue with a smaller average

queue length

$$p_1 = \frac{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\Big)}{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\Big) + \exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\Big)} \tag{40}$$

$$p_2 = \frac{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\Big)}{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\Big) + \exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\Big)}. \tag{41}$$

Here p_i is the probability of i^{th} queue being joined, $q_i(t)$ is the i^{th} queue length, and the integral expressions account for the uniformly distributed delay due to customers' travel time.

Given these probabilities we can describe the queue lengths as

$$\dot{q}_1 = \lambda \cdot \frac{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\Big)}{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\Big) + \exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\Big)} - \mu q_1(t) \tag{42}$$

$$\dot{q}_2 = \lambda \cdot \frac{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\Big)}{\exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds\Big) + \exp\Big(-\frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds\Big)} - \mu q_2(t), \tag{43}$$

where $\Delta, \lambda, \mu > 0$. The equations are simplified by the notation m_i, which itself satisfies a delay differential equation:

$$m_i(t,\Delta) = \frac{1}{\Delta}\int_{t-\Delta}^{t} q_i(s)ds, \tag{44}$$

$$\dot{m}_i(t,\Delta) = \frac{1}{\Delta}\cdot\big(q_i(t) - q_i(t-\Delta)\big), \quad i \in \{1,2\}. \tag{45}$$

The functional differential equations (42) - (43) can now be expressed as a

system of DDEs

$$\dot{q}_1 = \lambda \cdot \frac{\exp(-m_1(t))}{\exp(-m_1(t)) + \exp(-m_2(t))} - \mu q_1(t) \tag{46}$$

$$\dot{q}_2 = \lambda \cdot \frac{\exp(-m_2(t))}{\exp(-m_1(t)) + \exp(-m_2(t))} - \mu q_2(t) \tag{47}$$

$$\dot{m}_1 = \frac{1}{\Delta} \cdot (q_1(t) - q_1(t-\Delta)) \tag{48}$$

$$\dot{m}_2 = \frac{1}{\Delta} \cdot (q_2(t) - q_2(t-\Delta)). \tag{49}$$

Since the functions m_i represent the averages of q_i, the initial conditions of m_i must reflect this. With $f_1(t)$ and $f_2(t)$ being continuous and non-negative functions on $t \in [-\Delta, 0]$, the initial conditions are

$$q_1(t) = f_1(t), \quad q_2(t) = f_2(t), \quad t \in [-\Delta, 0]; \tag{50}$$

$$m_1(0) = \frac{1}{\Delta}\int_{-\Delta}^{0} f_1(s)ds, \quad m_2(0) = \frac{1}{\Delta}\int_{-\Delta}^{0} f_2(s)ds. \tag{51}$$

3.3 Asymptotic Behavior

The behavior of the queues in Equations (46) - (49) depends on the delay parameter Δ, but the dependence itself is more nuanced than in the constant delay model. To provide a qualitative understanding of the behavior, we will begin by establishing the existence and uniqueness of the equilibrium.

Theorem 3.1. *The unique equilibrium of Equations* (46) - (49) *is given by*

$$q_1^*(t) = q_2^*(t) = m_1^*(t) = m_2^*(t) = \frac{\lambda}{2\mu}. \tag{52}$$

Proof. See the proof in the Appendix. □

The stability of the equilibrium comes from the characteristic equation that is determined by the linearized system of equation. Section 7.2 in the Appendix linearizes the system of Equations (46) - (49) and separate the variables, reducing the system from four unknown functions to two:

$$\overset{\bullet}{\tilde{v}}_2(t) = -\frac{\lambda}{2}\tilde{v}_4(t) - \mu\tilde{v}_2(t) \tag{53}$$

$$\overset{\bullet}{\tilde{v}}_4(t) = \frac{1}{\Delta}\Big(\tilde{v}_2(t) - \tilde{v}_2(t-\Delta)\Big). \tag{54}$$

To determine the characteristic equation, we need to first consider a special scenario with the trivial eigenvalue. Under the assumption that $\tilde{v}_2 = e^{\Lambda t}$ with $\Lambda = 0$, both functions must be constant, so for some $c_2, c_4 \in \mathbb{R}$, $\tilde{v}_2(t) = c_2$, $\tilde{v}_4(t) = c_4$. By Equation (54), the initial condition for $\tilde{v}_2(t)$ must be a constant function on $t \in [-\Delta, 0]$ so $\tilde{v}_2(t) = c_2$ for all $t \geq -\Delta$. The initial condition for $\tilde{v}_4$ then implies that $\tilde{v}_4(0) = c_4 = \int_{-\Delta}^{0} c_2 ds = \Delta c_2$. Therefore $c_4 = \Delta c_2$, but from Equation (53) we also find that $c_2 = -\frac{\lambda c_4}{2\mu}$. The only way both equalities can hold is if $c_2 = c_4 = 0$. Thus the trivial eigenvalue can only exist as a solution when the initial conditions are exactly zero, meaning that both queues must be of equal length $q_1(t) = q_2(t) = \frac{\lambda}{2\mu}$ for all $t \in [-\Delta, 0]$.

Assuming $\tilde{v}_2 = e^{\Lambda t}$ and $\Lambda \neq 0$, the characteristic equation is given by

$$\Phi(\Lambda, \Delta) = \Lambda + \mu + \frac{\lambda}{2\Delta\Lambda} - \frac{\lambda}{2\Delta\Lambda} \cdot e^{-\Lambda\Delta} = 0. \tag{55}$$

The equilibrium is stable as long as all eigenvalues Λ have negative real parts. Proposition 7.2 in the Appendix shows that any real eigenvalue must be negative. However, since $\Delta > 0$ there are also infinitely many pairs of complex eigenvalues. The following proposition shows that, regardless of the parameters λ and μ, all complex eigenvalues have negative real parts when the delay is sufficiently small.

Proposition 3.2. *Let $\lambda, \mu, \Delta > 0$. There exists $\Delta^* > 0$ such that for any $\Delta < \Delta^*$, all complex eigenvalues of the characteristic equation* (55) *have negative real parts.*

Proof. Let $\Lambda = a + ib$ be a solution of Equation (55). Then a and b must satisfy

$$\cos(b\Delta) = \frac{e^{a\Delta}}{\lambda}(2a^2\Delta - 2b^2\Delta + \lambda + 2a\mu\Delta) \tag{56}$$

$$\sin(b\Delta) = -\frac{e^{a\Delta}}{\lambda} \cdot 2b\Delta(2a + \mu). \tag{57}$$

If b satisfies these equations, then $-b$ is a solution too. Hence without loss of generality we will assume that $b > 0$. Summing the squares of the two equations, we get

$$e^{-2a\Delta}\lambda^2 = \big(2a^2\Delta - 2b^2\Delta + \lambda + 2a\mu\Delta\big)^2 + \big(2b\Delta(2a + \mu)\big)^2, \tag{58}$$

from which b can be expressed as a continuous function of a and Δ, namely $b(a, \Delta)$. If $a = 0$ then $b(0, \Delta) = \sqrt{\frac{\lambda}{\Delta} - \mu^2}$, and when plugged into Equation

(57) we get

$$\sin(b(0,\Delta)\Delta) = -\frac{2\mu}{\lambda} \cdot b(0,\Delta)\Delta \tag{59}$$

$$\sin(x(0,\Delta)) = -\frac{2\mu}{\lambda} \cdot x(0,\Delta) \tag{60}$$

$$x(a,\Delta) = b(a,\Delta)\Delta, \quad x(0,\Delta) = \Delta\sqrt{\frac{\lambda}{\Delta} - \mu^2}. \tag{61}$$

The function x will be helpful in the proof. Note that x is a continuous function of b and therefore of a. Let us define $\Delta^* > 0$ as

$$\Delta^* = \begin{cases} \frac{\lambda}{2\mu^2}, & \frac{\lambda}{2\mu} \leq \pi \\ \frac{\lambda - \sqrt{\lambda^2 - 4\mu^2\pi^2}}{2\mu^2}, & \text{otherwise.} \end{cases} \tag{62}$$

This choice of Δ^* guarantees that for all $\Delta < \Delta^*$, the functions $b(0,\Delta)$ and $x(0,\Delta)$ are real. Further, Δ^* ensures that $0 < x(0,\Delta) < \min(\pi, \frac{\lambda}{2\mu})$ for all $\Delta < \Delta^*$, which can be checked from Equation (61). The condition $0 < x(0,\Delta) < \pi$ implies that

$$\sin(x(0,\Delta)) > 0 > -\frac{2\mu}{\lambda} \cdot x(0,\Delta). \tag{63}$$

However, for any $a \geq 0$, Equation (57) gives the inequality

$$\sin(x(a,\Delta)) = -\frac{e^{a\Delta}}{\lambda} \cdot 2x(a,\Delta)(2a+\mu) \leq -\frac{2\mu}{\lambda} \cdot x(a,\Delta), \tag{64}$$

therefore when $a = 0$ the inequality remains

$$\sin(x(0,\Delta)) \leq -\frac{2\mu}{\lambda} \cdot x(0,\Delta), \tag{65}$$

which is in contradiction with Equation (63). Hence a must be negative to satisfy the characteristic equation for $\Delta < \Delta^*$. □

The equilibrium becomes unstable when a pair of complex eigenvalues crosses the imaginary axis. If for some $\Delta = \Delta_{cr}$ there are purely imaginary eigenvalues, $\Lambda = \pm i\omega_{cr}$, $\omega_{cr} > 0$, then the characteristic equation gives the equalities

$$\sin(\omega_{cr}\Delta_{cr}) = -\frac{2\Delta_{cr}\mu\omega_{cr}}{\lambda}, \quad \cos(\omega_{cr}\Delta_{cr}) = 1 - \frac{2\Delta_{cr}\omega_{cr}^2}{\lambda}. \tag{66}$$

From the trigonometric identity $\sin^2(\omega_{cr}\Delta_{cr}) + \cos^2(\omega_{cr}\Delta_{cr}) = 1$, ω_{cr} can be found

$$\omega_{cr} = \sqrt{\frac{\lambda}{\Delta_{cr}} - \mu^2}. \tag{67}$$

Since ω_{cr} must be real and nonzero, the condition $\Delta_{cr} < \frac{\lambda}{2\mu}$ must hold. When ω_{cr} is substituted into Equation (66), we find that Δ_{cr} must satisfy the equation

$$\sin\left(\Delta_{cr} \cdot \sqrt{\frac{\lambda}{\Delta_{cr}} - \mu^2}\right) + \frac{2\mu\Delta_{cr}}{\lambda} \cdot \sqrt{\frac{\lambda}{\Delta_{cr}} - \mu^2} = 0. \tag{68}$$

We are now ready to formulate the conditions that determine the stability of the equilibrium.

Theorem 3.3. *If the Equation* (68) *has no positive roots* Δ_{cr} *then the equilibrium of Equations* (46) - (49) *is stable for all* $\Delta > 0$. *If there exists* $\Delta_{cr} > 0$ *satisfying Equation* (68) *then the equilibrium is stable when* Δ *is less than the smallest positive root* Δ_{cr} *or greater than the largest root* Δ_{cr}. *Further, the largest root* Δ_{cr} *is less than* $\frac{\lambda}{\mu^2}$.

Proof. See the proof in the Appendix. □

If and when Δ exceeds the smallest positive root Δ_{cr} of Equation (68), the equilibrium becomes unstable and a stable limit cycle emerges. Figures 22 and 23 show the transition. The change of behavior is due to a Hopf bifurcation, as shown in the next theorem. Further, since there can be multiple roots Δ_{cr} to Equation (68) for fixed parameters λ and μ, multiple Hopf bifurcations may occur.

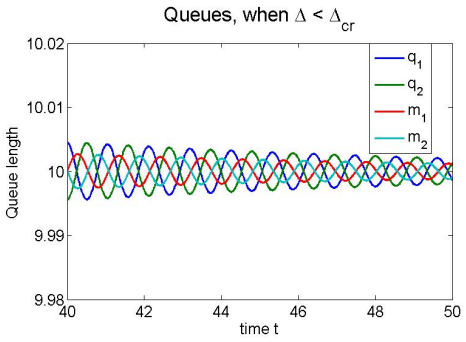

Figure 22: Before bifurcation. $\lambda = 10; \mu = 1$.

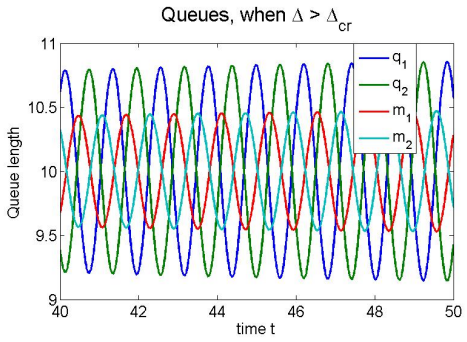

Figure 23: After bifurcation. $\lambda = 10; \mu = 1$.

Theorem 3.4. *If Δ_{cr} satisfies Equation* (68) *and $\Delta_{cr} \neq \frac{\lambda-2\mu}{2\mu^2}$, then the queues from Equations* (46) - (49) *undergo a Hopf bifurcation at Δ_{cr}.*

Proof. For each Δ_{cr} satisfying Equation (68), the characteristic equation (55) has two simple roots $\Lambda = \pm i\omega_{cr}$. Further, through implicit differentiation of Equation (55), it can be shown that $\text{Re}[\Lambda'(\Delta_{cr})] \neq 0$:

$$\text{Re } \Lambda'(\Delta_{cr}) = \frac{2\omega_{cr}^2(\lambda - 2\mu - 2\mu^2\Delta_{cr})}{4\omega_{cr}^2\Delta_{cr}(3 + 2\Delta_{cr}\mu) + \lambda(4 + \Delta_{cr}\lambda + 4\Delta_{cr}\mu)}. \tag{69}$$

The denominator of $\text{Re}[\Lambda'(\Delta_{cr})]$ is positive, and the assumption $\Delta_{cr} \neq \frac{\lambda-2\mu}{2\mu^2}$ guarantees the numerator to be nonzero. Further, all other eigenvalues Λ^* are complex with a nonzero real part, so $\Lambda^* \neq m\Lambda$. Therefore, a Hopf bifurcation occurs at Δ_{cr}. □

As was suggested by Figure 23, the limit cycle is stable. In fact, the following theorem shows that any Hopf bifurcation in our queueing system is supercritical.

Theorem 3.5. *Any Hopf bifurcation from Theorem 3.4 is supercritical.*

Proof. We will use the method of slow flow to determine whether the limit cycle is stable. The third order expansion of Equations (46) - (47) can be uncoupled, and the resulting equations of interest are given by Section 7.2 in

the Appendix:

$$\dot{\tilde{v}}_2 = \lambda\Big(- \frac{\tilde{v}_4(t)}{2} + \frac{\tilde{v}_4(t)^3}{24}\Big) - \mu\tilde{v}_2(t) \tag{70}$$

$$\dot{\tilde{v}}_4 = \frac{1}{\Delta}\Big(\tilde{v}_2(t) - \tilde{v}_2(t-\Delta)\Big). \tag{71}$$

The two variables are scaled by $\sqrt{\epsilon}$

$$\tilde{v}_2(t) = \sqrt{\epsilon}v(t), \qquad \tilde{v}_4(t) = \sqrt{\epsilon}u(t), \tag{72}$$

the delay and the frequency are expanded close to their critical values, and two time scales are introduced:

$$\Delta = \Delta_{cr} + \epsilon\alpha, \quad \omega = \omega_{cr} + \epsilon\beta, \quad \xi = \omega t, \quad \eta = \epsilon t. \tag{73}$$

The functions $v(t)$ and $u(t)$ are also expanded

$$v(\xi,\eta) = v_0(\xi,\eta) + \epsilon v_1(\xi,\eta), \qquad u(\xi,\eta) = u_0(\xi,\eta) + \epsilon u_1(\xi,\eta). \tag{74}$$

When the suggested transformations are made to the equations for $\dot{v}(t)$ and $\dot{u}(t)$, we can separate the resulting equations by collecting all the terms with the like orders of ϵ. The equations for the zeroth order terms are satisfied with a solution of the form

$$v_0(\xi,\eta) = A(\eta)\cos(\xi) + B(\eta)\sin(\xi), \tag{75}$$

which allows us to find the form of $u_0(\xi, \eta)$:

$$u_0(\xi, \eta) = -\frac{2(A(\eta) + B(\eta)\omega_{cr})}{\lambda}\cos(\xi) - \frac{2(B(\eta) - A(\eta)\omega_{cr})}{\lambda}\sin(\xi). \quad (76)$$

The terms involving the first order of ϵ comprise of (i) the differential operator acting on x_1, (ii) the non-resonant terms $\cos(3\xi)$ and $\sin(3\xi)$, and (iii) the resonant terms involving $\cos(\xi)$ and $\sin(\xi)$. For no secular terms, the coefficients of $\cos(\xi)$ and $\sin(\xi)$ must vanish, giving a slow flow on $A(\eta)$ and $B(\eta)$. By introducing the polar coordinates

$$A = R\cos(\Theta), \qquad B = R\sin(\Theta) \quad (77)$$

we find equation for the radial component $\frac{d}{d\eta}R(\eta)$

$$\frac{dR}{d\eta} = \frac{R(\lambda - \Delta_{cr}\mu^2)(R^2(\lambda + 2\mu) - 4\alpha\lambda(\lambda - 2\mu - 2\Delta_{cr}\mu^2))}{2\Delta_{cr}\lambda(-\Delta_{cr}\lambda^2 + 4\Delta_{cr}\mu^2(3 + 2\Delta_{cr}\mu) - 4\lambda(4 + 3\Delta_{cr}\mu))}. \quad (78)$$

Assuming $R \geq 0$, the equilibrium points are

$$R_0 = 0, \quad R_1 = \sqrt{\frac{4\alpha\lambda(\lambda - 2\mu - 2\Delta_{cr}\mu^2)}{\lambda + 2\mu}}. \quad (79)$$

From Theorem 3.4, we know $\Delta_{cr} \neq \frac{\lambda - 2\mu}{2\mu^2}$, so R_1 and R_0 are always two distinct points. When $\Delta_{cr} < \frac{\lambda - 2\mu}{2\mu^2}$ then in order for R_1 to be real, α must be positive. On the other hand, if $\Delta_{cr} > \frac{\lambda - 2\mu}{2\mu^2}$ then α must be negative for R_1 to be real. In both cases, the assumption $\frac{\lambda}{\Delta_{cr}} - \mu^2 > 0$ that arose from the

frequency ω_{cr} being positive, guarantees that $\frac{dR}{d\eta}$ is positive on the interval $R \in (0, R_1)$ and negative when $R > R_1$. Therefore the Hopf bifurcation is supercritical. $\square$

To summarize, for any fixed parameters λ and μ the queues converge to a stable equilibrium when the delay is sufficiently small. However, as the delay increases up to $\Delta = \frac{\lambda - 2\mu}{2\mu^2}$, finitely many pairs of complex eigenvalues may cross to the positive real side of the imaginary axis of the complex plane. Every pair of eigenvalues reaching the imaginary axis is indicated on Figure 24 by a Hopf curve. Note that the dashed orange line $\Delta = \frac{\lambda - 2\mu}{2\mu^2}$ from Figure 24 passes through the minimum of each Hopf curve, where each minimum represents a pair of eigenvalues that reaches the imaginary axis at $\Delta = \frac{\lambda - 2\mu}{2\mu^2}$ and then returns back to the negative real side of complex plane without crossing the imaginary axis.

Once the delay exceeds $\frac{\lambda - 2\mu}{2\mu^2}$ and the parameters are in the region to the right of the dashed orange line from Figure 24, every pair of eigenvalues with positive real parts will inevitably cross back the imaginary axis in the negative real direction. In fact, all eigenvalues will obtain negative real parts before the delay reaches a. This is guaranteed by the condition $0 \neq \omega_{cr} \in \mathbb{R}$ together with Proposition 7.3 in the Appendix. The condition $\Delta = \frac{\lambda}{\mu^2}$ is indicated on Figure 24 by the non-Hopf curve, and it is clear that the Hopf curves cannot cross the non-Hopf curve.

The equilibrium is stable whenever λ is below the Hopf 1 curve from Figure 24. To quantitatively describe the behavior of the queues after Hopf

1 curve is crossed, we will approximate the amplitude of the queue oscillations via Lindstedt's method.

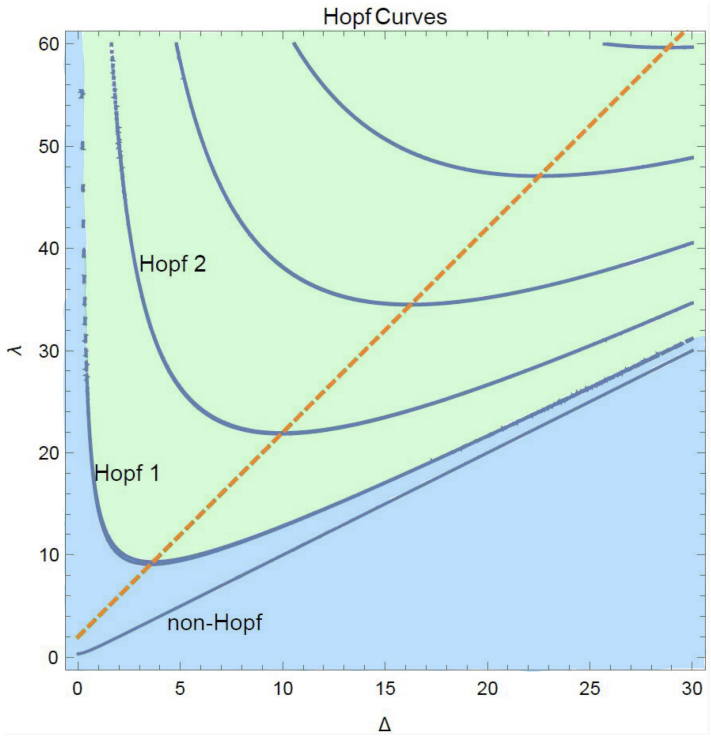

Figure 24: The Hopf curves for $\mu = 1$; green area - limit cycles; blue area - stable equilibrium; dashed orange line $\rightarrow \lambda = 2\mu^2\Delta + 2\mu$; the non-Hopf curve $\rightarrow \Delta = \frac{\lambda}{\mu^2}$.

3.4 Approximating the Amplitude of Oscillations

We apply Lindstedt's method according to the steps shown in Subsection 2.4. However, instead of working with one unknown function we are now working with two.

1. We start with the variables that represent third order polynomial ex-

pansion of q_1, q_2, m_1, and m_1 about the equilibrium. These four variables can be reduced to two by a change of variables. The details are in the Appendix 7.2. The functions of interest become $\tilde{v}_2$ and $\tilde{v}_4$ from Equations (70) - (70). We stretch the time and scale $\tilde{v}_2$ and $\tilde{v}_4$ by $\sqrt{\epsilon}$:

$$\tau = \omega t, \quad \tilde{v}_2 = \sqrt{\epsilon} v(t), \quad \tilde{v}_4 = \sqrt{\epsilon} u(t). \tag{80}$$

This ensures that the cubic terms will have one higher order of ϵ than linear terms,

$$\begin{aligned} \omega \dot{v}(\tau) &= \lambda\Big(-\frac{u(\tau)}{2} + \frac{\epsilon u(\tau)^3}{24}\Big) - \mu v(\tau) && (81) \\ \omega \dot{u}(\tau) &= \frac{1}{\Delta}\Big(v(\tau) - v(\tau - \omega\Delta)\Big). && (82) \end{aligned}$$

2. We approximate v and u by using asymptotic expansions in ϵ:

$$v(t) = v_0(t) + \epsilon v_1(t) + ..., \qquad u(t) = u_0(t) + \epsilon u_1(t) + ..., \tag{83}$$

$$\Delta = \Delta_0 + \epsilon\Delta_1 + ..., \qquad \omega = \omega_0 + \epsilon\omega_1 + ... \tag{84}$$

3. We separate each of the resulting equations by collecting all the terms of the like powers of ϵ. The terms of order ϵ^0 yield the following equalities

$$\begin{aligned} 0 &= \frac{1}{2}\lambda m_0(\tau) + \mu v_0(\tau) + \omega_0 \dot{v}_0(\tau) && (85) \\ 0 &= -v_0(\tau) + v_0(\tau - \Delta_0\omega_0) + \Delta_0\omega_0 \dot{m}_0(\tau), && (86) \end{aligned}$$

and the terms of order ϵ^1 yield the following equalities

$$\begin{aligned} 0 &= -\frac{1}{24}\lambda m_0(\tau)^3 + \frac{1}{2}\lambda m_1(\tau) + \mu v_1(\tau) + \omega_1 \dot{v}_0(\tau) + \omega_0 \dot{v}_1(\tau) \quad (87) \\ 0 &= \Delta_1\Big(v_0(\tau) - v_0(\tau - \Delta_0\omega_0)\Big) + \Delta_0^2\omega_1 \dot{m}_0(\tau) + \Delta_0^2\omega_0 \dot{m}_1(\tau) \quad (88) \\ &- \Delta_0\Big(v_1(\tau) - v_1(\tau - \Delta_0\omega_0) + (\Delta_1\omega_0 + \Delta_0\omega_1)\dot{v}_0(\tau - \Delta_0\omega_0)\Big). \end{aligned}$$

The function m_0 can be expressed through v_0 by Equation (85), and m_1 can be expressed through v_0 and v_1 from Equation (87). It can be verified that $v_0(\tau) = A_v \sin(\tau)$ satisfies Equations (85) - (86). Further, the homogeneous part of solution for v_1 is satisfied by $v_1^H(\tau) = a\sin(\tau) + b\cos(\tau)$. Therefore, to avoid secular terms $\sin(\tau)$ and $\cos(\tau)$, the coefficients of $\sin(\tau)$ and $\cos(\tau)$ from Equation (88) must vanish. This condition gives two equations for two unknowns, w_1 and A_v.

4. After some algebra, the amplitude A_v is found as a function of delay:

$$A_v(\Delta) = \sqrt{\Delta - \Delta_{cr}} \cdot \sqrt{\frac{4\lambda^2(-\lambda - 2\mu + 2\Delta_{cr}\omega_{cr}^2)}{\Delta_{cr}\big(\mu^2 + \omega_{cr}^2\big)\big(-\lambda + 2(\mu + \Delta_{cr}\mu^2 + \Delta_{cr}\omega_{cr}^2)\big)}}. \quad (89)$$

Amplitude of the Queues

The function A_v approximates the amplitude of oscillations for $v(t)$ from Equation (81). A change of variables reveals the amplitude of q_1 and q_2, showing that the steady state of queues is given up to a phase shift by

$$q_1(t) \to \frac{\lambda}{2\mu} + \frac{1}{2}A_v \sin(\omega t), \qquad q_2(t) \to \frac{\lambda}{2\mu} - \frac{1}{2}A_v \sin(\omega t), \quad (90)$$

where the amplitude is $\frac{1}{2}A_v$ and ω is the frequency of oscillations. Figures 25 - 26 use the predicted amplitude to bound the oscillations of queues near the bifurcation point, providing some validation to Lindstedt's method as well as our calculations.

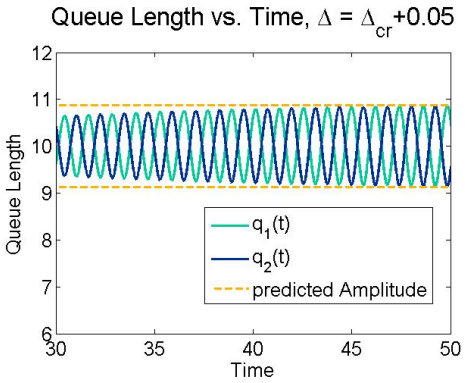

Figure 25: $\lambda = 20, \mu = 1$.

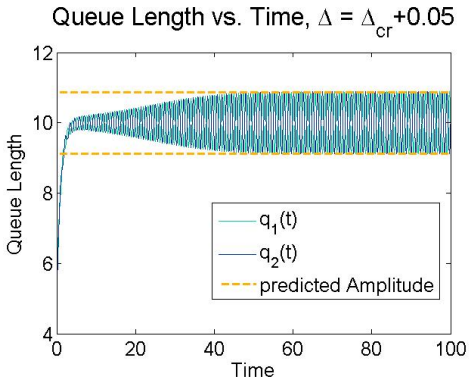

Figure 26: $\lambda = 20, \mu = 1$.

3.5 Numerical Results

We compare the approximations of amplitude from Lindstedt's method and the slope function method to the true behavior of the queueing system. The slope function is provided by the algorithm in Section 2.6 and Equation (39), so no additional work is needed. Also, we consider the queue lengths to be determined with sufficient accuracy by numerical integration of Equations (46) - (49) using MATLAB's 'dde23' function, so we will test our approximations against the numerical integration results.

Our key finding is that *the trends of the method performance are consistent* with those that were observed for the constant delay model in Subsection

2.5, both for Lindstedt's method and the slope function method. Hence, we avoid repeating the analysis of Subsection 2.5, and instead provide relevant figures with a summary of the key differences between the two methods.

- Lindstedt's method tends to be more accurate than the slope function method when $\Delta \to \Delta_{cr}$. For example, see Figure 28, where the amplitude is shown as a function of delay.

- Lindstedt's method loses accuracy when the delay increases, and it is outperformed by the slope function method for larger delay. See Figures 27 - 32.

- The error of Lindstedt's approximation is monotonic in λ, Δ, and μ. Hence, over the parameter space the error function has predictable and significant peaks around large λ and Δ and around small μ. See Figures 30 and 32.

- The error of the slope function method is relatively evenly distributed over the parameter space, and therefore there are no significant peaks in error. See Figures 29 and 31.

- The maximum error for the slope function method over a neighborhood of parameters is 3 - 4 times smaller than it is for Lindstedt's method; the maximum error is three times smaller for the constant delay model, and 4 times smaller for the model with the distributed delay. See Figures 29 - 32.

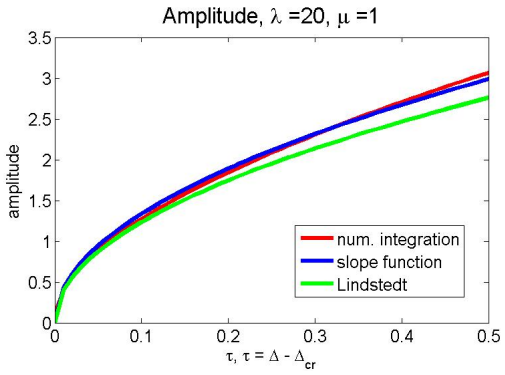

Figure 27: Comparison of approximations.

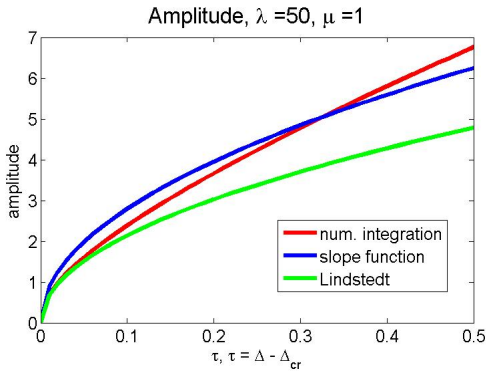

Figure 28: Comparison of approximations.

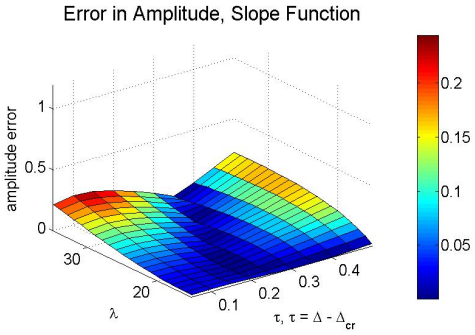

Figure 29: Absolute error, $\mu = 1$.

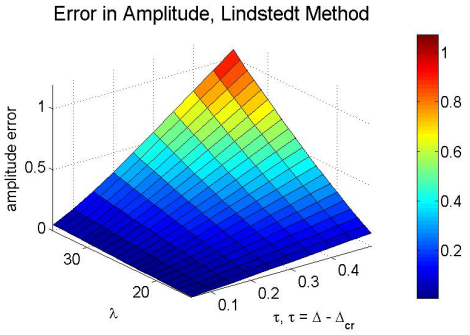

Figure 30: Absolute error, $\mu = 1$.

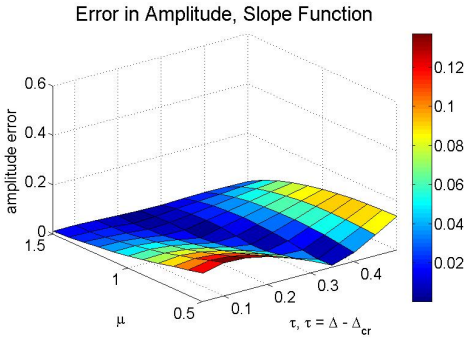

Figure 31: Absolute error, $\lambda = 20$.

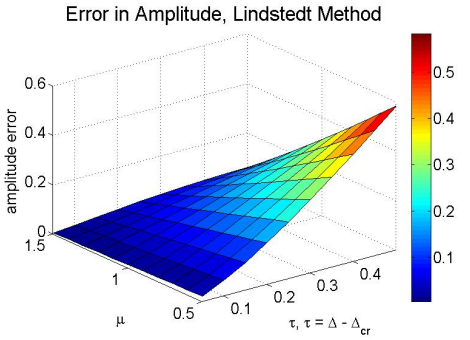

Figure 32: Absolute error, $\lambda = 20$.

3.6 Conclusion

So far we analyzed two queueing models that incorporate customer choice and delayed queue length information. The first model assumes a constant delay and while the second one uses uniformly distributed delay. We analyze the qualitative behavior of these queueing models and show the occurrence of supercritical Hopf bifurcations. Using Lindstedt's method, we construct an analytic approximation for the amplitude of oscillations that the queueing system exhibits after a Hopf bifurcation. Lindstedt's method works well where the delay is close the critical delay value, but it becomes less accurate for larger values of delay. We address this by proposing a new numerical technique, the slope function method, that estimates the slope of the amplitude as a function of the system's parameters.

The slope function method is conceptually intuitive and elementary in implementation. It can be used in a wide variety of models where a Hopf bifurcation is observed. Unlike the perturbations method, the slope function does not require complicated analytical work and can be implemented without a substantial mathematical background. Limit cycles are known to occur in models studied by social scientists and biologists, for which the slope function method can provide an easy way to numerically approximate the amplitude of oscillations. Although we give no theoretical guarantees on the method's performance, our paper demonstrates on two different models that the slope function method maintains a low error across a much wider range of the parameters than does Lindstedt's method. For our models, the

maximum error in approximation is $3-4$ times smaller over a large neighborhood of parameters than the maximum error from Lindstedt's method.

4 Limiting the Oscillations in Queues Through a Novel Type of Delay Announcement

Oscillations in queue lengths are generally undesirable both for the customers and the service managers. They cause a number of issues like longer waiting times and uneven workload for the service providers. The threshold at which the queues become unstable can be affected by the type of information revealed to the customers. This chapter analyzes what information the service managers should provide 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 getting 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 propose a new type of delay announcement that incorporates both pieces of available information, presenting the customers with a weighted sum of the queue length and the rate with which the queue length is changing (velocity). This delay announcement is considered in the context of the constant delay model studied in Chapter 2, except the model is generalized to include N queues.

The queueing system can undergo a Hopf bifurcation if the delay due to the customer travel time is sufficiently large. The exact point where the Hopf bifurcation occurs depends on the weight coefficient of velocity information.

In this chapter we show how to choose the weight coefficient 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.

The delay for which the queues remain stable can be increased, but only up to a point. We provide upper and lower bounds on the maximum delay that can be attained by incorporating the velocity information. When the oscillations in queues cannot be prevented, we use 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 harm to the queueing system by destabilizing it. We specify the threshold for the weight coefficient where the adverse effects take place.

4.1 Chapter Outline

In Section 4.2 we present a mathematical model for N queues and describe the qualitative behavior of the queueing system. 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 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 thresh-

old Δ_{cr}, the equilibrium becomes unstable. Section 4.4 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 Δ_{cr}.

Section 4.5 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.

4.2 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. Our model assumes that the queueing dynamics for the departure process are identical to that of an infinite-server queue with service rate $\mu > 0$. Figure 33 shows the queueing system for N queues. The queue length for the i^{th} queue is given by

$$\dot{q}_i(t) = \lambda p_i(q_1, \ldots, q_N, \dot{q}_1, \ldots, \dot{q}_N, \Delta) - \mu q_i(t), \quad \forall i \in \{1, ..., N\}, \qquad (91)$$

where the function p_i represents the probability that a customer chooses the i^{th} queue. Instead of being told the queue length, customers are given a weighed sum of the queue length and its velocity,

$$\text{Information about i}^{th} \text{ queue} = q_i(t-\Delta) + \delta \dot{q}_i(t-\Delta). \tag{92}$$

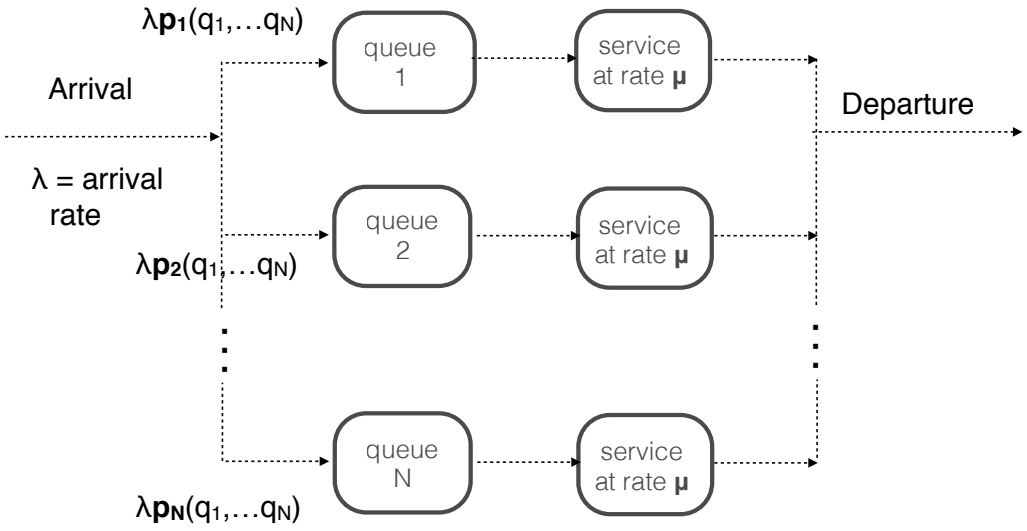

Figure 33: Customers going through a N-queue service system.

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

$$q(t) = q(t-\Delta) + \sum_{j=1}^{\infty} \frac{\Delta^j}{j!} q^{(j)}(t-\Delta), \tag{93}$$

where we define $q^{(j)}(t-\Delta)$ as the j^{th} time derivative of the queue length

process at the time $t - \Delta$. 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.

The probability of a customer choosing the i^{th} queue is given by the Multinomial Logit Model (MNL), which takes in as an input our delay announcement information; a customer chooses the i^{th} queue with probability

$$p_i(q_1, \ldots, q_N, \dot{q}_1, \ldots, \dot{q}_N, \Delta) = \frac{\exp\Big(-\theta\big(q_i(t-\Delta) + \delta\dot{q}_i(t-\Delta)\big)\Big)}{\sum_{j=1}^{N} \exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_j(t-\Delta)\big)\Big)}, \quad (94)$$

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 travelling to the service. The coefficient θ reflects the strength of the customers' preference for the shorter queue; when θ approaches zero, customers are ambivalent about the queue length and are equally likely to pick any queue, while θ approaching infinity means that customers will always go for the shortest queue. Obviously, when $\theta = 1$ and $\delta = 0$, we revert back to the old model from Chapter 2.

Incorporating the probabilities p_i into the queueing system provides a system of **neutral delay differential equations** (NDDEs) for the queue

lengths

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\Big(-\theta\big(q_i(t-\Delta) + \delta\dot{q}_i(t-\Delta)\big)\Big)}{\sum_{j=1}^{N} \exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_i(t-\Delta)\big)\Big)} - \mu q_i(t), \tag{95}$$

with the initial conditions specified by non-negative continuous functions f_i

$$q_i(t) = f_i(t), \quad \dot{q}_i(t) = \dot{f}_i(t), \quad t \in [-\Delta, 0]. \tag{96}$$

4.3 Conditions for Stability and Hopf Bifurcations

We begin by establishing the existence and uniqueness of solution to the initial value problem (95) - (96). We note that there exists an extensive analysis of functional differential equations, see for example [18, 7, 30]. The existence and uniqueness of solution for our specific model directly follows from Driver [11], as stated in the result below.

Theorem 4.1. *Let $f_i(t)$ from Equation* (96) *be absolutely continuous on $t \in [-\Delta, 0]$, and $\dot{f}_i(t)$ be bounded for almost all $t \in [-\Delta, 0]$ for every $1 \leq i \leq N$. Then there exists a solution $q_i, \ldots, q_N$ for all $t > 0$ that satisfies Equations* (95) - (96)*. Further, the solution is unique.*

Proof. The existence of the solution is given by Theorem 1 of Driver [11]. The uniqueness of the solution follows from Theorem 2 of Driver [11], but we first need to ensure that the conditions of Theorem 2 are fulfilled. The

theorem requires that the function

$$\lambda \cdot \frac{\exp\Big(-\theta\big(q_i(t-\Delta)+\delta\dot{q}_i(t-\Delta)\big)\Big)}{\sum_{j=1}^{N}\exp\Big(-\theta\big(q_j(t-\Delta)+\delta\dot{q}_i(t-\Delta)\big)\Big)} - \mu q_i(t), \quad \forall i \in \{1,...,N\} \tag{97}$$

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 (95) - (96) is guaranteed to be unique. □

In addition to a unique solution, we can show that this queueing model has a unique equilibrium state.

Theorem 4.2. *The unique equilibrium of $q_i(t)$ from Equation* (95) *is given by*

$$\lim_{t\to\infty} q_i(t) = q_i^* = \frac{\lambda}{N\mu}, \quad 1 \le i \le N. \tag{98}$$

Proof. See the Appendix for the proof. □

The stability of the equilibrium can be determined by the stability of the linearized system of equations [18, 40]. Hence, we proceed by linearizing q_i about the equilibrium, and finding the characteristic equation.

Proposition 4.3. *The characteristic equation of* (95) *is given by*

$$\Phi(R,\Delta) \;=\; -R - \frac{\lambda\theta}{N}\Big(e^{-R\Delta} + \delta R e^{-R\Delta}\Big) - \mu = 0. \tag{99}$$

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

$$u_i(t) = q_i(t) - q_i^* = q_i(t) - \frac{\lambda}{N\mu}. \tag{100}$$

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

$$\dot{u}_i(t) \approx -\frac{\lambda\theta}{N}\Big(u_i(t-\Delta) + \delta u_i'(t-\Delta)\Big) \tag{101}$$

$$+\frac{\lambda\theta}{N^2}\sum_{j=1}^{N}\Big(u_j(t-\Delta) + \delta u_j'(t-\Delta)\Big) - \mu u_i(t) \tag{102}$$

In the vector form, we have

$$\dot{\mathbf{u}}(t) = -\frac{\lambda\theta}{N}\cdot\big(\mathbf{u}(t-\Delta) + \delta\mathbf{u}'(t-\Delta)\big) \tag{103}$$

$$+\frac{\lambda\theta}{N^2}A\big(\mathbf{u}(t-\Delta) + \delta\mathbf{u}'(t-\Delta)\big) - \mu\mathbf{u}(t), \tag{104}$$

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:

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & -1 & \dots & -1 \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & & & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & 0 \end{bmatrix}$$

$$A = VDV^{-1}, \quad \text{where} \quad A, V, D \in \mathbb{R}^{N \times N}. \tag{105}$$

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

$$\mathbf{u}(t) = V\mathbf{w}(t). \tag{106}$$

That is an acceptable form of definition because V is invertible. Equation (103) becomes

$$V\dot{\mathbf{w}}(t) = -\frac{\lambda\theta}{N} V\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) \tag{107}$$

$$+\frac{\lambda\theta}{N^2} VDV^{-1}V\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) - \mu V\mathbf{w}(t). \tag{108}$$

Pre-multiplying this equation by V^{-1} yields the following simplification,

$$\dot{\mathbf{w}}(t) = -\frac{\lambda\theta}{N}\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) \tag{109}$$

$$+\frac{\lambda\theta}{N^2}D\big(\mathbf{w}(t-\Delta) + \delta\mathbf{w}'(t-\Delta)\big) - \mu\mathbf{w}(t). \tag{110}$$

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

$$\dot{w}_1(t) = -\mu w_1(t), \tag{111}$$

$$\dot{w}_i(t) = -\frac{\lambda\theta}{N}\big(w_i(t-\Delta) + \delta\dot{w}_i(t-\Delta)\big) - \mu w_i(t), \qquad i \neq 1. \tag{112}$$

Equation (111) 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 (112). □

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 4.4. *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:

$$-(a+\mu)N = e^{-a\Delta}\lambda\theta\Big((1+a\delta)\cos(b\Delta) + b\delta\sin(b\Delta)\Big) \tag{113}$$

$$bN = e^{-a\Delta}\lambda\theta\Big(-b\delta\cos(b\Delta) + (1+a\delta)\sin(b\Delta)\Big). \tag{114}$$

We find the expressions for sine and cosine to be

$$\cos(b\Delta) = -\frac{e^{a\Delta}\Big((a+\mu)(1+a\delta)N + b^2\delta N\Big)}{\lambda\theta\Big((a\delta+1)^2 + (b\delta)^2\Big)} \tag{115}$$

$$\sin(b\Delta) = -\frac{e^{a\Delta}Nb(\delta\mu - 1)}{\lambda\theta\big((1+a\delta)^2 + (b\delta)^2\big)}. \tag{116}$$

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

$$b = \sqrt{\frac{\lambda^2\theta^2(a\delta+1)^2 - e^{2a\Delta}N^2(a+\mu)^2}{e^{2a\Delta}N^2 - \delta^2\lambda^2\theta^2}}. \tag{117}$$

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 Equation (117) 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 be real. This yields two inequalities

$$\frac{\delta\theta\lambda}{N} > e^{a\Delta} > \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}. \tag{118}$$

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 inequality from Equation (118) 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 (118) hold, so by Equation (117) we have $0 \neq b \in R$ for all $a \in (a_1, a_2)$.

If $b \in \mathbb{R}$ satisfies Equation (116) 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 Equation (116) will oscillate between -1 and 1 an infinite number of times, while the left hand side of Equation (116) will converge to 0. Hence, there are infinitely many solutions to Equation (116) 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 Equation (117) is real-valued, we will determine an interval in the range of a where the numerator and the denominator of b are

positive. We obtain the condition

$$\frac{\delta\theta\lambda}{N} < e^{a\Delta} < \frac{\theta\lambda(1+a\delta)}{N(a+\mu)}. \tag{119}$$

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 Equation (116) 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 (99).

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

$$\cos(b\Delta) = -\frac{e^{a\Delta}N}{\lambda\theta\delta}, \quad \sin(b\Delta) = 0, \tag{120}$$

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 (99) has infinitely many eigenvalues with positive real parts. □

However, when the weight coefficient δ is sufficiently small, i.e. $\delta < \frac{N}{\lambda\theta}$, then given a sufficiently small delay the queues converge to a locally stable equilibrium. As the next result shows, the stability is due to all eigenvalues

having negative real parts.

Proposition 4.5. *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 (99). We can assume $b \geq 0$ without loss of generality. Plugging in R and separating the real and imaginary parts:

$$-(a+\mu)N = e^{-a\Delta}\lambda\theta\Big((1+a\delta)\cos(b\Delta) + b\delta\sin(b\Delta)\Big) \tag{121}$$

$$bN = e^{-a\Delta}\lambda\theta\Big(-b\delta\cos(b\Delta) + (1+a\delta)\sin(b\Delta)\Big). \tag{122}$$

Solving for sine and cosine, we find

$$\cos(b\Delta) = -\frac{e^{a\Delta}\Big((a+\mu)(1+a\delta)N + b^2\delta N\Big)}{\lambda\theta\Big((a\delta+1)^2 + (b\delta)^2\Big)} \tag{123}$$

$$\sin(b\Delta) = -\frac{e^{a\Delta}Nb(\delta\mu - 1)}{\lambda\theta\big((1+a\delta)^2 + (b\delta)^2\big)} \tag{124}$$

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

$$b = \sqrt{\frac{\lambda^2\theta^2(a\delta+1)^2 - e^{2a\Delta}N^2(a+\mu)^2}{e^{2a\Delta}N^2 - \delta^2\lambda^2\theta^2}}. \tag{125}$$

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 non-negative. Therefore we get inequalities

$$1 \leq e^{a\Delta} \leq \frac{\lambda\theta(a\delta+1)}{N(a+\mu)}, \quad e^{a\Delta} > \frac{\delta\lambda\theta}{N}. \tag{126}$$

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 (125) and (126) to find an upper bound on b:

$$b \leq B = \sqrt{\frac{2N\lambda^2\theta^2(1-\delta\mu)^2}{N^2 - \delta^2\lambda^2\theta^2}}, \quad B > 0. \tag{127}$$

However, we note from Equation (123) that $\cos(b\Delta) < 0$. Since b is non-negative 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 our assumption in Equation (127):

$$b > \frac{\pi}{2\Delta} > \frac{\pi}{2\Delta_0} > 2B. \tag{128}$$

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 (99), as all eigenvalues will be purely imaginary. We justify these findings in the result below.

Proposition 4.6. *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 4.4 - 4.5, 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

$$\sin(b\Delta) = -\frac{bNe^{a\Delta}(\mu N - \theta\lambda)}{N^2b^2 + (aN + \theta\lambda)^2} \tag{129}$$

$$\cos(b\Delta) = -\frac{Ne^{a\Delta}\left(a^2N + a\theta\lambda + a\mu N + b^2N + \theta\lambda\mu\right)}{N^2b^2 + (aN + \theta\lambda)^2} \tag{130}$$

We will address the three cases separately.

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

$$b = \frac{\sqrt{(aN + \theta\lambda)^2 - N^2e^{2a\Delta}(a + \mu)^2}}{\sqrt{N^2\left(e^{2a\Delta} - 1\right)}} \tag{131}$$

where the denominator is positive, so the numerator must be non-negative 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

$$aN + \theta\lambda - Ne^{a\Delta}(a + \mu) \leq aN + \theta\lambda - N(a + \mu) = \theta\lambda - N\mu < 0, \tag{132}$$

and we reach 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 Equation (131). 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 Equation (129) goes to zero. Therefore, Equation (129) will have infinitely many roots, while Equation (130) will be satisfied at each root automatically since b is given by (131). 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 the Equations (129) - (130) to be

$$\sin(b\Delta) = 0, \qquad \cos(b\Delta) = -e^{a\Delta}. \tag{133}$$

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 (99) are purely imaginary. □

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

$$\omega_{cr} = \sqrt{\frac{\lambda^2\theta^2 - N^2\mu^2}{N^2 - \delta^2\lambda^2\theta^2}} \tag{134}$$

and Δ_{cr} satisfies the transcendental equation

$$\cos\left(\Delta_{cr}\sqrt{\frac{\lambda^2\theta^2 - N^2\mu^2}{N^2 - \delta^2\lambda^2\theta^2}}\right) = -\frac{\delta\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta\mu)}. \tag{135}$$

Proof. Assume that R from the characteristic equation (99) is purely imaginary, $R = \pm i\omega_{cr}$. Plugging in R, the real and imaginary parts produce two equations

$$\mu = -\frac{\lambda\theta}{N}\cos(\omega_{cr}\Delta_{cr}) - \frac{\lambda\theta}{N}\delta\omega_{cr}\sin(\omega_{cr}\Delta_{cr}) \tag{136}$$

$$\omega_{cr} = \frac{\lambda\theta}{N}\sin(\omega_{cr}\Delta_{cr}) - \frac{\lambda\theta}{N}\delta\omega_{cr}\cos(\omega_{cr}\Delta_{cr}) \tag{137}$$

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

$$\cos(\omega_{cr}\Delta_{cr}) = -\frac{N(\mu + \delta\omega_{cr}^2)}{\lambda\theta(1+\delta^2\omega_{cr}^2)}, \quad \sin(\omega_{cr}\Delta_{cr}) = \frac{N\omega_{cr}(1-\delta\mu)}{\lambda\theta(1+\delta^2\omega_{cr}^2)} \tag{138}$$

and by the trigonometric identity $\sin^2(\omega_{cr}\Delta_{cr}) + \cos^2(\omega_{cr}\Delta_{cr}) = 1$, ω_{cr} is

found. The cosine equation from (138) then gives the equation for Δ_{cr}. $\square$

Proposition 4.7 provides the infinitely many critical delays Δ_{cr} as well as the necessary conditions on the other parameters ($\omega_{cr} \in \mathbb{R}$, $\omega_{cr} \neq 0$) for when Hopf bifurcations may occur. This information allows us to prove that a Hopf bifurcation occurs at every Δ_{cr}.

Theorem 4.8. *Suppose ω_{cr} from Equation* (134) *is real and nonzero. Then a Hopf bifurcation occurs at $\Delta = \Delta_{cr}$, where Δ_{cr} is any positive root of*

$$\Delta_{cr}(\lambda, \mu, \theta, N, \delta) = \arccos\left(- \frac{\delta\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta\mu)} \right) \cdot \sqrt{\frac{N^2 - \delta^2\lambda^2\theta^2}{\lambda^2\theta^2 - N^2\mu^2}}. \tag{139}$$

Proof. By Proposition 4.7, at each Δ_{cr} there is a pair of purely imaginary eigenvalues $R = i\omega_{cr}$, $\bar{R} = -i\omega_{cr}$. A Hopf bifurcation can only occur if $\frac{d}{d\Delta}\mathrm{Re}[R(\Delta_{cr})] \neq 0$. To verify this, we assume that $R(\Delta) = \alpha(\Delta) + i\omega(\Delta)$. The characteristic equation (99) is differentiated with respect to delay, and we find that at Δ_{cr} where $\alpha = 0$ and $\omega = \omega_{cr}$, $\frac{d}{d\Delta}\mathrm{Re}[R]$ is given by

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

The assumption $\omega_{cr} > 0$ guarantees the numerator of $\frac{d\alpha}{d\Delta}(\Delta_{cr})$ 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

$$\frac{dD}{d\Delta}(\Delta^*) = 0 \implies \Delta^* = \frac{\delta\lambda^2\theta^2 - N^2\mu}{\lambda^2\theta^2(1+\delta^2\omega_{cr}^2)}. \tag{140}$$

Once Δ^* and $\omega = \omega_{cr}$ from Equations (140) and (134) are substituted into the denominator $D(\Delta)$, we find that the minimum of D with respect to Δ is positive:

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

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

$$\frac{d\alpha}{d\Delta}(\Delta_{cr}) \neq 0. \tag{141}$$

In fact, if $\delta < \frac{N}{\lambda\theta}$ then $\frac{d\alpha}{d\Delta}(\Delta_{cr}) > 0$ so the eigenvalues always cross from left to right on the complex plane. If $\delta > \frac{N}{\lambda\theta}$ then $\frac{d\alpha}{d\Delta}(\Delta_{cr}) < 0$ 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 [18] are satisfied, so a Hopf bifurcation occurs at every root Δ_{cr}. □

Theorem 4.8 provides an explicit expression for the critical delay Δ_{cr}, which implies that if the delay $\Delta > \Delta_{cr}$, 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 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 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.

In the proof of Theorem 4.8, 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 the conditions for the local stability of the equilibrium.

Theorem 4.9. *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 4.6, 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 4.5 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, $\omega_{cr} = \sqrt{\frac{\lambda^2\theta^2 - N^2\mu^2}{N^2 - \delta^2\lambda^2\theta^2}} \in \mathbb{R}$, $\omega_{cr} \neq 0$ must hold. In case when $\lambda\theta \leq N\mu$ and $\delta < \frac{N}{\lambda\theta}$, then either $\omega_{cr} \notin \mathbb{R}$ or $\omega_{cr} = 0$ 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 4.6 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 Equation (95) can be categorized by 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 Figure 34. In the following discussion of the two cases, we will refer to this diagram and will explain it in detail.

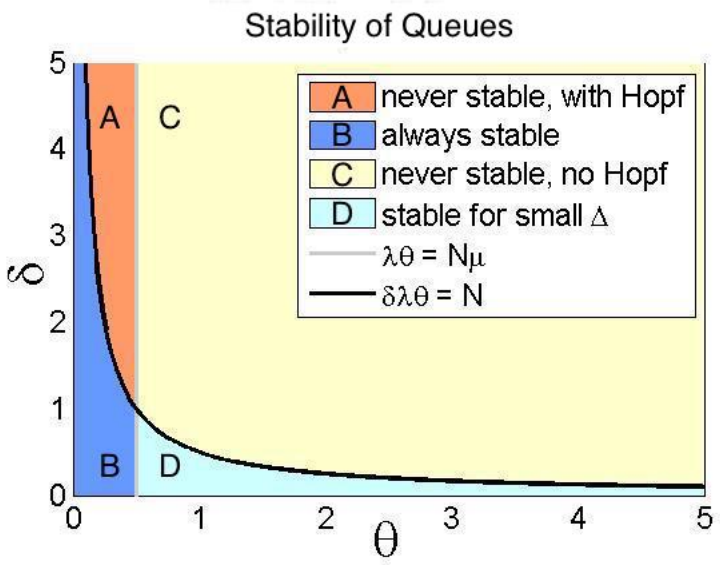

Figure 34: The four stability cases.

Before explaining Figure 34 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 the system becomes unstable 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 similar to the no velocity case and we have oscillations if Δ is large enough as usual. From these different 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 Figure 34. 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 Figure 35. However, when $\delta > \frac{N}{\lambda\theta}$, which is region A of Figure 34, 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 will increase, the complex pairs will cross the imaginary axis from right to left, causing Hopf bifurcations to occur as shown in Figure 36. Note, however, that queues will never gain stability because for any delay Δ there will be eigenvalues with positive real parts.

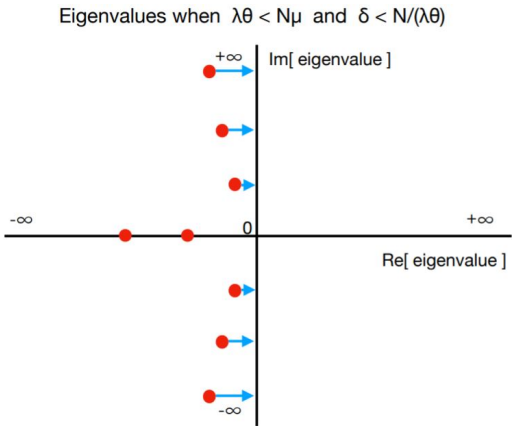

Figure 35: Eigenvalues have negative real parts for all Δ.

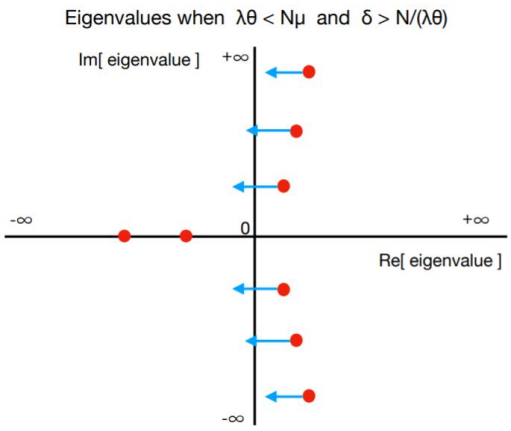

Figure 36: Eigenvalues cross the imaginary axis as Δ increases.

Case 2: $\lambda\theta > N\mu$**.** This case is represented by the regions C and D in Figure 34. 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 Figure 37. During the crossing of each pair, a Hopf bifurcation occurs. When $\delta \geq \frac{N}{\lambda\theta}$, which is region C of the diagram from Figure 34, 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 Figure 38, so there will never be a stable equilibrium.

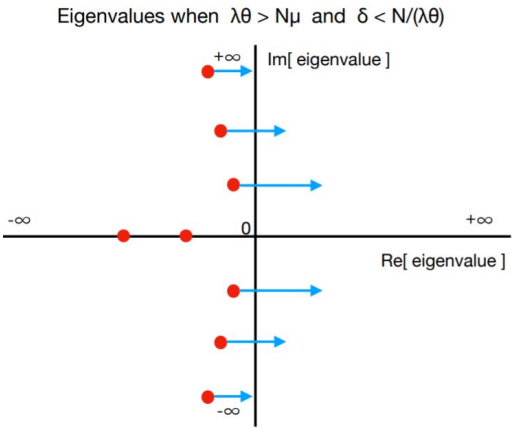

Figure 37: Eigenvalues cross the imaginary axis as Δ increases.

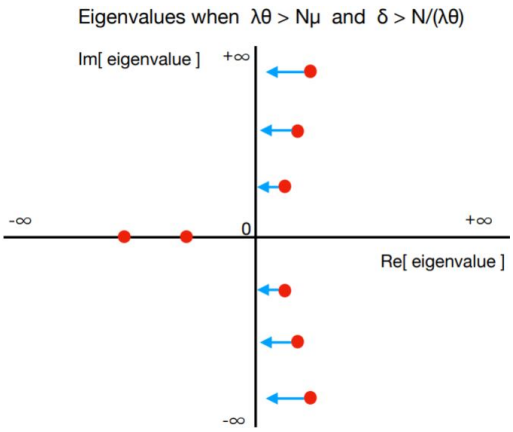

Figure 38: Eigenvalues have positive real parts for all Δ.

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 region B of Figure (34), 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 Figure (34), so the queues will always be unstable.

4.4 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 experience. It is therefore useful to know when providing extra information helps make queues more stable. For example, consider a numerical example from Figures 39 - 42, with two queues and

fixed parameters λ, θ, μ, and Δ. In Figures 39 and 41, $\Delta < \Delta_{cr}$ so the queues converge to an equilibrium over time. However, in Figures 40 and 42 we have $\Delta > \Delta_{cr}$, so the queues oscillate indefinitely. Although the delay Δ is the same, the change of behavior results from tweaking the parameter δ, which consequently regulates the bifurcation threshold Δ_{cr}.

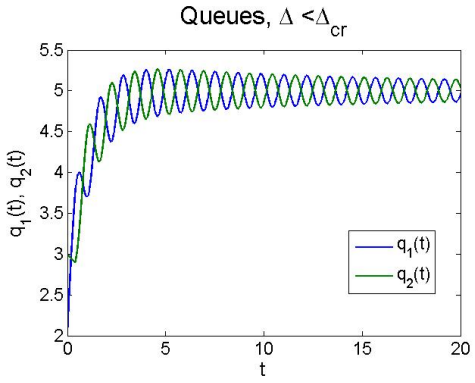

Figure 39: Queues before bifurcation; $\delta = 0.08$, $\lambda = 10$, $\mu = 1$, $\theta = 1$.

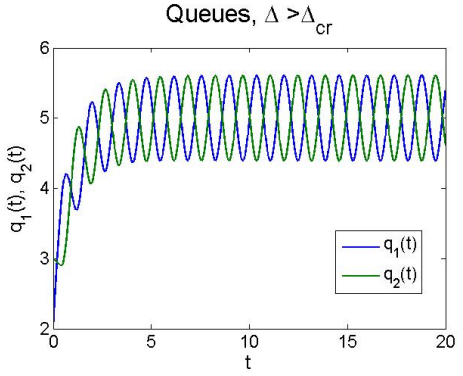

Figure 40: Queues after bifurcation; $\delta = 0$, $\lambda = 10$, $\mu = 1$, $\theta = 1$.

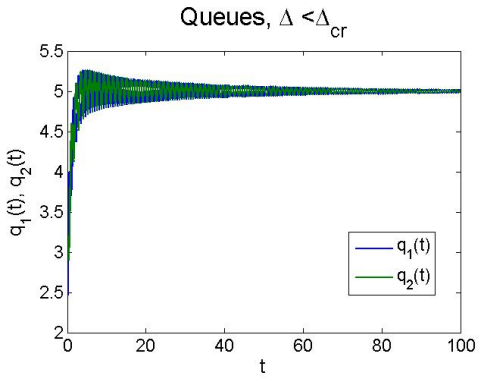

Figure 41: Queues before bifurcation; $\delta = 0.08$, $\lambda = 10$, $\mu = 1$, $\theta = 1$.

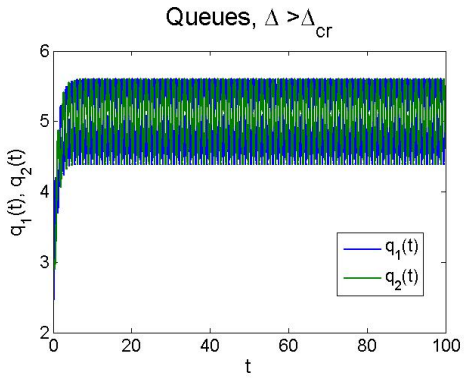

Figure 42: Queues after bifurcation; $\delta = 0$, $\lambda = 10$, $\mu = 1$, $\theta = 1$.

In this Subsection, 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 4.10. *Suppose $\lambda\theta > N\mu$. Then the function $\Delta_{cr}(\delta)$ is concave for all $\delta \in [0, \frac{N}{\lambda\theta})$.*

Proof. The critical delay Δ_{cr} is given by Equation (139). It is clear that the second derivative $\frac{d^2\Delta_{cr}}{d\delta^2}$ is negative for all $\delta \in [0, \frac{N}{\lambda\theta})$:

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

$$C_1 = (N^2 - \delta^2\lambda^2\theta^2)(\lambda^2\theta^2 - N^2\mu^2)(\delta\lambda^2\theta^2 + N^2\mu) > 0$$

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

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

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

□

Proposition 4.10 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 δ_{max}, and it is such that $\delta = \delta_{max}$ maximizes the threshold Δ_{cr}. In Proposition 4.11, we give an equation that determines δ_{max}, and provide closed-form expressions for an upper and a lower bound of δ_{max}.

Proposition 4.11. *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 solution of*

$$\frac{\sqrt{N^2 - \delta_{max}^2\lambda^2\theta^2}}{1+\delta_{max}\mu} = \frac{\delta_{max}\lambda^2\theta^2}{\sqrt{\lambda^2\theta^2 - N^2\mu^2}} \cdot \arccos\bigg(- \frac{\delta_{max}\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta_{max}\mu)}\bigg). \quad (142)$$

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

$$\begin{aligned}
\delta_1 &= \frac{-\big(\Delta_0 + \frac{N}{\lambda\theta}\big)\lambda\theta}{2\lambda\theta\Big(1 + \big(\Delta_0 + \frac{N}{\lambda\theta}\big)\mu\Big)} && (143)\\
&+ \frac{\sqrt{\lambda^2\theta^2\big(\Delta_0 + \frac{N}{\lambda\theta}\big)^2 + 4N^2\big(\Delta_0 + \frac{N}{\lambda\theta}\big)\mu + 4N^2}}{2\lambda\theta\Big(1 + \big(\Delta_0 + \frac{N}{\lambda\theta}\big)\mu\Big)} && (144)\\
\delta_2 &= \frac{-\Delta_0\lambda\theta + \sqrt{\lambda^2\theta^2\Delta_0^2 + 4N^2\Delta_0\mu + 4N^2}}{2\lambda\theta(1+\Delta_0\mu)} && (145)\\
\Delta_0 &= \arccos\bigg(- \frac{N\mu}{\lambda\theta}\bigg) \cdot \sqrt{\frac{N^2}{\lambda^2\theta^2 - N^2\mu^2}}. && (146)
\end{aligned}$$

Proof. We can treat Δ_{cr} as a function of δ. The implicit differentiation of (135) gives the rate with which Δ_{cr} changes:

$$\begin{aligned}
\frac{d}{d\delta}\Delta_{cr}(\delta) &= \frac{N^2 - \delta\lambda^2\theta^2\Big(\delta + \Delta_{cr}(\delta) + \delta\mu\Delta_{cr}(\delta)\Big)}{(N^2 - \delta^2\lambda^2\theta^2)(1+\delta\mu)} = \frac{1}{1+\delta\mu} - \frac{\delta\lambda^2\theta^2\Delta_{cr}(\delta)}{N^2 - \delta^2\lambda^2\theta^2} && (147)\\
&= \frac{1}{1+\delta\mu} - \frac{\delta\lambda^2\theta^2}{N^2 - \delta^2\lambda^2\theta^2} \cdot \arccos\bigg(- \frac{\delta\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta\mu)}\bigg) \cdot \sqrt{\frac{N^2 - \delta^2\lambda^2\theta^2}{\lambda^2\theta^2 - N^2\mu^2}} && (148)
\end{aligned}$$

By Proposition 4.10, $\Delta_{cr}(\delta)$ is concave on the interval $[0, \frac{N}{\lambda\theta})$. Further, it can be shown that $\frac{d}{d\delta}\Delta_{cr}(0) = 1 > 0$ and $\lim_{\delta \to \frac{N}{\lambda\theta}} \frac{d}{d\delta}\Delta_{cr}(\delta) = -\infty < 0$, so

there is a point δ_{max} where $\frac{d}{d\delta}\Delta_{cr}(\delta_{max}) = 0$. Therefore $\Delta_{cr}(\delta)$ reaches its absolute maximum at $\delta_{max} \in (0, \frac{N}{\lambda\theta})$. Figure 43 plots $\frac{d}{d\delta}\Delta_{cr}(\delta)$.

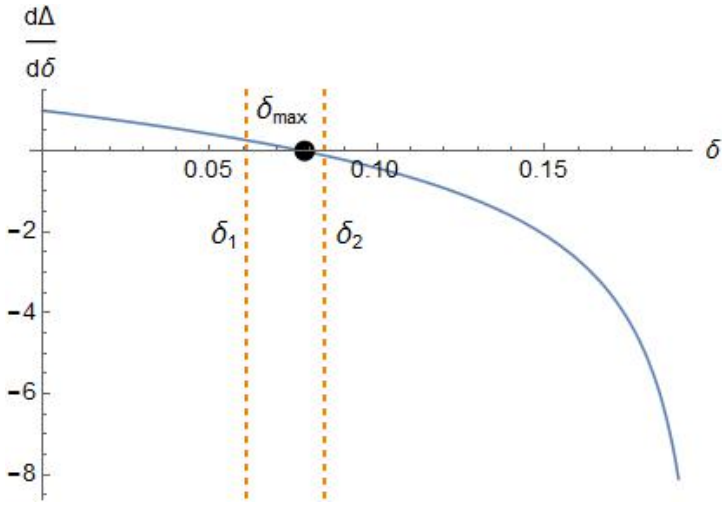

Figure 43: δ_{max} and its bounds $\delta_1 < \delta_{max} < \delta_2$.

The value δ_{max} can be found numerically by solving $\frac{d}{d\delta}\Delta_{cr}(\delta_{max}) = 0$ from Equation (148), alternatively written as (142). It is left to find closed-form expressions for the bounds on δ_{max}. By Equation (147), we can find

$$\frac{d}{d\delta}\Delta_{cr}(\delta_{max}) = \frac{1}{1+\delta_{max}\mu} - \frac{\delta_{max}\lambda^2\theta^2\Delta_{cr}(\delta_{max})}{N^2 - \delta_{max}^2\lambda^2\theta^2} = 0, \tag{149}$$

$$\frac{1}{1+\delta_{max}\mu} - \frac{\delta_{max}\lambda^2\theta^2\Delta_0}{N^2 - \delta_{max}^2\lambda^2\theta^2} > 0, \tag{150}$$

where $\Delta_0 = \Delta_{cr}(0) < \Delta_{cr}(\delta_{max})$. When solved for δ_{max}, the inequality (150) produces an upper bound condition $\delta_{max} < \delta_2$ given by Equation (145).

To find the lower bound, we note that $\frac{d}{d\delta}\Delta_{cr}(\delta)$ is monotonically decreasing. Thus, $\frac{d}{d\delta}\Delta_{cr}(\delta) < \frac{d}{d\delta}\Delta_{cr}(0) = 1$ for all $\delta \in (0, \frac{N}{\lambda\theta})$, and $\Delta_{cr}(\delta) \leq$

$\delta + \Delta_{cr}(0) < \frac{N}{\lambda\theta} + \Delta_0$. Therefore by Equation (149), we get

$$\frac{1}{1+\delta_{max}\mu} - \frac{\delta_{max}\lambda^2\theta^2\left(\Delta_0 + \frac{N}{\lambda\theta}\right)}{N^2 - \delta_{max}^2\lambda^2\theta^2} < 0, \tag{151}$$

which produces the bound $\delta_{max} > \delta_1$ from Equation (143). □

Figures 44 - 45 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, Figures 46 and 47 show Δ_{cr} as a function of μ and δ with the two curves δ_1 and δ_2. As seen in the Figures 44 - 47, the bounds on δ_{max} are tight.

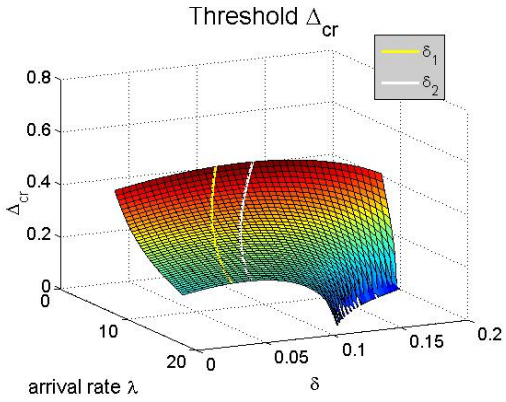

Figure 44: For each λ, the maximum Δ_{cr} is achieved when $\delta \in (\delta_1, \delta_2)$.

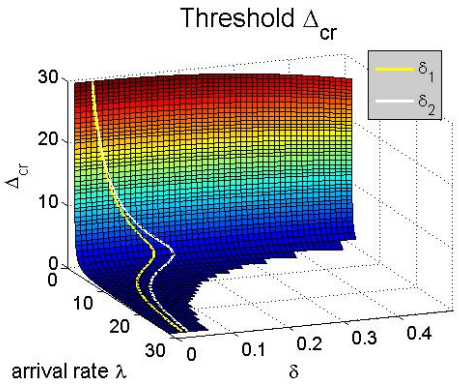

Figure 45: For each λ, the maximum Δ_{cr} is achieved when $\delta \in (\delta_1, \delta_2)$.

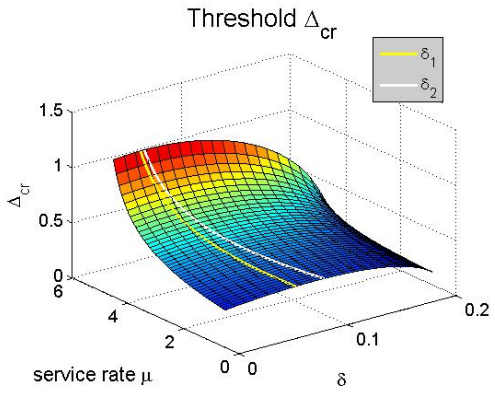

Figure 46: For each μ, the maximum Δ_{cr} is achieved when $\delta \in (\delta_1, \delta_2)$.

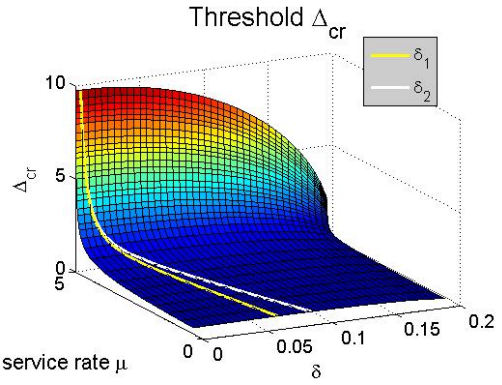

Figure 47: For each μ, the maximum Δ_{cr} is achieved when $\delta \in (\delta_1, \delta_2)$.

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.

Proposition 4.12. *The maximum value of a root Δ_{cr} for fixed parameters λ, μ, and N, is attained at δ_{max} and is bounded by $\Delta_1 < \Delta_{cr}(\delta_{max}) < \Delta_2$, where*

$$\Delta_1 = \max[\Delta_{cr}(\delta_1), \Delta_{cr}(\delta_2)], \quad \Delta_2 = \min[\Delta_{2a}, \Delta_{2b}], \tag{152}$$

$$\Delta_{2a} = \Delta_{cr}(\delta_1) + (\delta_2 - \delta_1) \cdot \frac{d}{d\delta}\Delta_{cr}(\delta_1), \tag{153}$$

$$\Delta_{2b} = \Delta_{cr}(\delta_2) - (\delta_2 - \delta_1) \cdot \frac{d}{d\delta}\Delta_{cr}(\delta_2). \tag{154}$$

Proof. By Proposition 4.11, $\Delta_{cr}(\delta)$ attains its maximum at $\delta = \delta_{max}$. Hence the lower bound $\Delta_1 < \Delta_{cr}(\delta_{max})$ trivially follows, since $\delta_{max} \neq \delta_1, \delta_2$. To find

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

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

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

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

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

Figure 48 illustrates $\Delta_{cr}(\delta) - \Delta_{cr}(0)$ as a function of δ, with the maximum attained at δ_{max} and the bounds on the maximum given by Δ_1 and Δ_2. Further, it is evident from Figure 48 that there is a threshold value, which we call δ_{cap}, that places a cap on the potential utility of the velocity information. When δ is less than δ_{cap}, the queueing system becomes more stable from the velocity information because $\Delta_{cr}(\delta) > \Delta_{cr}(0)$. However, when δ exceeds δ_{cap}, the queues become more unstable in the sense that $\Delta_{cr}(\delta) < \Delta_{cr}(0)$. The result below provides an equation for δ_{cap}.

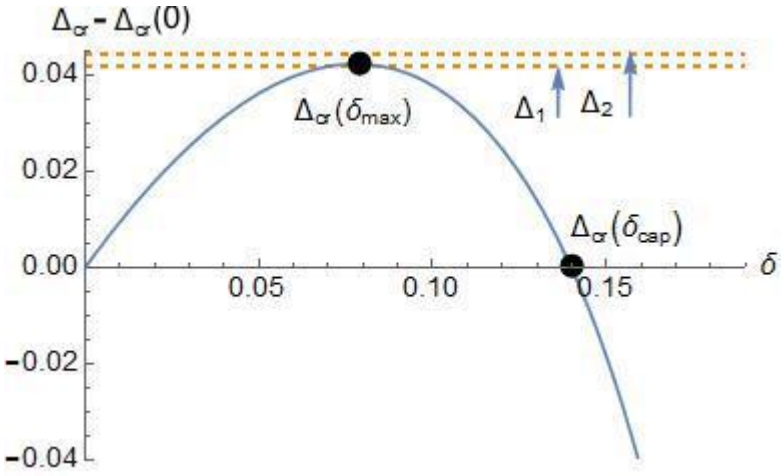

Figure 48: δ_{max} maximizes Δ_{cr}.

Proposition 4.13. *Suppose $\lambda\theta > N\mu$. There exists a unique $\delta_{cap} > 0$ such that $\Delta_{cr}(\delta) > \Delta_{cr}(0)$ for all $\delta < \delta_{cap}$, and $\Delta_{cr}(\delta) < \Delta_{cr}(0)$ for all $\delta > \delta_{cap}$. It is given by the solution to*

$$\begin{aligned} &\arccos\left(-\frac{N\mu}{\lambda\theta}\right)\sqrt{\frac{N^2}{\lambda^2\theta^2 - N^2\mu^2}} \\ &= \arccos\left(-\frac{\delta_{cap}\lambda^2\theta^2 + N^2\mu}{N\lambda\theta(1+\delta_{cap}\mu)}\right)\sqrt{\frac{N^2 - \delta_{cap}^2\lambda^2\theta^2}{\lambda^2\theta^2 - N^2\mu^2}}. \end{aligned} \tag{155}$$

Proof. As previously shown, $\Delta_{cr}(\delta)$ is monotonically increasing on $\delta \in [0, \delta_{max})$ and monotonically decreasing on $\delta \in (\delta_{max}, \frac{N}{\lambda\theta})$. Further, $\lim_{\delta \to \frac{N}{\lambda\theta}} \Delta_{cr}(\delta) = 0 < \Delta_{cr}(0)$ since $\Delta_{cr}(0) > 0$ by assumption, so there exists exactly one point δ_{cap} on the interval $(\delta_{max}, \frac{N}{\lambda\theta})$ where $\Delta_{cr}(\delta_{cap}) = \Delta_{cr}(0)$, and it also follows that $\Delta_{cr}(\delta_{cap}) > \Delta_{cr}(0)$ for all $\delta < \delta_{cap}$ and $\Delta_{cr}(\delta_{cap}) < \Delta_{cr}(0)$ for all $\delta > \delta_{cap}$. By substituting the expression for Δ_{cr} from (139) into $\Delta_{cr}(0) - \Delta_{cr}(\delta_{cap}) = 0$

we get Equation (155). □

To summarize, when $\lambda\theta > N\mu$, the queues are stable when the delay is less than Δ_{cr}. We can therefore provide the most stability for the queues by choosing δ that maximizes Δ_{cr}, i.e. δ_{max}. Proposition 4.11 proves the existence of δ_{max}, gives an equation describing δ_{max}, and provides closed-form expressions for bounds δ_1 and δ_2 such that $\delta_1 < \delta_{max} < \delta_2$. Proposition 4.12 also provides bounds Δ_1 and Δ_2 for the maximum value that Δ_{cr} can take as a function of δ, so $\Delta_1 < \Delta_{cr}(\delta_{max}) < \Delta_2$. Lastly, we show that even if $\delta \neq \delta_{max}$, it is still beneficial to include the velocity information as long as $\delta < \delta_{cap}$. When δ exceeds δ_{cap}, however, $\Delta_{cr}(\delta)$ becomes less than $\Delta_{cr}(0)$, so the queues are less likely to be stable than if the velocity information was omitted altogether. Proposition 4.13 proves the existence of δ_{cap} and provides the equation for it.

4.5 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. As seen in Chapters 2 and 3, we must use third order Taylor expansions to obtain information about the amplitude. Thus, many of the matrix techniques we exploited so far to linearize the NDDEs, 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

$$\dot{q}_1(t) = \lambda \cdot \frac{\exp\Big(-\theta\big(q_1(t-\Delta) + \delta\dot{q}_1(t-\Delta)\big)\Big)}{\sum_{j=1}^{2}\exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_j(t-\Delta)\big)\Big)} - \mu q_1(t) \tag{156}$$

$$\dot{q}_2(t) = \lambda \cdot \frac{\exp\Big(-\theta\big(q_2(t-\Delta) + \delta\dot{q}_2(t-\Delta)\big)\Big)}{\sum_{j=1}^{2}\exp\Big(-\theta\big(q_j(t-\Delta) + \delta\dot{q}_j(t-\Delta)\big)\Big)} - \mu q_2(t), \tag{157}$$

where as usual $\Delta, \lambda, \mu, \theta > 0$ and $\delta \geq 0$. Similar to Section 4.4, we will consider the scenario with $\lambda\theta > 2\mu$, where even for a small δ the equilibrium of queues becomes unstable for sufficiently large delay. Our first result shows that the Hopf bifurcations that occur at each root Δ_{cr} are supercritical.

Theorem 4.14. *Suppose $\omega_{cr} \in \mathbb{R}$ and $\omega_{cr} \neq 0$. The NDDE system* (156) - (157) *undergoes a supercritical Hopf bifurcation at each root Δ_{cr}. If $\delta\mu < 1$ then the limit cycle is born when $\Delta \leq \Delta_{cr}$. If $\delta\mu > 1$ then the limit cycle is born when $\Delta \geq \Delta_{cr}$.*

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 4.8. This method is often applied to systems of delay differential equations (DDEs) [10, 6, 29]. 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. [19].

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

$$w_1(t) = q_1(t) + q_2(t), \quad w_2(t) = q_1(t) - q_2(t) \tag{158}$$

$$\dot{w}_1(t) = -\mu w_1(t) \tag{159}$$

$$\dot{w}_2(t) = -\mu w_2(t) - \frac{\lambda\theta}{2}(w_2(t-\Delta) + \delta\dot{w}_2(t-\Delta)) \tag{160}$$

$$+\frac{\lambda\theta^3}{24}(w_2(t-\Delta) + \delta\dot{w}_2(t-\Delta))^3 + O(w_2^4). \tag{161}$$

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:

$$\dot{x}(t) = -\mu x(t) - \frac{\lambda\theta}{2}(x(t-\Delta) + \delta\dot{x}(t-\Delta)) + \frac{\sqrt{\epsilon}\lambda\theta^3}{24}(x(t-\Delta) + \delta\dot{x}(t-\Delta))^3 \tag{162}$$

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:

$$\Delta = \Delta_{cr} + \epsilon\alpha, \quad \omega = \omega_{cr} + \epsilon\beta. \tag{163}$$

The time derivative $\dot{x}$ becomes

$$\dot{x} = \frac{dx}{dt} = \frac{\partial x}{\partial \xi}\frac{d\xi}{dt} + \frac{\partial x}{\partial \eta}\frac{d\eta}{dt} = \frac{\partial x}{\partial \xi} \cdot (\omega_{cr} + \epsilon\beta) + \frac{\partial x}{\partial \eta} \cdot \epsilon. \tag{164}$$

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

$$x(t - \Delta) = x(\xi - \omega\Delta, \eta - \epsilon\Delta) \tag{165}$$

$$= x(\xi - (\omega_{cr} + \epsilon\beta)(\Delta_{cr} + \epsilon\alpha), \eta - \epsilon(\Delta_{cr} + \epsilon\alpha)) + O(\epsilon^2) \tag{166}$$

$$= \tilde{x} - \epsilon(\omega_{cr}\alpha + \Delta_{cr}\beta) \cdot \frac{\partial \tilde{x}}{\partial \xi} - \epsilon\Delta_{cr}\frac{\partial \tilde{x}}{\partial \eta} + O(\epsilon^2), \tag{167}$$

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

$$\frac{dx}{dt} = \omega_{cr}\frac{\partial x_0}{\partial \xi} + \epsilon\beta\frac{\partial x_0}{\partial \xi} + \epsilon\frac{\partial x_0}{\partial \eta} + \epsilon\omega_{cr}\frac{\partial x_1}{\partial \xi}. \tag{168}$$

After these substitutions are made into (162), 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

$$L(x_0) = \mu x_0 + \frac{\lambda}{2}\tilde{x}_0 + \omega_{cr}\frac{\partial x_0}{\partial \xi} + \frac{\delta\lambda\omega_{cr}}{2}\frac{\partial \tilde{x}_0}{\partial \xi} = 0, \tag{169}$$

which is satisfied with a solution of the form

$$x_0(t) = A(\eta)\cos(\xi) + B(\eta)\sin(\xi). \tag{170}$$

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 (170), 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

$$\begin{aligned}
\frac{dR}{d\eta} &= -\frac{R\big(c_1 R^2 - c_2\big)}{c_3}, \quad \text{where} && (171)\\
c_1 &= (\mu^2 + \omega_{cr}^2)(\mu^2 + \omega_{cr}^2 + \delta^2\omega_{cr}^2\mu^2 + \delta^2\omega_{cr}^4)\Delta_{cr} && (172)\\
&+ (\mu^2 + \omega_{cr}^2)(\mu - \delta\mu^2 + \delta^2\omega_{cr}^2\mu - \delta\omega_{cr}^2), && (173)\\
c_2 &= 4\alpha\lambda^2\omega_{cr}^2(1 - \delta^2\mu^2), && (174)\\
c_3 &= \Delta_{cr}^2 \cdot 4\lambda^2(\mu^2 + \omega_{cr}^2 + \delta^2\mu^2\omega^2 + \delta^2\omega_{cr}^2) && (175)\\
&+ \Delta_{cr} \cdot 8\lambda^2(\mu - \delta\mu^2 - \delta\omega_{cr}^2 + \delta^2\mu\omega_{cr}^2) + 4\lambda^2(1 - \delta\mu)^2. && (176)
\end{aligned}$$

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

$$\Delta^*_{cr} = \frac{\delta}{1+\delta^2\omega^2_{cr}} - \frac{\mu}{\mu^2+\omega^2_{cr}} \tag{177}$$

$$c_3 = c_3(\Delta_{cr}) \geq c_3(\Delta^*_{cr}) = \frac{4\lambda^2\omega^2_{cr}(1-\delta^2\mu^2)^2}{(\mu^2+\omega^2_{cr})(1+\delta^2\omega^2_{cr})} > 0 \tag{178}$$

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, Equation (138) must be satisfied so $\cos(\omega_{cr}\Delta_{cr}) < 0$, which implies that $\omega_{cr}\Delta_{cr} > \frac{\pi}{2}$ and

$$\Delta_{cr} > \frac{\pi}{2\omega_{cr}}. \tag{179}$$

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

$$\Delta^*_{cr} = \frac{\delta}{1+\delta^2\omega^2_{cr}} - \frac{\mu}{\mu^2+\omega^2_{cr}}. \tag{180}$$

Using the inequality in (179), we can show by contradiction that Δ_{cr} is always greater than Δ^*_{cr}. Suppose that for some parameters, we have $\Delta^*_{cr} > \frac{\pi}{2\omega_{cr}}$.

From the equation (134), this implies that

$$\frac{\pi}{2\omega_{cr}} < \Delta^*_{cr} = \frac{\delta}{1+\delta^2\omega_{cr}^2} - \frac{\mu}{\mu^2+\omega_{cr}^2} < \frac{\delta}{1+\delta^2\omega_{cr}^2} \tag{181}$$

$$\frac{\pi}{2}(1+\delta^2\omega_{cr}^2) < \delta\omega_{cr} \tag{182}$$

$$\frac{2\pi(1-\delta^2\mu^2)}{4-\delta^2\lambda^2\theta^2} < \delta\sqrt{\frac{\lambda^2\theta^2-4\mu^2}{4-\delta^2\lambda^2\theta^2}} \tag{183}$$

$$4\pi^2\cdot\frac{(1-\delta^2\mu^2)^2}{(4-\delta^2\lambda^2\theta^2)^2} < \delta^2\cdot\frac{\lambda^2\theta^2-4\mu^2}{4-\delta^2\lambda^2\theta^2} \tag{184}$$

$$4\pi^2\Big(1-\delta^2\mu^2\Big)^2 < \delta^2(\lambda^2\theta^2-4\mu^2)(4-\delta^2\lambda^2\theta^2) \tag{185}$$

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

$$f(\bar{\delta}) = \Big(\lambda^4\theta^4 - 4\lambda^2\theta^2\mu^2 + 4\pi^2\mu^4\Big)\bar{\delta}^2 \tag{186}$$

$$- \ 4\Big(\lambda^2\theta^2 + 2\pi^2\mu^2 - 4\mu^2\Big)\bar{\delta} + 4\pi^2 < 0. \tag{187}$$

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 $\lambda^4\theta^4\Big(1-\frac{1}{\pi^2}\Big)$ so it's clearly positive. This means that $f(\bar{\delta})$ is a convex function, with a minimum at $\bar{\delta}^*$

$$\bar{\delta}^* = \frac{2(\lambda^2\theta^2+2\pi^2\mu^2-4\mu^2)}{\lambda^4\theta^4-4\lambda^2\theta^2\mu^2+4\pi^2\mu^4} \tag{188}$$

$$f(\bar{\delta}) \geq f(\bar{\delta}^*) = \frac{4(\pi^2-1)(\lambda^2\theta^2-4\mu^2)^2}{\lambda^4\theta^4-4\lambda^2\theta^2\mu^2+4\pi^2\mu^4} > 0, \tag{189}$$

where the denominator is the same as the coefficient of $\bar{\delta}^2$ from Equation (187), so it must be positive. The inequalities (187) and (189) contradict

each other, and so $\Delta_{cr}^* \leq \frac{\pi}{2\omega_{cr}}$ for all parameters. Hence by Equation (179), $\Delta_{cr} > \Delta_{cr}^*$, which implies that c_1 must be positive.

Since c_1 is positive, the only way for R from Equation (171) to have a nonzero equilibrium point is for c_2 to be also positive. This produces the conditions on the direction of the Hopf

$$\delta\mu < 1 \implies \alpha > 0 \tag{190}$$

$$\delta\mu > 1 \implies \alpha < 0. \tag{191}$$

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

$$R_0 = 0, \quad R_1 = \sqrt{\frac{c_2}{c_1}} > 0. \tag{192}$$

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

$$R_1 = \sqrt{\frac{4\alpha(\lambda^2\theta^2 - 4\mu^2)(4 - \delta^2\lambda^2\theta^2)^2}{\theta^2(1 - \delta^2\mu^2)(16\mu + \lambda^2\theta^2(4\Delta_{cr} - 4\delta + \delta^3\lambda^2\theta^2 - 4\delta^2\mu - 4\delta^2\Delta_{cr}\mu^2))}} \tag{193}$$

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

Theorem 4.14 establishes that as Δ increases, the equilibrium becomes unstable and a stable limit cycle is born. We also observe from Equation

(193) that the amplitude of the oscillations depends heavily on the model parameters. 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.

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 4.15. *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 Equation* (193)*.*

Proof. The radius of the limit cycle from (193) approximates the amplitude of the oscillations of $w_2(t)$ from (335). By the change of variables given in Equation (158), as $t \to \infty$, the behavior of the queues up to a phase shift is

$$q_1 = \frac{1}{2}(w_1 + w_2) \to \frac{1}{2}R_1 \sin(\omega \Delta t) \tag{194}$$

$$q_2 = \frac{1}{2}(w_1 - w_2) \to -\frac{1}{2}R_1 \sin(\omega \Delta t). \tag{195}$$

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

Therefore, when Δ exceeds Δ_{cr} and $\Delta - \Delta_{cr}$ is small, the amplitude of

oscillations can be approximated to first order by

$$\text{first order amplitude approximation} = \sqrt{\frac{(\Delta-\Delta_{cr})(\lambda^2\theta^2-4\mu^2)(4-\delta^2\lambda^2\theta^2)^2}{\theta^2(1-\delta^2\mu^2)(16\mu+\lambda^2\theta^2(4\Delta_{cr}-4\delta+\delta^3\lambda^2\theta^2-4\delta^2\mu-4\delta^2\Delta_{cr}\mu^2))}}. \quad (196)$$

The approximation is accurate when δ is substantially smaller than the ratio $\frac{2}{\lambda\theta}$. For example, in Figure 49 the queues oscillate throughout time, and the two horizontal lines provide a good approximation of the amplitude of oscillations based on Equation (196). However, the approximation becomes inaccurate when δ approaches $\frac{2}{\lambda\theta}$. As demonstrated in Figure 50, 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 Figures 51 - 54 as well. The surface plot in Figure 51 shows the true amplitude based on numerical integration as a function of the delay Δ and the coefficient δ, while the surface plot in Figure 52 shows the amplitude's first-order approximation. Furthermore, the surface plot in Figure 53 shows the error of first-order approximation, where the error increases with δ. Finally, Figure 54 provides intuition for why the approximation fails as δ approaches $\frac{2}{\lambda\theta}$. Figure 54 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 $\sqrt{\Delta - \Delta_{cr}}$, while the true amplitude appears to be a linear function of $(\Delta - \Delta_{cr})$ (even though it is not exactly linear).

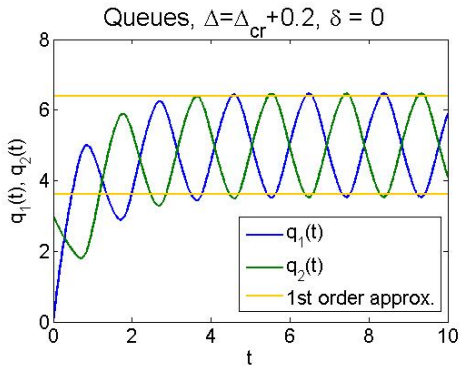

Figure 49: Amplitude approximation, $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{cr} + 0.2$, $\delta = 0$.

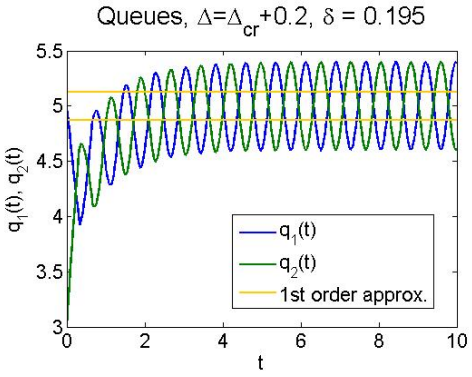

Figure 50: Amplitude approximation, $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{cr} + 0.2$, $\delta = 0.195$.

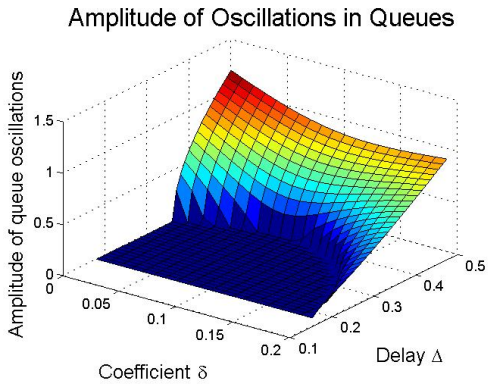

Figure 51: Amplitude of oscillations; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

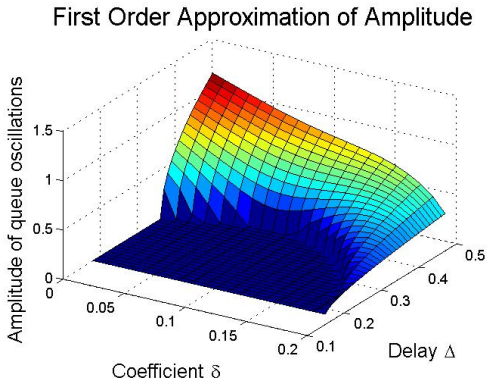

Figure 52: First-order approximation; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

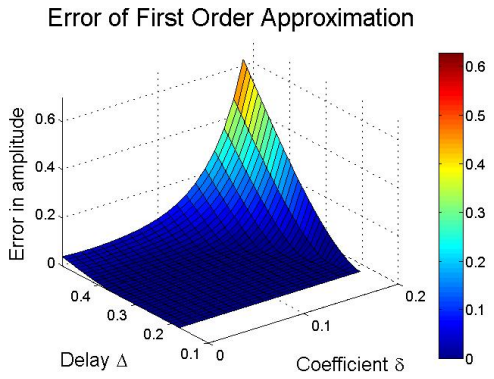

Figure 53: Error of approximation; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

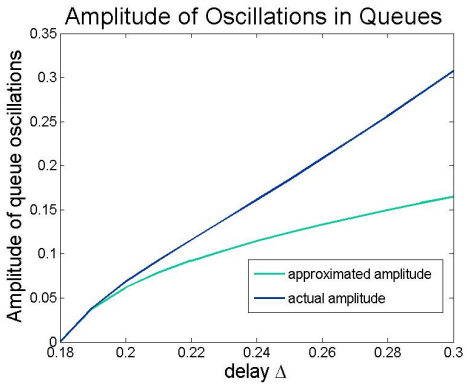

Figure 54: First-order approximation when $\delta = 0.19$, $\theta = 1$, $\lambda = 10$, $\mu = 1$.

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 Figure 52, 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 of the approximation.

Second-Order Approximation of Amplitude

The first-order approximation, as seen from Equation (196), is of the form

$$\text{Amplitude} \approx c_0(\Delta - \Delta_{cr})^{0.5}, \tag{197}$$

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

$$\text{Amplitude} \approx c_0(\Delta - \Delta_{cr})^{0.5} + c_1(\Delta - \Delta_{cr})^{1.5}, \tag{198}$$

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 the Appendix 7.3 for the expressions as well as a discussion on how the second-order approximation is obtained. As shown in Figures 55 and 56, the second order approximation performs just as well as the first order approximation when δ is significantly smaller than $\frac{2}{\lambda\theta}$, but is much more accurate when δ approaches $\frac{2}{\lambda\theta}$. Figures 57 and 58 confirm that this trend holds throughout the parameter space in δ and delay Δ. Figure 59 compares the true amplitude with the two approximations when $\delta = 0.1$. The next plot in Figure 60 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 61 and 62 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.

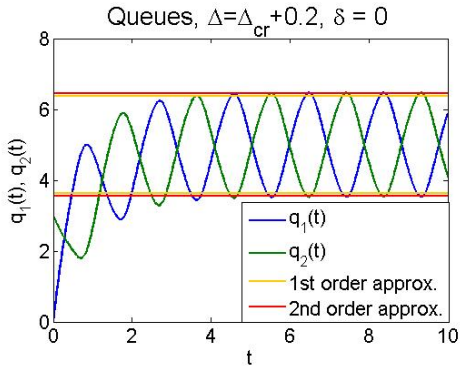

Figure 55: Amplitude approximation, $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{cr} + 0.2$, $\delta = 0$.

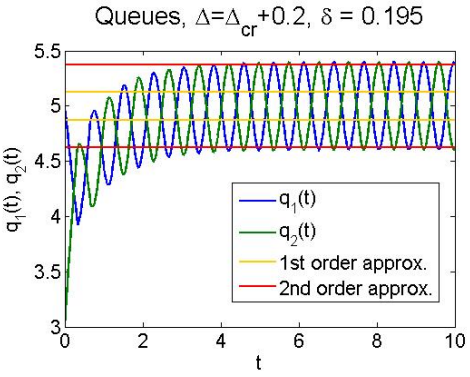

Figure 56: Amplitude approximation, $\frac{N}{\lambda\theta} = 0.2$, $\Delta = \Delta_{cr} + 0.2$, $\delta = 0.195$.

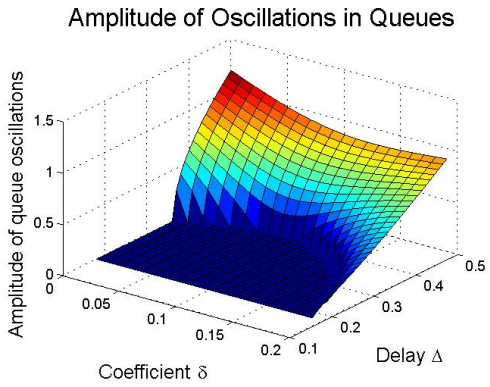

Figure 57: Amplitude of oscillations; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

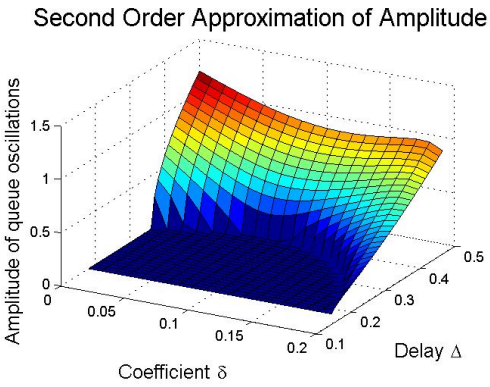

Figure 58: First-order approximation; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

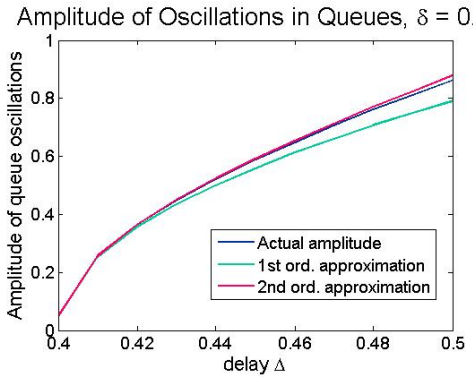

Figure 59: Comparison when $\delta = 0.10$; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

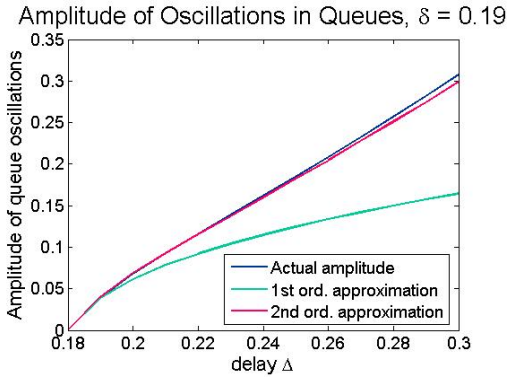

Figure 60: Comparison when $\delta = 0.19$; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

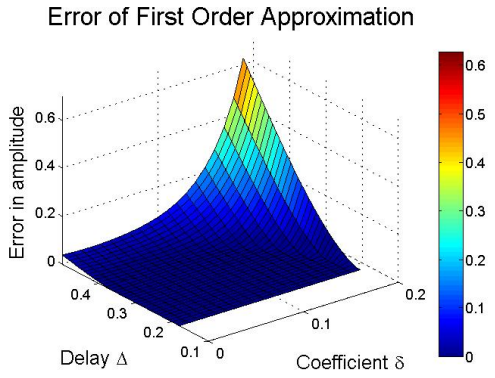

Figure 61: First-order error; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

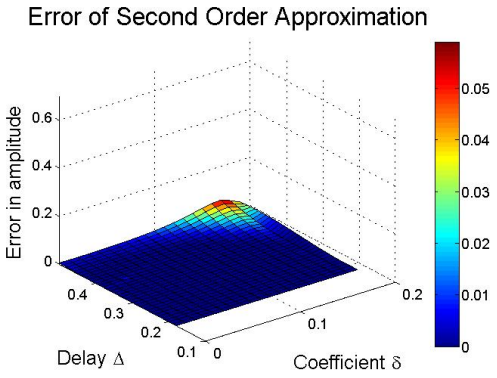

Figure 62: Second-order error; $\theta = 1$, $\lambda = 10$, $\mu = 1$.

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 63 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.

Figure 63 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 63 also reveals an important finding. **The value δ_{max} for the coefficient δ that maximizes Δ_{cr} is not the same as δ_{amp} that minimizes the amplitude of oscillations**. Specifically, δ_{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, $\delta_{max} = \delta_{amp}$, is when δ_{amp} is computed for the delay eqal to the maximum possible Δ_{cr}, i.e. $\Delta = \Delta_{cr}(\delta_{max})$. Therefore, one should use δ_{max} as the weight coefficient as long as the delay is less than the bifurcation threshold Δ_{cr} evaluated at δ_{max}, but when the delay exceeds Δ_{cr} one should use δ_{amp} for the weight coefficient instead. Thus, 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 $\delta = \delta_{amp}$.

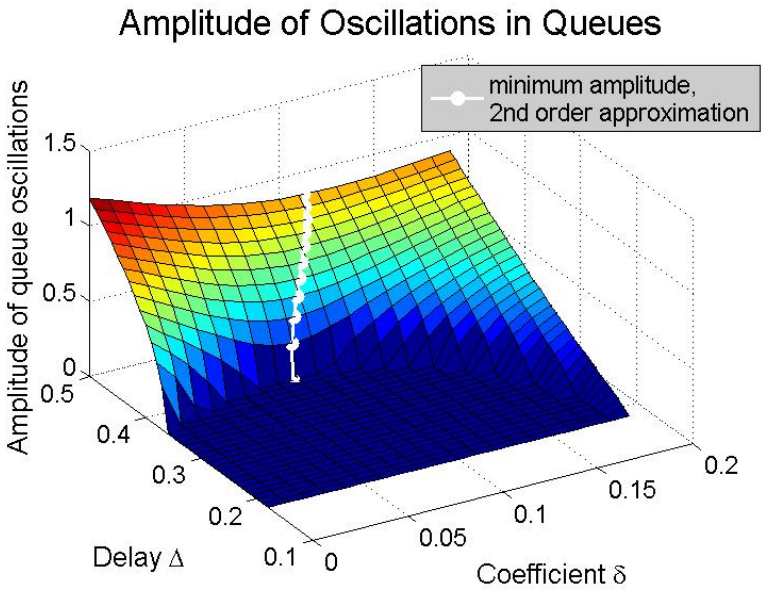

Figure 63: For any delay, the amplitude can be minimized as a function of δ.

4.6 Conclusion

This chapter answers important questions with regards to businesses incorporating the queue length velocity into the 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 that queue is moving, or the queue velocity,

$$\text{Delay announcement about } i^{th} \text{ queue} = q_i(t-\Delta) + \delta \dot{q}_i(t-\Delta), \quad (199)$$

with the delay Δ being the time of customers travelling 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 $\Delta = \Delta_{cr}$ and becoming unstable. Alternatively, if $\delta > \frac{N}{\lambda\theta}$, then the queues will never be stable even when the delay in information is infinitesimally small. The reader can refer to Figure 34 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 $\delta \leq \delta_{cap}$, 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 become unstable 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 is is clear that $\delta > \delta_{cap}$, then the queues bifurcate almost immediately because $\Delta_{cr} \to 0$ 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}.

Further, there exists an optimal value for δ called δ_{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 δ_{max} can be found numerically, as well as closed form expressions for upper and lower bounds on δ_{max}. Choosing δ within those bounds is a safe choice for the service managers.

This leads to an assessment of the limitations of providing the queue velocity information. The threshold Δ_{cr} where the queues become unstable 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 δ_{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 when unstable. From numerical integration of the queues such as in Figure 51, 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 that this coefficient as a function of delay, and is not

necessarily equal to the coefficient δ_{max} that maximizes the delay threshold.

In the future, this model can be extended 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 this data

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

The equations describing such a queueing system will no longer be neutral, and may be more complicated. However, such 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.

5 Queues With Randomly Delayed Information

The queueing models in the previous chapters make strict assumptions about the customers' commute to the queues, which in turn dictates the delay in information. The constant delay model assumes all customers have exactly the same travel time, while the model from Chapter 3 assumes a uniform distribution of travel times. The question arises: what can be said about the behavior of the queueing system in the general case when nothing is assumed about the delay distribution?

To answer this question, this chapter formulates a novel mathematical model where the individual's delay due to travel is represented as a random variable drawn from a fixed, but not necessarily known, distribution. We study the asymptotic behavior of this generalized queueing system, and prove that there exists a unique equilibrium state. Regardless of the delay distribution, the equilibrium is guaranteed to be locally stable when a certain relationship between the system's parameters is met ($\lambda\theta < N\mu$). The equilibrium can become unstable if and only if that parameter relationship is broken and a Hopf bifurcation occurs.

Whether or not a Hopf bifurcation occurs, and where it occurs on the parameter space, depends on the delay distribution. We derive the characteristic equation for the generalized queueing model and use it to determine the stability of the queues for several common distributions of the delay.

Constant delay, multiple discrete delays, continuous bounded, and continuous unbounded distributions are considered.

In the case where the delay distribution might be unknown or difficult to calculate, we develop a novel data-driven approximation technique. It uses Taylor expansions of the Laplace transform to determine stability based on the central moments of the delay distribution. Central moments can have the advantage of being easy to estimate by data from service systems through sampling the travel times of the incoming customers. The accuracy of this method of approximation is demonstrated on several numerical examples.

5.1 Chapter Outline

The generalized queueing model is presented in Section 5.2, and its asymptotic behavior is studied. We show that, independent of the delay distribution, there exists a unique equilibrium that is guaranteed to be stable if a certain parameter relationship is met. Further, the equilibrium can become unstable only if a Hopf bifurcation occurs. Unsurprisingly, whether or not a Hopf bifurcation occurs depends on the system's parameters as well as on the distribution of the delay. Section 5.4 proceeds to consider specific common distributions for the delay, and for each distribution we map out the stability region in the model's parameter space. The generalized model allows us to look into queueing systems with a constant delay, multiple discrete delays, infinitely many delays, as well as continuously distributed delays. Lastly, we consider in Section 5.5 the case where the delay distribution is unknown to

the service manager. By using information that can be gathered by sampling the customers, such as the average delay and the central moments of the delay distribution, we propose a technique that approximately determines the stability region for a queueing system with unknown delay distribution.

5.2 The Queueing Model

Figure 64 shows the structure of the queueing system. Customers appear at a constant arrival rate $\lambda > 0$ to a system of N queues. Each customer selects one queue to join, and upon arrival receives service at a rate $\mu > 0$. The model assumes that the rate of departure is a linear function of the queue length. This is equivalent to an infinite server queue which are quite important in the operations research and applied probability literature [13, 28, 12, 15, 24, 38]. With these assumptions, the queue lengths can be described by a system of N functional differential equations

$$\dot{q}_i(t) = \lambda p_i(q_1, \ldots, q_N) - \mu q_i(t), \quad \forall i \in \{1, ..., N\}, \tag{200}$$

where p_i is the proportion of customers that will join the i^{th} queue.

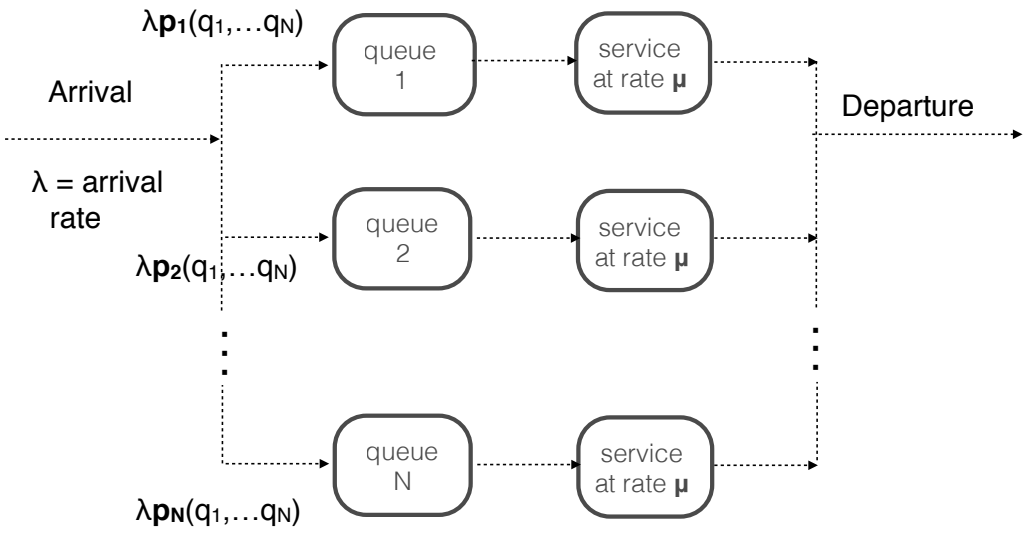

Figure 64: Customers going through a N-queue service system.

All queues offer identical service, but the queue length reported may differ depending on the number of customers in the queue. Being informed of the current length of each queue, the customer decides which queue he or she is going to join and gives higher preference to the shorter queue according to the Multinomial Logit Model (MNL). A customer who appeared at time $t - X$, chose a queue, and then travelled for the time X, at time t has the following probability of joining the i^{th} queue:

$$\text{probability of joining } i^{th} \text{ queue} = \frac{\exp(-\theta q_i(t-X))}{\sum_{j=1}^{N} \exp(-\theta q_j(t-X))}. \tag{201}$$

As desired, the probability that he joins any of the N queues sums up to 1. The shorter a given queue is at time $t - X$, the more likely it is going to be

chosen by the customer.

When customers' individual commute time is a random variable X drawn from a distribution with probability density function $f(X = s)$, the proportion of all customers who arrive to the i^{th} queue at time t is

$$\text{proportion joining } i^{th} \text{ queue}, p_i = \frac{\exp\left(-\theta \int_0^\infty q_i(t-s)f(s)ds\right)}{\sum_{j=1}^N \exp\left(-\theta \int_0^\infty q_j(t-s)f(s)ds\right)}. \quad (202)$$

The queueing system (200) can therefore be described as

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\left(-\theta \int_0^\infty q_i(t-s)f(s)ds\right)}{\sum_{j=1}^N \exp\left(-\theta \int_0^\infty q_j(t-s)f(s)ds\right)} - \mu q_i(t), \quad \forall i \in \{1, ..., N\} \quad (203)$$

5.3 Stability

When we look at the generalized queueing system from Equation (203), a couple of things can be said about its stability prior to considering specific probability distribution functions f. Theorem 5.1 shows that regardless of the delay distribution, there is a unique equilibrium state. For convenience, we will reformulate Equation (203) as

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta g(q_i(t))\big)}{\sum_{j=1}^N \exp\big(-\theta g(q_j(t))\big)} - \mu q_i(t), \quad (204)$$

where $g(x)$ is a monotonically increasing function.

Theorem 5.1. *The unique equilibrium to the system of equations* (204) *where* $g(x)$ *is a monotonically increasing function is given by* $q_i(t) = q_i^* = \frac{\lambda}{N\mu}$.

Proof. It is easy to check that if $q_i^* = \frac{\lambda}{N\mu}$, then $\dot{q}_i(t) = 0$ for every i, so q_i^* is indeed an equilibrium. The uniqueness of the equilibrium can be verified by contradiction. We will suppose that there is another distinct equilibrium state, $\bar{q}_i$ for $i = 1, \ldots, N$. We note that $\sum_{i=1}^{N} \dot{\bar{q}}_i = 0 = \lambda - \mu \sum_{i=1}^{N} \bar{q}_i$, so $\sum_{i=1}^{N} \bar{q}_i = \frac{\lambda}{N} = \sum_{i=1}^{N} q_i^*$. Since the equilibrium state $\bar{q}_i$ is distinct from q_i^* then $\bar{q}_i$ cannot all be $\frac{\lambda}{N\mu}$, and therefore for some indices k and l,k $1 \leq k, l \leq N$, the inequalities below must hold

$$\bar{q}_k < \frac{\lambda}{N\mu}, \quad \text{and} \quad \bar{q}_l > \frac{\lambda}{N\mu}. \tag{205}$$

Because the function g is monotonically increasing, $-g(\bar{q}_k) > -g(\bar{q}_l)$ and

$$\frac{\exp\big(-\theta g(\bar{q}_l)\big)}{\exp\big(-\theta g(\bar{q}_k)\big)} < 1. \tag{206}$$

Further, $\dot{\bar{q}}_k(t) = 0$ so

$$\dot{\bar{q}}_k(t) = \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_k)\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big)} - \mu \bar{q}_k(t) = 0 \tag{207}$$

$$\lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_k)\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big)} = \mu \bar{q}_k(t) < \frac{\lambda}{N} \tag{208}$$

$$\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big) > N \exp\big(-\theta g(\bar{q}_k)\big). \tag{209}$$

Finally, we use inequalities (206) and (209) to show that $\dot{\bar{q}}_l(t) \neq 0$:

$$\dot{\bar{q}}_l(t) = \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_l)\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big)} - \mu \bar{q}_l(t) \tag{210}$$

$$< \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_l)\big)}{N \exp\big(-\theta g(\bar{q}_k)\big)} - \frac{\lambda}{N} < \frac{\lambda}{N} - \frac{\lambda}{N} = 0. \tag{211}$$

Hence, $\bar{q}_i$ is not an equilibrium state, so the equilibrium is unique. □

To determine when the queueing system is locally stable, we need to find the characteristic equation.

Finding the characteristic equation. The system of delay differential equations can be expressed in the following manner

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta m_i(t)\big)}{\sum_{j=1}^{N} \exp\big(-\theta m_j(t)\big)} - \mu q_i(t) \tag{212}$$

$$m_i(t) = \int_0^\infty q_i(t-s) f(s) ds. \tag{213}$$

$$\dot{q}_i(t) \approx -\frac{\lambda\theta}{N} m_i(t) + \frac{\lambda\theta}{N^2} \sum_{j=1}^{N} m_j(t) - \mu q_i(t) \tag{214}$$

$$\dot{m}_i(t) = \frac{d}{dt} \int_0^\infty q_i(t-s) f(s) ds, \tag{215}$$

which can be expressed in a vector form

$$\dot{\bar{q}}(t) = -\frac{\lambda\theta}{N} \bar{m}(t) + \frac{\lambda\theta}{N^2} A\bar{m}(t) - \mu \bar{q}(t) \tag{216}$$

$$\dot{\bar{m}} = \frac{d}{dt} \int_0^\infty \bar{q}(t-s) f(s) ds, \tag{217}$$

where $\bar{q} = [q_1, \ldots, q_N]^T \in \mathbb{R}^N$, $\bar{m} = [m_1, \ldots, m_N]^T \in \mathbb{R}^N$, and $A \in \mathbb{R}^{N \times N}$ with $A_{ij} = 1$ for $1 \leq i, j \leq N$. The matrix can be diagonalized,

$$A = VDM, \quad \text{where } D, M, V \in \mathbb{R}^{N \times N} \quad \text{and } D_{ij} = \begin{cases} 1 & i = j = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{218}$$

Let $\bar{q}(t) = V\bar{w}(t)$ and $\bar{m}(t) = V\bar{u}(t)$. Note that since $VM = MV = I$, we are guaranteed that such vectors $\bar{w}$ and $\bar{u}$ exist. With this transformation of variables, the equations become

$$\begin{aligned} V\dot{\bar{w}}(t) &= -\frac{\lambda\theta}{N} V\bar{u}(t) + \frac{\lambda\theta}{N^2} VD\bar{u}(t) - \mu V\bar{w}(t) & (219) \\ V\dot{\bar{u}}(t) &= \frac{d}{dt}\int_0^\infty V\bar{w}(t-s)f(s)ds. & (220) \end{aligned}$$

Once the two equations are pre-multiplied by M, we find

$$\begin{aligned} \dot{\bar{w}}(t) &= -\frac{\lambda\theta}{N} \bar{u}(t) + \frac{\lambda\theta}{N^2} D\bar{u}(t) - \mu\bar{w}(t) & (221) \\ \dot{\bar{u}}(t) &= \frac{d}{dt}\int_0^\infty \bar{w}(t-s)f(s)ds. & (222) \end{aligned}$$

Only one element of the matrix D is nonzero, which simplifies our equations

$$\begin{aligned} \dot{w}_1(t) &= -\mu w_1(t) & (223) \\ \dot{w}_i(t) &= -\frac{\lambda\theta}{N} u_i(t) - \mu w_i(t), \quad i = 2, 3, \ldots, N & (224) \\ \dot{u}_i(t) &= \frac{d}{dt}\int_0^\infty w_i(t-s)f(s)ds, \quad i = 1, 2, \ldots, N & (225) \end{aligned}$$

All w_i functions are now uncoupled, and w_1 from Equation (223) has a solution that's always stable. To find the characteristic equation, we assume $w_i(t) = e^{rt}$. Then $u_i(t) = \int_0^\infty e^{r(t-s)} f(s)ds = e^{rt}F(r)$, where $F(r)$ is the Laplace transform of the delay distribution function $f(t)$. Equation (224) yields the characteristic equation

$$\Phi(r, \Delta) = r + \mu + \frac{\lambda\theta}{N} F(r) = 0. \tag{226}$$

When the arrival rate λ is sufficiently small or the service rate μ is sufficiently high, the queueing system is guaranteed to be stable. We formulate this result in the following theorem.

Theorem 5.2. *If $\lambda\theta < N\mu$, the equilibrium from Theorem 5.1 is locally stable.*

Proof. The equilibrium is locally stable if every eigenvalue r that satisfies the characteristic equation has a negative real part, i.e. $\mathrm{Re}[r] < 0$. Plug in the explicit formulation for the Laplace transform $F(r) = \int_0^\infty e^{-rs} f(s)ds$ into the characteristic equation, and rewrite the eigenvalue as $r = a + ib$ where $a, b \in \mathbb{R}$.

$$\Phi(r, \Delta) = a + ib + \mu + \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \big(\cos(bs) - i \sin(bs) \big) f(s)ds = 0. \tag{227}$$

After separating the real and imaginary parts, we arrive at two equations

$$a + \mu + \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \cos(bs) f(s) ds = 0 \tag{228}$$

$$b - \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \sin(bs) f(s) ds = 0. \tag{229}$$

To reach contradiction, let us suppose that $\lambda < N\mu/\theta$ and there exists $a \geq 0$ that satisfies Equations (228) - (229). Since $\int_0^\infty f(s)ds = 1$, then

$$\left| \int_0^\infty e^{-as} \cos(bs) f(s) ds \right| \leq \left| \int_0^\infty e^{-as} f(s) ds \right| \leq \left| \int_0^\infty f(s) ds \right| = 1. \tag{230}$$

From Equation (228) it then follows that

$$a = -\mu - \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \cos(bs) f(s) ds \tag{231}$$

$$\leq -\mu + \frac{\lambda\theta}{N} < -\mu + \frac{N\mu\theta}{N\theta} = 0, \tag{232}$$

which contradicts our assumption that a can be non-negative. It follows that when $\lambda\theta < N\mu$ the real part any eigenvalue satisfying the characteristic equation must be negative, and therefore the equilibrium from Theorem 5.1 is locally stable. □

When the system's parameters change and the relationship $\lambda\theta < N\mu$ is no longer true, the equilibrium may become unstable. The next result proves that the equilibrium can become locally unstable only if a pair of complex eigenvalues reaches the imaginary axis, since any real eigenvalues

are guaranteed to be negative regardless of the parameter values.

Lemma 5.3. *Any real eigenvalue of the characteristic equation* (226) *is negative.*

Proof. Suppose $r \in \mathbb{R}$. Then $F(r) = \int_0^\infty e^{-rs} f(s) ds \geq 0$, and from equation (226) it follows that r must be negative

$$r = -\mu - \frac{\lambda\theta}{N} F(r) \leq -\mu < 0. \tag{233}$$

□

Since the real-valued eigenvalues remain negative for all parameters λ, θ, μ, and N, the queueing system can become unstable only if a complex-valued eigenvalue has a positive real part. This can be caused by a Hopf bifurcation.

Theorem 5.4. *If one pair of eigenvalues is purely imaginary, a Hopf bifurcation occurs as λ or μ change.*

Proof. The infinite-dimensional version of the Hopf Theorem from Hale and Lunel [18] states that a Hopf bifurcation occurs with respect to a parameter x at $x = x^*$ when the following three conditions hold.

- When $x = x^*$, there must be a pair of purely imaginary eigenvalues r^+ and r^- that satisfy the characteristic equation.
- Any other eigenvalue $r \neq r^+, r^-$ is not an integer multiple of the imaginary eigenvalue, so $r \neq mr^+, mr^-$ for any $m \in \mathbb{Z}$.

- The derivative of the real part of the eigenvalue with respect to the bifurcation parameter at the point of bifurcation is non-zero, i.e. $\frac{d}{dx}$ Re $r^+(x^*) \neq 0$.

In the case of our queueing model, both λ and μ can be viewed as the bifurcation parameter x. Suppose that for some given value of λ^* and μ^*, one pair of eigenvalues is purely imaginary. Thus if we think of the eigenvalue r as a function of λ and μ, $r(\lambda, \mu) = a(\lambda, \mu) + ib(\lambda, \mu)$ where a and b are real-valued, then $r(\lambda^*, \mu^*) = ib(\lambda^*, \mu^*)$. The other imaginary eigenvalue must be its complex conjugate, $r(\lambda^*, \mu^*) = -ib(\lambda^*, \mu^*)$. Without loss of generality, assume that $b(\lambda^*, \mu^*)$ is positive. Equations (228) - (229) at (λ^*, μ^*) become

$$\mu^* + \frac{\lambda^*\theta}{N} \int_0^\infty \cos(bs) f(s) ds = 0 \tag{234}$$

$$b - \frac{\lambda^*\theta}{N} \int_0^\infty \sin(bs) f(s) ds = 0. \tag{235}$$

To show that the third condition of the Hopf Theorem is fulfilled, we differentiate Equation (228) with respect to λ:

$$\frac{da}{d\lambda} + \frac{\theta}{N} \int_0^\infty e^{-as} \cos(bs) f(s) ds + \frac{\lambda\theta}{N} \cdot \frac{d}{d\lambda} \int_0^\infty e^{-as} \cos(bs) f(s) ds = 0. \tag{236}$$

As long as the probability density function f is continuous everywhere besides countably many points, we can interchange the derivative and integral,

$$\frac{da}{d\lambda} + \frac{\theta}{N} \int_0^\infty e^{-as} \cos(bs) f(s) ds + \frac{\lambda\theta}{N} \int_0^\infty (-s) e^{-as} \cos(bs) f(s) ds \cdot \frac{da}{d\lambda} = 0 \tag{237}$$

When $(\lambda, \mu) = (\lambda^*, \mu^*)$ we recall that $a = 0$ and by Equation (234) we find $\int_0^\infty \cos(bs) f(s) ds = -\frac{N\mu^*}{\lambda^* \theta}$. Substituting this into (237)

$$\frac{d}{d\lambda} a(\lambda^*, \mu^*) - \frac{\mu^*}{\lambda^*} - \frac{\lambda^* \theta}{N} \int_0^\infty s \cos(bs) f(s) ds \cdot \frac{d}{d\lambda} a(\lambda^*, \mu^*) = 0 \qquad (238)$$

Since $\frac{\mu^*}{\lambda^*} \neq 0$, it is evident that

$$\frac{d}{d\lambda} \mathrm{Re}[r(\lambda^*, \mu^*)] = \frac{d}{d\lambda} a(\lambda^*, \mu^*) \neq 0, \qquad (239)$$

so the third condition of the Hopf Theorem holds with respect to the parameter λ.

Analogously, it can be shown that $\frac{d}{d\mu} \mathrm{Re}[r(\lambda^*, \mu^*)] \neq 0$ as well. □

Based on our previous results, we can conclude that the smallest Hopf curve determines the stability of the queueing system.

Theorem 5.5. *The equilibrium becomes locally unstable at the smallest first Hopf bifurcation as λ increases or μ decreases.*

Proof. Suppose that for a given delay distribution, a Hopf bifurcation may occur. When $\lambda\theta < N\mu$, all eigenvalues have a negative real part by Theorem 5.2. Further, Theorem 5.4 shows that when a pair of complex eigenvalues becomes purely imaginary, the derivatives of the real part of these eigenvalues with respect to λ and μ are nonzero, so they must cross the imaginary axis. Thus, the first pair of eigenvalues that reaches the imaginary axis on the complex plane must cross the axis from negative to positive real side. Once

there is a pair of eigenvalues with positive real parts, the equilibrium must be locally unstable. □

For a fixed delay distribution, as the arrival rate of customers changes, the queueing system can qualitatively change its behavior. With a sufficiently small arrival rate ($\lambda\theta < N\mu$), the queue will stabilize over time and all queues will converge to the same length of $\frac{\lambda}{N\mu}$. However, as the arrival rate will increase such that $\lambda\theta \geq N\mu$, the queueing system may undergo a Hopf bifurcation, causing the queues to oscillate throughout time, never reaching an equilibrium state.

Whether or not a Hopf bifurcation occurs and the equilibrium becomes unstable depends on the distribution of the delay. In the next section, we will consider several common distributions and discuss how the queueing system behaves in each case.

5.4 Common Delay Distributions

Queueing systems may exhibit different behavior depending on the distribution of the delay from the customers' commute time. Based on the results from Section 5.3, however, any queueing system of the form given by Equation (203) is guaranteed to be stable unless a Hopf bifurcation occurs. Our goal for this section is to consider some common distributions for the time delay, and see under what conditions (if ever) the resulting queueing systems become unstable. We will begin by reviewing the simplest possible distri-

bution, and then will gradually introduce more complexity to get a better understanding of how the distribution of the delay affects the underlying dynamics of the queueing system.

5.4.1 Constant Delay

For our first model, we will suppose that there is no variation in the commute time of the customers. They all travel from the same population center to the queue of their choice, where the queues are equidistant from the customers' initial location. This is the constant delay model from Chapter 2. Figure 65 provides a geographical representation of such a queueing system.

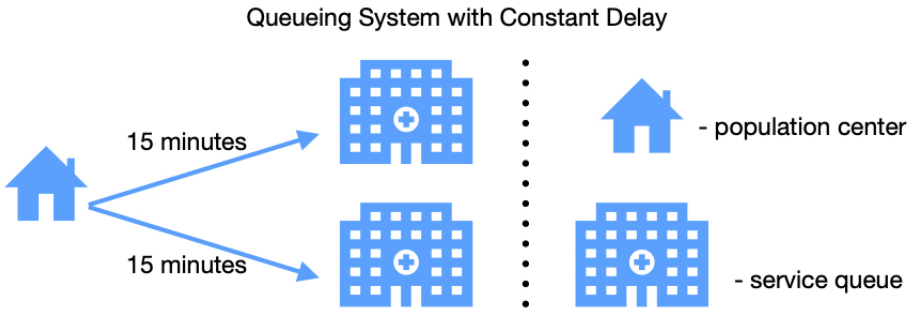

Figure 65: All customers have the same commute time.

The system of differential equations from (203) simplifies to

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta q_i(t-\Delta)\big)}{\sum_{j=1}^{N} \exp\big(-\theta q_j(t-\Delta)\big)} - \mu q_i(t) \qquad \forall i = 1, 2, \ldots, N, \quad (240)$$

where the parameter $\Delta > 0$ represents the time delay that the customers experience. The delay distribution is a Dirac delta function $f(s) = \delta(s - \Delta)$

with the Laplace transform $F(r) = \exp(-r\Delta)$. Therefore the characteristic equation (226) becomes

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N}\exp(-r\Delta) = 0. \tag{241}$$

The queueing system (240) undergoes a Hopf bifurcation, and the point of bifurcation can be found analytically. We set the eigenvalue to be purely imaginary, $r = ib$, and separate the real and imaginary parts of the characteristic equation to find

$$\mu + \frac{\lambda\theta}{N}\cos(b\Delta) = 0, \quad b - \frac{\lambda\theta}{N}\sin(b\Delta) = 0. \tag{242}$$

The trigonometric identity $\sin^2(b\Delta) + \cos^2(b\Delta) = 1$ gives expressions for b and Δ

$$b = \sqrt{\frac{\lambda^2\theta^2}{N^2} - \mu^2}, \qquad \Delta = \frac{N\arccos(-N\mu/\ (\lambda\theta))}{\sqrt{\lambda^2\theta^2 - N^2\mu^2}}. \tag{243}$$

By Theorem 5.4, the queues undergo a Hopf bifurcation when the equations (243) hold. In the parameter space λ and Δ, the Hopf curve is given by Figure 66. The queues are stable in the region to the left of the Hopf curve, and unstable in the region to the right.

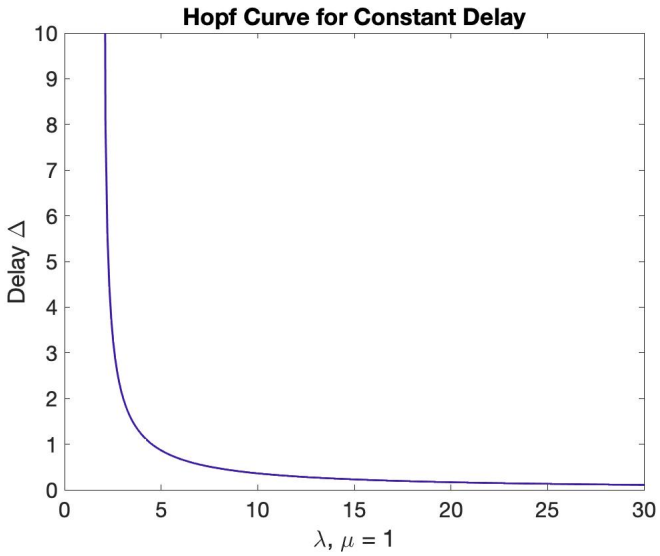

Figure 66: Hopf curve given a constant delay.

5.4.2 Real-Time and Delay

Let us now slightly generalize the constant delay model, so that not all customers have the same commute. Suppose that some customers are located at a population center that requires a fixed commute just like in the constant delay model, but now there is another group of customers that are near the service and don't require any commute. Figure 67 gives a geographical picture for the resulting queueing system. There are two population centers present, (i) and (ii), where the customers from (i) experience a delay Δ while the customers from (ii) have no delay. We will refer to this set-up as a *real-time and delay* model.

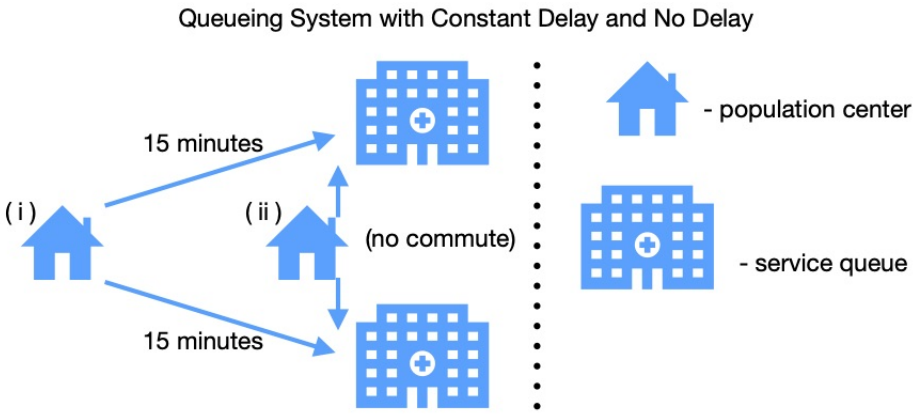

Figure 67: Customers experience either a constant delay or no delay, depending on their initial location.

When a proportion p of customers ($0 \leq p \leq 1$) have delay Δ and the rest of the population have no delay, the queueing system can be represented as

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta(1-p)q_i(t) - \theta p q_i(t-\Delta)\big)}{\sum_{j=1}^{N} \exp\big(-\theta(1-p)q_j(t) - \theta p q_j(t-\Delta)\big)} - \mu q_i(t), \quad i = 1, 2, \ldots, N.$$

Here the delay distribution is a linear combination of Dirac delta functions, $f(s) = p\delta(s-\Delta) + (1-p)\delta(s)$, which has the Laplace transform of $F(r) = p\exp(-r\Delta) + (1-p)$. The characteristic equation becomes

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N}\exp(-r\Delta) + \frac{\lambda\theta}{N} \cdot (1-p) = 0. \tag{244}$$

To determine when the queueing system becomes unstable, we solve the characteristic equation assuming a purely imaginary eigenvalue $r = ib$, $0 < b \in \mathbb{R}$. Using the same technique as for the constant delay model, we find a

closed-form expression for b and the delay Δ where a bifurcation occurs in the following Theorem.

Theorem 5.6. *The real-time and delay model has the following expression for the Hopf bifurcation critical delay value:*

$$\Delta_{cr} = \frac{N\mu \arccos\Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)}{\lambda\theta p\sqrt{1 - \Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)^2}}. \tag{245}$$

Moreover, the model is always stable unless

$$p \;>\; \frac{1}{2} + \frac{N\mu}{2\lambda\theta}. \tag{246}$$

$$\Delta = \frac{\arccos\Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)}{b}, \quad b = \frac{\lambda\theta p}{N\mu} \cdot \sqrt{1 - \Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)^2}. \tag{247}$$

The dynamics of this model are more interesting than in the constant delay case. Specifically, the queues may remain stable for all values of λ and μ, and that the Hopf bifurcation will never occur. To see this, notice that denominator from Equation (245) must be real and positive, which imposes

the condition i.e.

$$\frac{\lambda\theta p}{N\mu} \cdot \sqrt{1 - \Big(- \frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)^2} > 0 \tag{248}$$

$$\Big(\frac{N\mu}{\lambda\theta p} + \frac{1}{p}\Big)\Big(2 - \frac{N\mu}{\lambda\theta p} - \frac{1}{p}\Big) > 0 \tag{249}$$

$$p > \frac{N\mu + \lambda\theta}{2\lambda\theta} = \frac{1}{2} + \frac{N\mu}{2\lambda\theta}. \tag{250}$$

This implies that if at least half of the customers experience no delay, then the queueing system will remain stable despite all other factors. The queueing system, however, will become unstable if the condition (250) holds, meaning that a significant proportion of the population experienced a delay.

Figure 68 shows the Hopf curves where different proportions of the customers experience a delay. When the delay affects all customers ($p = 1$), the Hopf curve becomes exactly constant delay Hopf curve from Figure 65. As the proportion of the delayed customers shrinks (p decreases), the Hopf curve moves to the right, increasing the region of the parameter space where the queueing system is stable.

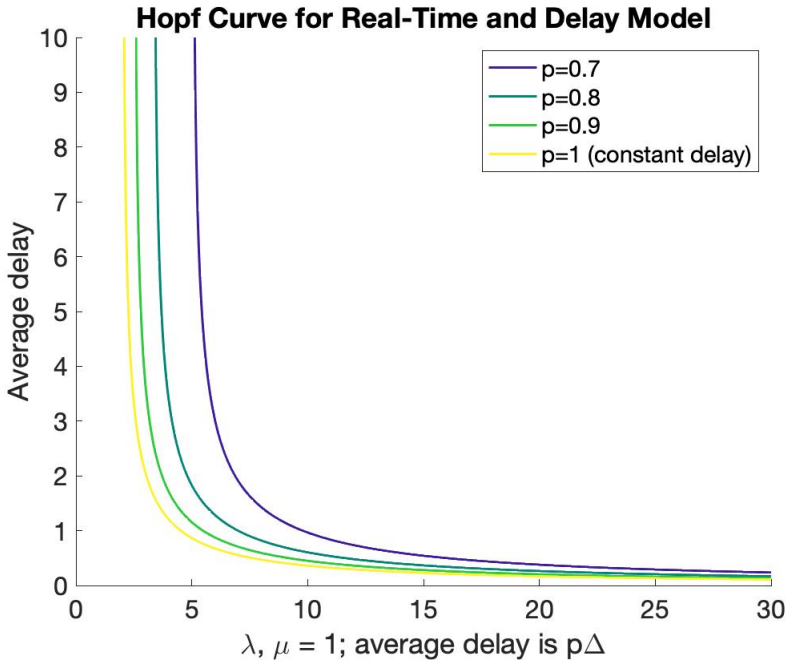

Figure 68: Hopf curve for different values of p.

5.4.3 M Discrete Delays

To further generalize the delay distribution, we will consider a queueing system where M different delays are present, and each delay is experienced by some proportion of the incoming customers. Thus, we analyze a finite discrete distribution as the distribution of travel times. We know that the discrete distribution has the following pmf and is specified by the probabilities $p_1, p_2, ..., p_m$ and values $a_1, a_2, ..., a_m$, and is specified by

$$\mathbb{P}(X = a_k) = p_k, \tag{251}$$

with the Laplace transform

$$F(r) = \sum_{k=1}^{m} p_k e^{-ra_k}. \tag{252}$$

The eigenvalue equation then follows directly from (226),

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N} \sum_{k=1}^{m} p_k e^{-ra_k} = 0. \tag{253}$$

Although this representation of the eigenvalue equation does not give us any information on how to solve it at first. It does show us that the multiple delays is equivalent to the assuming the delay is random and follows a discrete distribution. Thus, if we can develop approximations for random delays, then it will immediately give us a way of understanding the properties of ddes with multiple constant delays.

Now that the delay distribution is more complicated, we cannot extract useful closed-form expressions to determine where the system of queues becomes unstable. However, (253) can be solved numerically. We will set $r = ib$ in the characteristic equation and separate the real and imaginary parts. The resulting system of two equations can be solved for the unknown b and λ, given that the other parameters are known. Below in Figure 69 we give an example of a Hopf curve resulting from a system with three delays. A proportion $p_1 = 1/6$ of the customers experience delay of $\Delta_1 = 0.8$, and $p_2 = 1/3$ of the customers has a delay of $\Delta_2 = 0.9$. The plot shows how large the final delay must be in order for the queueing system to become unstable. On the

left plot $\mu = 0.5$ and on the right $\mu = 1$.

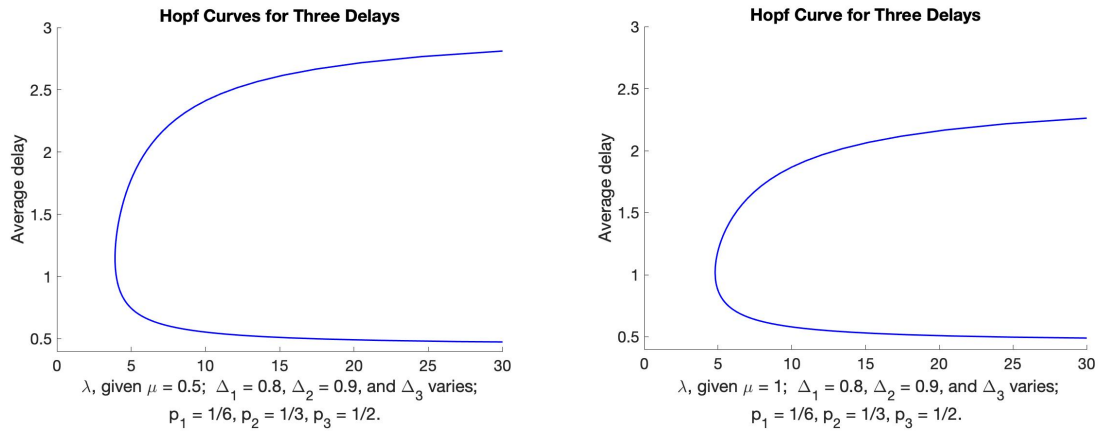

Figure 69: Hopf curve for three discrete delays.

A curious phenomenon to observe is that for a fixed arrival rate, the queueing system can go from stable to unstable and then back to stable again as the average delay increases. We do not have a good intuition for why this is the case, but such behavior is common for more complicated delay distributions.

5.4.4 Discrete Uniform On Bounded Interval

With M discrete delays, one particular distribution that one can analyze with our methodology is the case when all of the delays are equidistant and have equal probability of occurring. For example we let X be the following random variable where each outcome has the same probability

$$X = \frac{2\Delta k}{M} \quad \text{withprobability } \frac{1}{M+1} \quad \text{where } k \in \{0, 1, 2, ..., M\}.$$

Lemma 5.7. *The Laplace transform for the random variable X above is*

$$F_X^M(r) = \mathbb{E}[e^{-rX}] = \frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})}. \tag{254}$$

Proof.

$$\begin{aligned} F_X^M(r) &= \mathbb{E}[e^{-rX}] && (255)\\ &= \frac{1}{M+1}\sum_{k=0}^{M} e^{\frac{-2r\Delta k}{M}} && (256)\\ &= \frac{1}{M+1}\sum_{k=0}^{M} \left(e^{\frac{-2r\Delta}{M}}\right)^k && (257)\\ &= \frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})}, \quad \text{by the truncated geometric sum.} \end{aligned}$$

□

By Lemma 5.7, the characteristic equation becomes

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N}\cdot\frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})} = 0. \tag{258}$$

As the number of delays goes to infinity, $M \to \infty$, the Laplace transform of the discrete uniform distribution converges to the Laplace transform of a continuous uniform distribution. It follows from Lemma 5.8.

Lemma 5.8. *As the number of points M tends to infinity, the Laplace transform for the random variable X has the following expression*

$$\lim_{M\to\infty} F_X^M(r) = \frac{1-e^{-2r\Delta}}{2r\Delta} \tag{259}$$

Proof.

$$\begin{aligned} \lim_{M\to\infty} F_X^M(r) &= \lim_{M\to\infty} \frac{1-e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1-e^{-\frac{2r\Delta}{M}})} && (260)\\ &= \frac{1-e^{-2r\Delta}}{2r\Delta}. && (261) \end{aligned}$$

□

Therefore, the characteristic equation of the discrete uniform delays converges to the characteristic equation based on a uniformly distributed delay, which we will consider later on in Section 5.4.5. Figure 70 shows the convergence as $M \to \infty$ of the Hopf curves from Equation (258) to the uniform distribution Hopf curve.

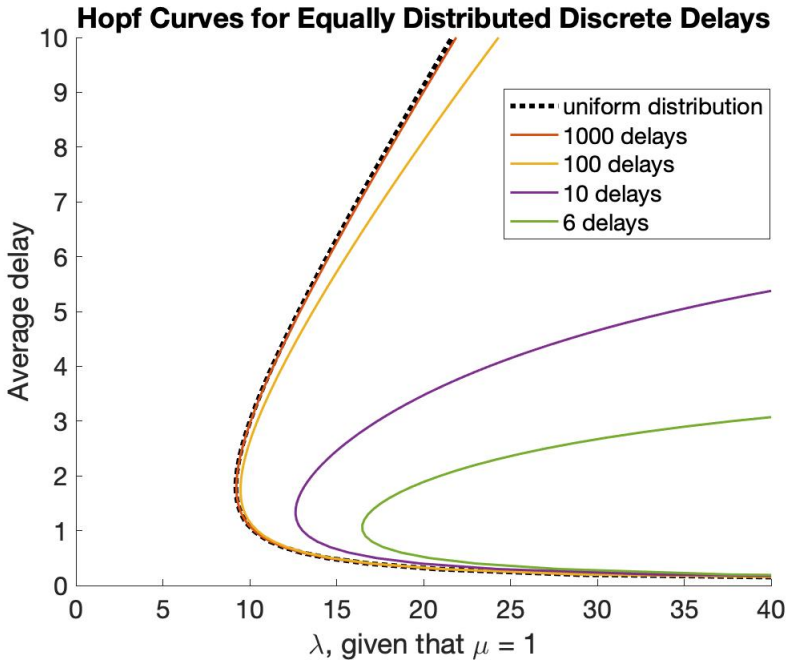

Figure 70: Discrete delay system Hopf curves converge to the uniformly distributed delay Hopf curve. The latter yields the least stable system.

5.4.5 Uniform Distribution

When M discrete delays are weighed equally and M increases, the delay distribution begins to resemble a continuous distribution. This serves as a motivation to study a queueing system with uniformly distributed delay on an interval $[0, 2\Delta]$, $\Delta > 0$. Figure 71 shows a queueing system where customers' commute to the queues may be approximated by a uniform distribution.

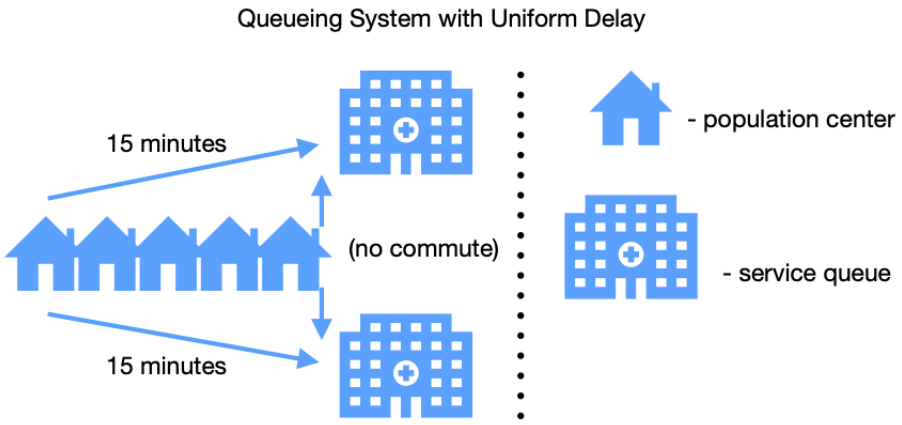

Figure 71: Customers going through a N-queue service system.

The probability density function is $f(s) = \begin{cases} \frac{1}{2\Delta} & 0 \leq s \leq 2\Delta \\ 0 & \text{otherwise} \end{cases}$, with the Laplace transform given by

$$F(r) = \int_0^{2\Delta} e^{-rt} \cdot \frac{1}{2\Delta} dt = \frac{1}{2\Delta r}\left(1 - e^{-2r\Delta}\right). \tag{262}$$

The characteristic equation (226) for the queueing system is therefore

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N \cdot 2\Delta r} - \frac{\lambda\theta}{N \cdot 2\Delta r} \cdot e^{-2r\Delta} = 0. \tag{263}$$

We can determine when a Hopf bifurcation occurs by solving for a purely imaginary eigenvalue, that is $r = ib$. By separating the real and imaginary parts of the characteristic equation we find expressions for sine and cosine

$$\sin(2b\Delta) = -\frac{4\Delta\mu b}{\lambda\theta}, \quad \cos(2b\Delta) = 1 - \frac{4\Delta b^2}{\lambda\theta}. \tag{264}$$

The identity $\sin^2(2b\Delta) + \cos^2(2b\Delta) = 1$ produces a closed-form expression

$$b = \sqrt{\frac{\lambda\theta}{2\Delta} - \mu^2}. \tag{265}$$

When b is substituted into Equation (264), we get a transcendental equation,

$$\sin\left(2\Delta \cdot \sqrt{\frac{\lambda\theta}{2\Delta} - \mu^2}\right) + \frac{4\mu\Delta}{\lambda\theta} \cdot \sqrt{\frac{\lambda\theta}{2\Delta} - \mu^2} = 0. \tag{266}$$

When Equation (266) is solved numerically for Δ, the Hopf curve from Figure 72 can be found. As with the three discrete delays from Figure 69, the Hopf curve is not necessarily uniquely determined for a fixed customer arrival rate λ. As seen from plot on the right in Figure 72, when $\lambda = 15$, the queueing system is stable when the average delay is $\Delta < 0.5$ or $\Delta > 7$, and unstable roughly when the average delay is in the range $[1, 6]$.

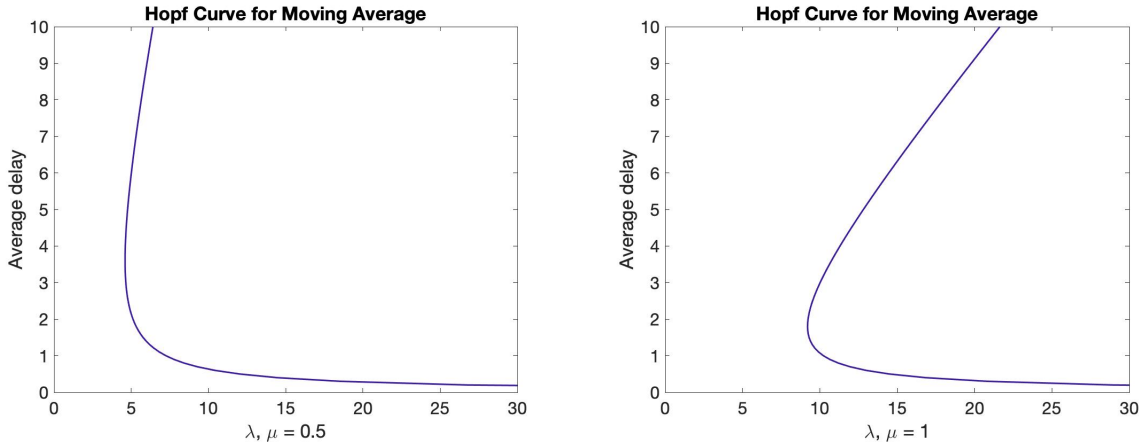

Figure 72: Hopf curve for uniformly distributed delay on $[0, 2\Delta]$. The average delay is Δ.

Uniform distribution does not have to be restricted to the interval $[0, 2\Delta]$. A generalized case can be considered, with the delay being distributed on the interval $[\Delta - a, \Delta + a]$. The parameter $a \in \mathbb{R}^+$ is no greater than Δ so that $\Delta - a \geq 0$. The characteristic equation is then given by

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{2arN} \cdot e^{-r\Delta}(e^{ra} - e^{-ra}) = 0. \tag{267}$$

5.4.6 Gamma Distribution

Another continuous distribution of interest is the gamma distribution. Unlike the uniform distribution, gamma distribution is unbounded and allows for more versatility in the actual shape of the probability density function, as its shape is determined by two parameters k and a.

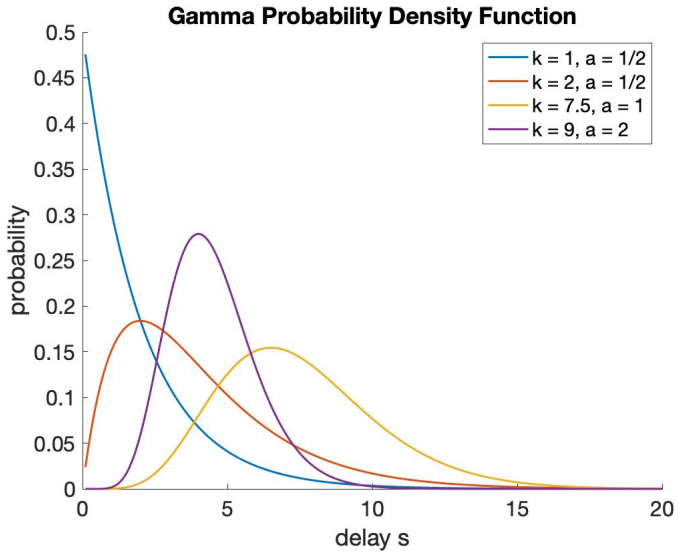

Figure 73: Probability density function of gamma distribution.

This density is specified by $f(s)$ for $s \geq 0$ as

$$f(s) = \frac{a^k}{\Gamma(k)} s^{k-1} e^{-as}, \tag{268}$$

with the Laplace transform $F(r) = \frac{a^k}{(r+a)^k}$. The eigenvalue equation follows from (226),

$$\Phi(r) = r + \mu + \frac{\lambda\theta a^k}{N(r+a)^k} = 0. \tag{269}$$

To determine where the queues may become unstable, we solve for $\lambda\theta/N$ and μ at the point where an eigenvalue is purely imaginary, so $r = ib$. We set $\tan(\phi) = \frac{b}{a}$:

$$(ib+\mu)\Big(\frac{ib}{a}+1\Big)^k + \frac{\lambda\theta}{N} = (ib+\mu)\Big(i\tan(\phi)+1\Big)^k + \frac{\lambda\theta}{N} = 0. \tag{270}$$

This equation can be simplified through the de Moivre's formula:

$$(ib+\mu)\Big(i\sin(\phi)+\cos(\phi)\Big)^k = -\frac{\lambda\theta}{N}\cos^k(\phi) \tag{271}$$

$$(ib+\mu)\Big(i\sin(k\phi)+\cos(k\phi)\Big) = -\frac{\lambda\theta}{N}\cos^k(\phi). \tag{272}$$

Separating the real and the imaginary parts of the equation we find

$$-b\sin(k\phi) + \mu\cos(k\phi) = -\frac{\lambda\theta}{N}\cos^k(\phi) \tag{273}$$

$$\mu\sin(k\phi) = -b\cos(k\phi), \tag{274}$$

so μ and $\frac{\lambda\theta}{N}$ can be expressed as functions of b,

$$\mu = -b\cot(k\phi) \tag{275}$$

$$\frac{\lambda\theta}{N} = (b\sin(k\phi) + b\cot(k\phi)\cos(k\phi))/\cos^k(\phi). \tag{276}$$

Based on these equalities, the Hopf curve can be found numerically. Figure 74, for example, is a result of setting $k = 2$ and varying the other parameter a. This in turn changes the average delay $E[f(s)] = \frac{k}{a}$.

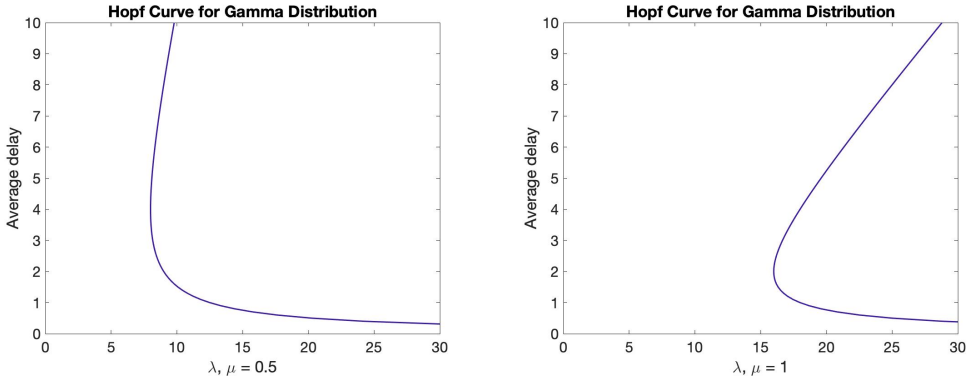

Figure 74: Hopf curve for gamma distribution, given that $k = 2$.

Exponential Distribution

Exponential distribution is a special case of gamma distribution where $k = 1$. For exponentially distributed delay, the queues are always stable as seen by the following proposition.

Proposition 5.9. *When the delay distribution is given by an exponential distribution, the queueing system given in Equation* (203) *is always stable.*

Proof. The characteristic equation given in Equation (269) simplifies to a quadratic equation with respect to the eigenvalue r, namely,

$$r^2 + r(\mu + a) + \left(\mu a + \frac{\lambda\theta}{N} a\right) = 0, \tag{277}$$

$$r = \frac{1}{2}\left(-(\mu + a) \pm \sqrt{(\mu + a)^2 - 4a\left(\mu + \frac{\lambda\theta}{N}\right)}\right). \tag{278}$$

If the discriminant is non-positive, then $\mathrm{Re}[r] = -(\mu + a) < 0$ so the queues are locally stable. If the discriminant is positive then

$$(\mu + a)^2 - 4a\left(\mu + \frac{\lambda\theta}{N}\right) = (\mu - a)^2 - \frac{4a\lambda\theta}{N} < (\mu - a)^2, \tag{279}$$

which reveals that the eigenvalue must be real and negative:

$$r = \frac{1}{2}\left(-(\mu + a) \pm \sqrt{(\mu + a)^2 - 4a\left(\mu + \frac{\lambda\theta}{N}\right)}\right) \tag{280}$$

$$< \frac{1}{2}\Big(- (\mu + a) + |\mu - a|\Big) < \max[-\mu, -a] < 0. \tag{281}$$

□

When the delay is exponentially distributed, the queues are always stable. Figure 75 shows that the queueing model with gamma distribution becomes more stable as k goes to 1.

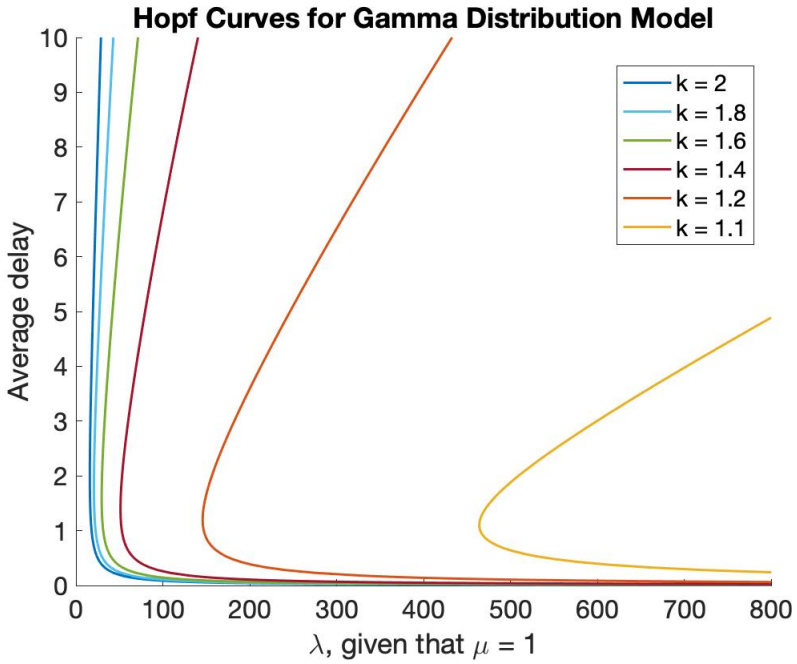

Figure 75: As $k \to 1$, gamma distribution model becomes more stable.

Gamma Distribution Converges to Constant Delay

When $k \to \infty$, a special scaling of the gamma distribution converges to a Dirac delta function, and the delay becomes a deterministic value. Hence, the gamma distribution model converges to the constant delay model. To show this, we rename the average delay to be Δ, or $\frac{k}{a} = \Delta$, setting therefore a to be k/Δ. The term from the characteristic equation (269) with k then becomes

$$\lim_{k\to\infty} \frac{a^k}{(r+a)^k} = \lim_{k\to\infty} \left(\frac{k}{r\Delta + k}\right)^k = \lim_{k\to\infty} \exp\left(k \ln\left(\frac{k}{r\Delta + k}\right)\right). \quad (282)$$

The limit in the exponent can be evaluated by the L'Hopital's rule

$$\lim_{k\to\infty} k \ln\left(\frac{k}{r\Delta+k}\right) = \lim_{k\to\infty} \frac{\ln\left(\frac{k}{r\Delta+k}\right)}{\frac{1}{k}} = \lim_{k\to\infty} \frac{\frac{r\Delta+k}{k}\cdot\frac{r\Delta}{(r\Delta+k)^2}}{-k^{-2}} \tag{283}$$

$$= \lim_{k\to\infty} -\frac{r\Delta k}{r\Delta+k} = -r\Delta. \tag{284}$$

Combining (284) with (282), we find that

$$\lim_{k\to\infty} \frac{a^k}{(r+a)^k} = e^{-r\Delta}, \tag{285}$$

which confirms that in the limit as $k \to \infty$ the characteristic equation of the gamma distribution model (269) converges to the characteristic equation of the constant delay model (241):

$$r + \mu + \frac{\lambda\theta a^k}{N(r+a)^k} \to r + \mu + \frac{\lambda\theta}{N} e^{-r\Delta}. \tag{286}$$

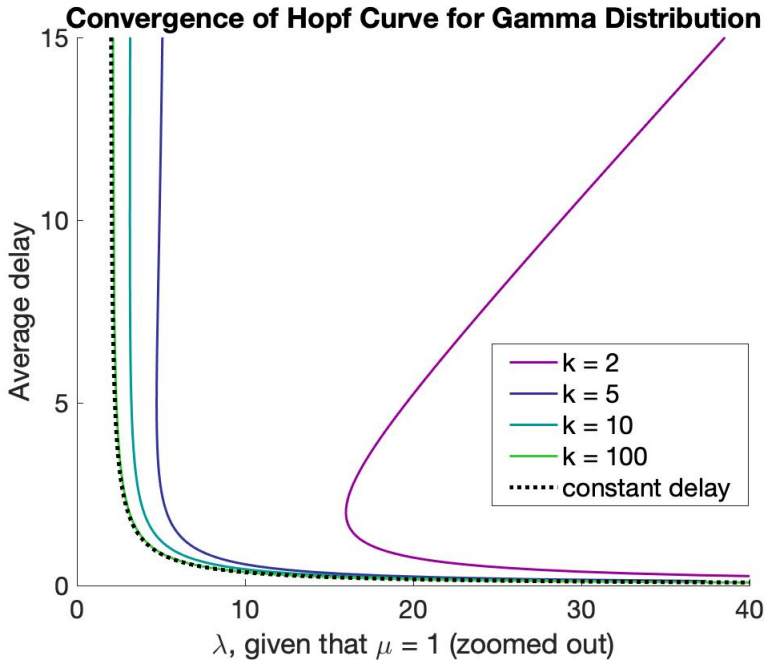

Figure 76: As $k \to \infty$, gamma distribution model converges to constant delay model.

5.5 Data-Driven Approximations of Stability Regions for Unknown Distributions

In a physical setting the service managers of a queueing system may not know the distribution of their customers' commute time. This motivates us to study the stability regions of a queueing system with distributed delay based on the moments of the delay distribution, which can be determined from physical setting.

First, we will summarize our findings about upper and lower bounds on the system's stability, based on the knowledge of the mean delay, symmetry

of the distribution, and whether or not it is bounded.

- If the average delay is known, any model with a symmetric distribution will be at least as stable as the constant delay model with that average delay. This is proven in Theorem 4.0.5 by Bernard, et al. [8]. So if $\int_0^\infty sf(s)ds = \Delta$ and $\lambda\theta > N\mu$, the queues are asymptotically stable when

$$\Delta < \frac{\arccos(-N\mu/(\lambda\theta))}{\sqrt{\lambda^2\theta^2/N^2 - \mu^2}}. \tag{287}$$

  The condition for Δ is derived in Equation (243), and it provides a lower bound on stability of a queueing system. Further, we hypothesize that the same bound holds for non-symmetric delay distributions as well, but we have not been able to prove it.

- If the delay distribution is unbounded, the upper bound on stability based on the average delay does not exist. This is evident from the following example. For any average delay, one can consider the delay being exponentially distributed, which results in a stable queueing system, as shown in Section 5.4.6.

- Even if the distribution is known to be bounded, an upper bound on stability based on average delay is still not guaranteed. One can consider the distribution from Section 5.4.2, where some customers experience a fixed delay and the rest experience no delay. This delay distribution is

bounded, but still the queueing system is always stable if at least half of the customers receive no delay.

Next, we propose a method that approximates the Hopf curve using the moments of the delay distribution. The service manager has the freedom to choose how many (or how few) moments to incorporate. Generally, higher moments will produce a more accurate approximation, but it is sufficient to know just the average delay and its variance based on sampling the customers in order to get a rough idea for when the queues will become unstable.

The approximation method relies on expanding the Laplace transform of the delay distribution in an infinite series, and then using the truncated series in order to find imaginary eigenvalues of the characteristic equation (226). This allows us to avoid dealing with the probability density function $f(s)$, and instead use its moments. Lemma 5.10 below establishes the connection between the Laplace transform and the moments of a given probability distribution.

Lemma 5.10. *The Laplace transform of the non-negative random variable X can be expanded in the following Taylor series around a generic point a*

$$E[e^{-rX}] \;=\; F(r) = e^{-ra}\left(\sum_{j=0}^{\infty} \frac{(-r)^j \cdot E[(X-a)^j]}{j!}\right). \tag{288}$$

Moreover, when $a = 0$ we have

$$E[e^{-rX}] \;=\; F(r) = \sum_{n=0}^{\infty} \frac{(-r)^n}{n!} \cdot E[X^n]. \tag{289}$$

and when $a = E[X]$, we have

$$E[e^{-rX}] \;=\; F(r) = e^{-rE[X]} \left(1 + \sum_{j=2}^{\infty} \frac{(-r)^j \cdot E[(X - E[X])^j]}{j!} \right). \quad (290)$$

Proof. This follows immediately from standard Taylor expansions. □

The random variable X from Lemma 5.10 represents the delay of an individual customer, and it is specified by the probability density function $f(s)$. Therefore the average delay can be found as $\Delta = E[X] = \int_0^\infty s f(s) ds$, and the j-th centered moment can be expressed as

$$E[(X - \Delta)^j] = \int_0^\infty (s - \Delta)^j f(s) ds, \quad \text{where} \quad \Delta = \int_0^\infty s f(s) ds. \quad (291)$$

It follows that the characteristic equation (226) can be expressed as

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N} \cdot e^{-r\Delta} \left(1 + \sum_{j=2}^{\infty} \frac{(-r)^j \cdot E[(X - \Delta)^j]}{j!} \right) = 0. \quad (292)$$

If the data about the first K central moments is available, then the characteristic equation can be approximated by the truncated series

$$\Phi(r) \approx r + \mu + \frac{\lambda\theta}{N} \cdot e^{-r\Delta} \left(\sum_{j=0}^{K} \frac{(-r)^j}{j!} \int_0^\infty (s - \Delta)^j f(s) ds \right), \quad (293)$$

and solving numerically $\Phi(ib) = 0$, $b \in \mathbb{R}$, will produce an approximation to the Hopf curve.

When the distribution of the delay is symmetric, Equation (293) can be

further simplified. The odd central moments are zero, $E[(X-\Delta)^{2j+1}] = 0$, which so the Laplace transform can be expressed as

$$F(r) = E[e^{-rX}] = e^{-r\Delta}\Big(1 + \sum_{j=1}^{\infty} \frac{(-r)^j}{j!} E[(X-\Delta)^j]\Big) \tag{294}$$

$$= e^{-r\Delta}\Big(1 + \sum_{j=1}^{\infty} \frac{r^{2j}}{2j!} E[(X-\Delta)^{2j}]\Big). \tag{295}$$

At the point of a Hopf bifurcation when the eigenvalue r becomes purely imaginary, i.e. $r = ib$, the expression for the Laplace transform takes the form of alternating series

$$F(r) = \Big(\cos(b\Delta) - i\sin(b\Delta)\Big) \tag{296}$$

$$\cdot \sum_{j=0}^{\infty} \Big((-1)^j \cdot \frac{b^{2j}}{2j!} \cdot E\big[(X-\Delta)^{2j}\big]\Big)$$

$$= \Big(\cos(b\Delta) - i\sin(b\Delta)\Big) \cdot \sum_{j=0}^{\infty} (-1)^j k_j, \tag{297}$$

$$\text{where} \quad k_j = \frac{b^{2j}}{2j!} \cdot E\big[(X-\Delta)^{2j}\big].$$

The characteristic equation at the Hopf takes the form

$$\Phi(ib) = ib + \mu + \frac{\lambda\theta}{N} \cdot \Big(\cos(b\Delta) - i\sin(b\Delta)\Big) \cdot \sum_{j=0}^{\infty} (-1)^j k_j = 0. \tag{298}$$

Therefore, an approximation based on K central moments to where the Hopf

bifurcation occurs is given by the solution of the system of equations

$$\begin{cases} \mu + \frac{\lambda\theta}{N} \cdot \cos(b\Delta) \cdot \sum_{j=0}^{K} (-1)^j k_j = 0 \\ b - \frac{\lambda\theta}{N} \cdot \sin(b\Delta) \cdot \sum_{j=0}^{K} (-1)^j k_j = 0. \end{cases} \tag{299}$$

If the Taylor expansion of the Laplace transform consists of terms with decreasing magnitude, then we can have an upper and a lower bound on the Laplace transform. Additionally, the upper and the lower bound are guaranteed to be tighter as more terms of the truncated Taylor series are included.

Theorem 5.11. *Suppose X is a symmetric non-negative random variable and $a = E[X]$. Further, the terms $k_j = \frac{b^{2j}}{2j!} \cdot E\big[(X-a)^{2j}\big]$ are decreasing, or $k_j \geq k_{j+1}$ for all $j \geq 0$. Then we can derive an upper and lower bound the Laplace transform $F(r)$ at the bifurcation point $r = ib$:*

$$-\sin(ba) \sum_{j=0}^{2N} (-1)^j k_j \leq \mathrm{Im}\big(F(r)\big) \leq -\sin(ba) \sum_{j=0}^{2N+1} (-1)^j k_j \tag{300}$$

$$\cos(ba) \sum_{j=0}^{2N} (-1)^j k_j \leq \mathrm{Re}\big(F(r)\big) \leq \cos(ba) \sum_{j=0}^{2N+1} (-1)^j k_j, \tag{301}$$

where $N \geq 0$ is an arbitrary integer. Furthermore, the bounds are guaranteed to be tighter for larger N.

Proof. By Equation (297),

$$\lim_{n\to\infty}\Big(\cos(ba) - i\sin(ba)\Big)\cdot S_n = F(ib), \tag{302}$$

$$\text{where}\quad S_n = \sum_{j=0}^{n}(-1)^j k_j. \tag{303}$$

For any n, S_n is a real quantity, so by separating the real and imaginary parts of (302), we see that

$$\lim_{n\to\infty}\cos(ba)\cdot S_n = \mathrm{Re}[F(ib)] \tag{304}$$

$$\lim_{n\to\infty} -\sin(ba)\cdot S_n = \mathrm{Im}[F(ib)] \tag{305}$$

Since $k_n = \frac{b^{2n}}{2n!}\cdot E\big[(X-a)^{2n}\big]$ is non-negative for all non-negative integers n and $\{k_n\}$ is a decreasing sequence by assumption, the sequence $\{S_{2n}\}$ is monotonically decreasing. Specifically, for any $n \geq 0$

$$S_{2(n+1)} = \sum_{j=0}^{2(n+1)}(-1)^j k_j = \sum_{j=0}^{2n}(-1)^j k_j + \big(-k_{2n+1} + k_{2n+2}\big) \tag{306}$$

$$\leq \sum_{j=0}^{2n}(-1)^j k_j = S_{2n}. \tag{307}$$

Similarly, the sequence of partial sums $\{S_{2n+1}\}$ is monotonically increasing,

$$S_{2(n+1)+1} = \sum_{j=0}^{2(n+1)+1} (-1)^j k_j = \sum_{j=0}^{2n+1} (-1)^j k_j + \big(k_{2n+2} - k_{2n+3}\big) \tag{308}$$

$$\geq \sum_{j=0}^{2n+1} (-1)^j k_j = S_{2n+1}. \tag{309}$$

Lastly, we note that $b > 0$ must satisfy the characteristic equation (226), meaning that

$$\begin{cases} \mu + \mathrm{Re}\big[\frac{\lambda\theta}{N} F(ib)\big] &= 0 \\ b + \mathrm{Im}\big[\frac{\lambda\theta}{N} F(ib)\big] &= 0, \end{cases} \tag{310}$$

$$\begin{cases} \mu + \frac{\lambda\theta}{N} \cdot \cos(ba) \sum_{j=0}^{\infty} (-1)^j k_j &= 0 \\ b - \frac{\lambda\theta}{N} \sin(ba) \sum_{j=0}^{\infty} (-1)^j k_j &= 0. \end{cases} \tag{311}$$

The series $\sum_{j=0}^{\infty} (-1)^j k_j > 0$ since the sequence $k_j > 0$ for every j and $\{k_j\}$ is decreasing, which therefore dictates that

$$\cos(ba) < 0 \quad \text{and} \quad \sin(ba) > 0. \tag{312}$$

Since $\cos(ba) < 0$ and $\{S_{2n}\}$ is a decreasing sequence, we can conclude that for any non-negative n

$$\cos(ba) S_{2n} \geq \cos(ba) S_{2(n+1)} \geq \mathrm{Re}[F(ib)] \tag{313}$$

Further, since $\{S_{2n+1}\}$ is an increasing sequence, we prove the upper bound for $\mathrm{Re}[F(ib)]$:

$$\cos(ba)S_{2n+1} \leq \cos(ba)S_{2(n+1)+1} \leq \mathrm{Re}[F(ib)]. \tag{314}$$

These inequalities state an upper and lower bound on the real part of the Laplace transform, as given in Equation (301). Similarly, because $\sin(ba) > 0$, we get the upper and lower bounds on the imaginary part of the Laplace transform as in Equation (300):

$$-\sin(ba)S_{2n} \leq -\sin(ba)S_{2(n+1)} \leq \mathrm{Im}[F(ib)] \tag{315}$$

$$-\sin(ba)S_{2n+1} \geq -\sin(ba)S_{2(n+1)+1} \geq \mathrm{Im}[F(ib)]. \tag{316}$$

Note that Equations (313)-(316) also demonstrate that choosing a larger n provides tighter bounds on both the real and the imaginary parts of $F(r)$. $\square$

Since the exact point of the Hopf bifurcation is given by

$$\mu + \frac{\lambda\theta}{N}\,\mathrm{Re}(F(ib)) = 0 \tag{317}$$

$$b - \frac{\lambda\theta}{N}\,\mathrm{Im}(F(ib)) = 0, \tag{318}$$

or when the arrival rate of the customers satisfies the equation

$$\lambda = -\frac{\mu N}{\theta\,\mathrm{Re}(F(ib))} > 0, \tag{319}$$

then by Theorem 5.11 for every $n \geq 0$ we have the following upper and lower bounds on λ:

$$\lambda \geq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2(n+1)} (-1)^j k_j} \geq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2n} (-1)^j k_j}, \tag{320}$$

$$\lambda \leq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2n+3} (-1)^j k_j} \leq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2n+1} (-1)^j k_j}. \tag{321}$$

In other words, when the first $4n$ central moments are included to approximate λ where the Hopf bifurcation occurs, we get a lower bound. When the first $4n+2$ central moments are used, we get an upper bound. Additionally, these bounds get tighter as more moments are incorporated.

5.5.1 Examples Using the Approximation Method

We demonstrate numerically the performance of the approximation from Equations (299) based on several delay distributions.

Discrete uniform delays on a bounded interval Recall the distribution from Section 5.4.4, where there are $M+1$ evenly distributed discrete delays $\{0, \frac{2\Delta}{M}, \frac{4\Delta}{M}, \ldots, 2\Delta\}$ with equal probability of occurring $\frac{1}{M+1}$. Since we already know how a queueing system behaves when $M = 1$ from Section 5.4.2, we will assume that $M \geq 2$ so there are three or more delays total. The odd central moments are zero, while the even central moments are given

by the formula

$$E[(X-\Delta)^{2n}] = \begin{cases} \frac{2}{M+1} \cdot \sum_{j=1}^{M/2} \left(\frac{2\Delta j}{M}\right)^{2n}, & M \text{ is even} \\ \frac{2}{M+1} \cdot \sum_{j=1}^{(M-1)/2} \left(\frac{2\Delta j}{M} - \frac{\Delta}{M}\right)^{2n}, & M \text{ is odd.} \end{cases} \tag{322}$$

In Figure 77, we consider a distribution with six delays or $M = 5$ in the left plot, and ten delays of $M = 9$ in the right plot. For each distribution, we include $2, 4$, and 20 central moments, and plot the Hopf curves resulting from the system of equations (299). The second order approximation for both distributions predicts the queues to be more stable than they actually are (for a fixed average delay), while the fourth order approximation predicts the queues to be less stable. The twentieth order, however, approximates the Hopf curve very accurately.

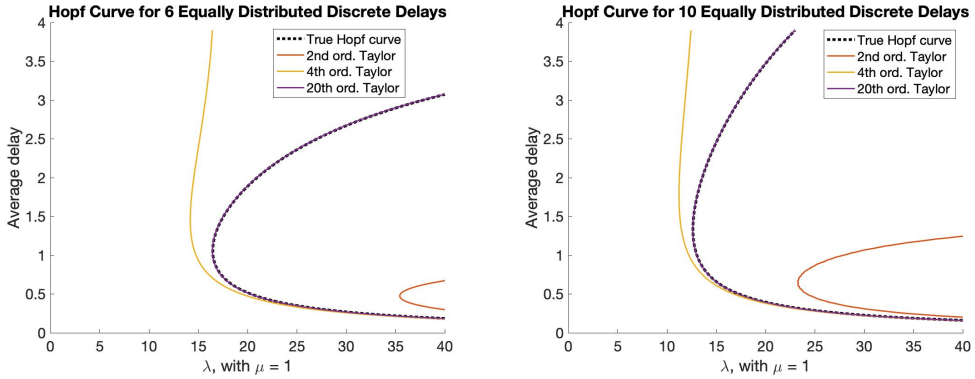

Figure 77: The approximation method applied to a system with 6 delays (on the left) and 10 delays (on the right).

Uniform distribution The uniform distribution on interval $[0, 2\Delta]$ has even central moments given by

$$E[(X-\Delta)^{2n}] = \frac{\Delta^{2n}}{2n+1}. \tag{323}$$

Figure 78 shows the Hopf curve approximations when the two, four, six, and twenty central moments are included in the system of equations (299). Based on the plot, as the number of utilized moments increases, the approximation becomes more accurate and converges to the true Hopf curve. Further, the second and the sixth order approximations give an upper bound with respect to λ, whereas the fourth order gives a lower bound.

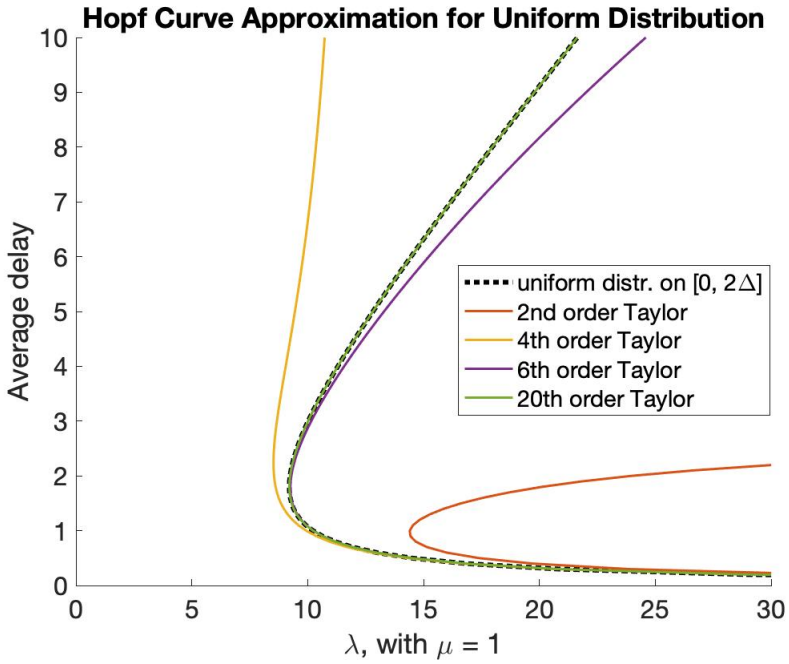

Figure 78: Approximations to a model with uniformly distributed delay on $[0, 2\Delta]$.

If the delay is uniformly distributed on the interval $[\Delta - a, \Delta + a]$ where $0 < a \leq \Delta$, the even central moments are

$$E[(X - \Delta)^{2n}] = \frac{a^{2n}}{2n + 1}. \tag{324}$$

Below we consider a queueing system where the delay is distributed proportionally to the average delay on the interval $[0.5\Delta, 1.5\Delta]$ (so $a = 0.5\Delta$ from Equation (324)). The left plot in Figure 79 shows that the second, fourth, and twentieth order approximations are so close to the true Hopf curve that is difficult to even distinguish the curves. The plot on the right shows a zoomed in version of the same plot, where one can see the second and fourth order approximations deviating from the true solution.

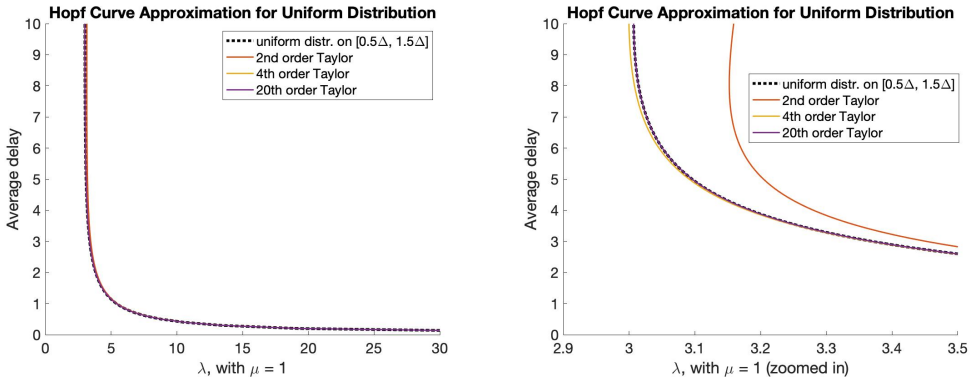

Figure 79: An approximation to a system with uniformly distributed delay on $[.5\Delta, 1.5\Delta]$.

5.6 Conclusion

The generalized queueing system from Equation (203) has a unique equilibrium state, the local stability of which is determined by the distribution of the customer delay. Under certain distributions, such as the exponential distribution or real-time and delay distribution, the queues may remain locally and asymptotically stable regardless of the size of the delay, the service rate, and the arrival rate of the customers to the queues. For other delay distributions, however, the queues are guaranteed to be asymptotically stable when the relationship $\lambda\theta < N\mu$ holds, but once the parameters change a Hopf bifurcation may occur and the equilibrium may become unstable. Common delay distributions are considered in Section 5.4, where for each distribution we study the stability region of the queues.

The stability is uniquely determined from the characteristic equation when the delay distribution is known. However, we also consider the scenario when only certain moments of the distribution are known. This is motivated by physical settings where the moments can be approximated by sampling the incoming customers. We propose an approximation method that utilizes the moments of the delay distribution in order to determine whether or not the queues are stable.

In the course of writing this paper, however, we discovered more questions than answers. A natural extension of this paper is to conduct further study of the queueing models with specific distributions. For example, one can ask how does the delay distribution affect the amplitude of oscillations in queues

that result from a Hopf bifurcation. Or, when the queueing system is locally stable, what is the rate of convergence of the queue lengths to the equilibrium. It may also be possible to show that, regardless of the delay distribution, if the equilibrium is locally unstable, then it will remain unstable for larger customer arrival rate λ and for a lower service rate μ.

Another research direction is to investigate whether the queueing systems can be ranked for their stability based on the characteristics of their delay distributions. We hypothesize that under a fixed average delay, the increased entropy of delay distribution leads to a more stable queueing system. Having such a criteria would be incredibly useful, and it could help in developing lower and upper bounds on the stability of queueing systems that are difficult to analyze analytically.

The approximation method proposed in Section 5.5 could be further developed improved. It would be great to develop guarantees on the accuracy of this method. By considering specific distributions, we noticed that the method performs much better for queueing systems with symmetric delay distributions. It would be interesting to understand why the method loses accuracy when the distribution is non-symmetric. We also noticed that Taylor series around the average delay produce by far the most accurate results, and we would like to learn why that is the case.

6 Thesis Conclusion

This thesis begins by considering a system of two queues with customers choosing which queue to join based on delayed information. We study the asymptotic behavior of this system and show that delay can cause oscillations in queues. To describe the behavior quantitatively, we develop a novel numerical method - the slope function method - for approximating the amplitude of oscillations in queues. This method is compared to a classic perturbations technique, showing its relative advantages in performance. The slope function method is not specific to queueing systems and can be applied to other dynamical systems where a Hopf bifurcation occurs.

We expand on this topic by considering a modified two-queue system, where the delay is uniformly distributed rather than constant, which reflects variation in the commute time of the customers to the queue. Again, the presence of delay can cause oscillations of unknown amplitude. This model serves as another test case for confirming the accuracy of approximation by the slope function method.

Beyond quantifying the amplitude, we offer a way to control and reduce the oscillations by tweaking the information that customers receive about the queues. When the information accounts for the rate of change of the queue length, the queueing system provably becomes more stable. Given the same delay, the oscillations are either reduced in size or eliminated completely.

Lastly, we formulate a generalized model of N queues where the delay

from the customer's commute is specified as a random variable from an arbitrary distribution. This model shows that, regardless of the delay distribution, any queueing system will follow a certain pattern of behavior. For example, the queues have a unique equilibrium state and under a certain parameter restriction the equilibrium is locally stable. Further, the queues can start oscillating due to a Hopf bifurcation. Whether or not the bifurcation occurs depends on the distribution of the delay. However, given a particular distribution, the stability region in the parameter space can be mapped out by numerically solving the characteristic equation, or the region can be approximately determined through a numerical method proposed in Chapter 5. The proposed method is handy when the specific distribution of the delay is unknown, and instead only the average delay and some central moments are available.

This thesis connects the field of queueing theory to nonlinear dynamics and, in particular, delay-differential equations. Our work opens doors for many other queueing models to be considered with mathematical techniques that may be new to the queueing community. Simultaneously, our work places queueing theory on the radar of the dynamical systems experts as a potential application area for their research with direct relevance to industry.

6.1 Future Research

Many ideas about future research have already been discussed in the conclusions of Chapters 3, 4, and 5. Now, however, we gained very powerful insights

from Chapter 5, so we can revisit the results of the previous chapters and view them from a new perspective.

Let us begin by discussing the slope function method from Chapters 2 - 3. Back when we initially proposed this approximation technique, we only tested its accuracy on two queueing systems - where the delay distribution is a deterministic constant, and where the delay is uniformly distributed. These two distributions were perhaps the simplest nontrivial examples to study, which is why they were chosen. The results from Chapter 5, however, can allow a more thorough testing of accuracy for the slope function method. First, we now know that a Hopf bifurcation is not unique to the queueing systems from Chapters 2 - 3, and oscillations are present for systems with many other delay distributions. Second, the characteristic equation for the generalized queueing system of Chapter 5 makes it easy to find whether the bifurcation occurs (and if so, where in the parameter space) for any delay distribution with a known Laplace transform. Therefore, the slope function method can be systematically tested for accuracy across queueing systems with a variety of delay distributions.

The results from Chapter 5 can also help study the new delay announcement technique proposed in Chapter 4, which helps to limit the size of oscillations. The usefulness of the novel delay announcement has been demonstrated on the constant delay model, which serves as a proof of concept. However, the generalized queueing system can be used to see how the new delay announcement affects queueing systems where the customer delay isn't

constant. It is possible that the new announcement is more beneficial for some delay distributions over others. However, it is potentially useful regardless of the delay distribution.

Lastly, this thesis only considers deterministic fluid-like models. These models in turn have a direct connection to queueing systems modeled as stochastic processes. Pender et al. [33] presents two stochastic models and proves that in a certain limiting regime their behavior is modeled by the queueing systems from Chapters 2 and 3. No such analysis has been done for the generalized queueing model of Chapter 5, and we recommend to explore this question.

References

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

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

[3] Hans Armbruster, Christian Ringhofer, and Tae Chang Jo. Continuous models for production flows. In *Proceedings of the American Control Conference*, volume 5, pages 4589–4594, 2004. ISBN 0780383354. doi: 10.1109/ACC.2004.182675.

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

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

[6] Mohamed Belhaq and Si Mohamed Sah. Fast parametrically excited van der pol oscillator with time delay state feedback. *International Journal of Non-Linear Mechanics*, 43(2):124–130, 2008.

[7] Alfredo Bellena and Nicola Guglielmi. Solving neutral delay differential equations with state-dependent delays. *Journal of Computational and Applied Mathematics*, 229(2):350–362, 2009.

[8] Samuel Bernard, Jacques Bélair, and Michael C Mackey. Sufficient conditions for stability of linear differential equations with distributed delay. *Discrete and Continuous Dynamical Systems Series B*, 1(2):233–256, 2001.

[9] Shui Nee Chow and Hans Otto Walther. Characteristic multipliers and stability of symmetric periodic solutions of $\dot{x}(t) = g(x(t-1))$. *Transactions of the American Mathematical Society, Vol. 307, No. 1, pp. 127-142*, 1988.

[10] S. L. Das and A. Chatterjee. Multiple scales without center manifold

reductions for delay differential equations near Hopf bifurcations. *Nonlinear Dynamics*, 30:323–335, 2002.

[11] Rodney D. Driver. Existence and continuous dependence of solutions of a neutral functional-differential equation. *Archive for Rational Mechanics and Analysis*, 19(2):149–166, 1965. ISSN 1432-0673. doi: 10.1007/BF00282279. URL `https://doi.org/10.1007/BF00282279`.

[12] Stephen G Eick, William A Massey, and Ward Whitt. Mt/g/ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252, 1993.

[13] Stephen G Eick, William A Massey, and Ward Whitt. The physics of the mt/g/ queue. *Operations Research*, 41(4):731–742, 1993.

[14] Thomas Erneux. *Applied Delay Differential Equations*. Springer Science, 2009.

[15] Brian H. Fralix and Ivo J. B. F. Adan. An infinite-server queue influenced by a semi-markovian environment. *Queueing Systems*, 61(1): 65–84, 2009.

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

[17] Pengfei Guo and Paul Zipkin. The impacts of customers' delay-risk sensitivities on a queue with balking. *Probability in the Engineering and Informational Sciences*, 23(3):409–432, 2009.

[18] Jack Hale and Verduyn Lunel. *Introduction to Functional Differential Equations*. Springer Science, 1993.

[19] Brian D. Hassard, N. D. Kazarinoff, and Y. H. Wan. *Theory and Applications of Hopf Bifurcation*. Cambridge University Press, 1981. URL https://books.google.com/books?id=3wU4AAAAIAAJ.

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

[21] Jerry Hausman and Daniel McFadden. Specification tests for the multinomial logit model. *Econometrica*, 52(5):1219–1240, 1984.

[22] Dirk Helbing. Improved fluid-dynamic model for vehicular traffic. *Physical Review E*, 51:3164–3169, 1995. doi: 10.1103/PhysRevE.51.3164. URL https://link.aps.org/doi/10.1103/PhysRevE.51.3164.

[23] Rouba Ibrahim, Mor Armony, and Achal Bassamboo. Does the past predict the future? the case of delay announcements in service systems. *Management Science*, 63(6):1657–2048, 2017.

[24] Donald L. Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, 2(2):429–441, 1965.

[25] Anatoli Ivanov, Bernhard Lani-Wayda, and Hans-Otto Walther. Unsta-

ble hyperbolic periodic solutions of differential delay equations. *World Scientific Series in Applicable Analysis: Volume 1, pp. 301-316*, 1992.

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

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

[28] Young Myoung Ko and Jamol Pender. Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes. *Stochastic Models*, 34(2):186–206, 2018.

[29] Lauren Lazarus, Matthew Davidow, and Richard Rand. Periodically forced delay limit cycle oscillator. *International Journal of Non-Linear Mechanics*, 94:216–222, 2017.

[30] D. Lipshutz and R. J. Williams. Existence, uniqueness, and stability of slowly oscillating periodic solutions for delay differential equations with nonnegativity constraints. *SIAM Journal on Mathematical Analysis*, 47 (6):4467–4535, 2015.

[31] Daniel McFadden. Modelling the choice of residential location. Cowles Foundation Discussion Papers 477, Cowles Foundation for Research in

Economics, Yale University, 1977. URL https://EconPapers.repec.org/RePEc:cwl:cwldpp:477.

[32] Samantha Nirenberg, Andrew Daw, and Jamol Pender. 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.

[33] Jamol Pender, Richard H. Rand, and Elizabeth Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 27(4), 2017.

[34] Jamol Pender, Richard H Rand, and Elizabeth Wesson. An asymptotic analysis of queues with delayed information and time varying arrival rates. *Nonlinear Dynamics*, 91:2411–2427, 2018.

[35] James Perkins and Poornachandran Kumar. Optimal control of pull manufacturing systems. *IEEE Transactions on Automatic Control*, 40 (12):2040–2051, 1995.

[36] Richard Rand. Differential-delay equations. In *Complex Systems: Fractionality, Time-Delay and Synchronization*, chapter 3, pages 83–117. A.C.J. Luo and J-Q Sun, eds., Springer, 2011.

[37] Richard Rand. *Lecture Notes on Nonlinear Vibrations*. University Readers, version 53, 2012. URL http://www.math.cornell.edu/~rand/randdocs/.

[38] Sidney Resnick and Gennady Samorodnitsky. Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Systems*, 33(1-3):43–71, 1999.

[39] Alexander Skubachevskii and Hans Otto Walther. On the floquet multipliers of periodic solutions to non-linear functional differential equations. *Journal of Dynamics and Differential Equations, Vol. 18, No. 2*, 2006.

[40] Hal Smith. *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer Science, 2011.

[41] Ying So and Warren Kuhfeld. Multinomial logit models. *SUGI 20 conference proceedings*, 1995.

[42] Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

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

[44] Xianwen Xie. Uniqueness and stability of slowly oscillating periodic solutions of delay equations with bounded nonlinearity. *Journal of Dynamics and Differential Equations, Vol. 3, No. 4*, 1990.

[45] Xianwen Xie. The multiplier equation and its application to s-solutions of a delay equation. *Journal of Differential Equations, 95*, 1992.

[46] Xianwen Xie. Uniqueness and stability of slowly oscillating periodic solutions of delay equations with unbounded nonlinearity. *Journal of Differential Equations, 103*, 1993.

7 Appendix

7.1 Approximating Amplitude of Oscillations for Queueing System with Constant Delay

Showing the existence and uniqueness of equilibrium

Proof of Theorem 98. When $q_i(t) = q_i(t-\Delta) = \frac{\lambda}{N\mu}$ for each $1 \leq i \leq N$, all functions q_i are constant with respect to time

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp(-\frac{\lambda}{N\mu})}{\sum_{j=1}^{N} \exp(-\frac{\lambda}{N\mu})} - \mu \frac{\lambda}{N\mu} = 0. \tag{325}$$

Therefore $q_i^* = \frac{\lambda}{N\mu}$ is an equilibrium.

To show uniqueness, we will argue by contradiction. Suppose there is another equilibrium given by $\bar{q}_i$, $1 \leq i \leq N$, and for some i we have $q_i^* \neq \bar{q}_i$. Without loss of generality, let us assume that it is the N'th queue, so $q_N^* \neq \bar{q}_N$. Also, without loss of generality let us assume that $q_N^* > \bar{q}_N$, and since both are constants with respect to time, we can conclude that $\bar{q}_N(t) = \frac{\lambda}{N\mu} + \epsilon$ for some $\epsilon > 0$.

From the condition $0 = \sum_{i=1}^{N} \dot{\bar{q}}_i$, the sum of the queues has to be $\sum_{i=1}^{N} \bar{q}_i =$

$\frac{\lambda}{\mu}$, so the average queue length is $\frac{1}{N}\sum_{i=1}^{N}\bar{q}_i = \frac{\lambda}{N\mu}$. Since $\bar{q}_N$ is greater than the average, then there must be some queue $\bar{q}_k$, $1 \leq k \leq N-1$, that is less than the average, so $\bar{q}_k = \frac{\lambda}{\mu N} - \delta$ for some $\delta > 0$. We can use this together with the condition $\dot{\bar{q}}_i = 0$ to get an expression

$$\sum_{i=1}^{N} \exp\big(-\bar{q}_i(t-\Delta)\big) = \frac{\lambda}{\mu} \cdot \frac{\exp\big(-\frac{\lambda}{N\mu} + \delta\big)}{(\frac{\lambda}{N\mu} - \delta)}, \tag{326}$$

which can now be used to show contradiction:

$$\begin{aligned} \dot{\bar{q}}_N(t) &= \lambda \frac{\exp\big(-\frac{\lambda}{N\mu} - \epsilon\big)}{\frac{\lambda}{\mu} \cdot \frac{\exp(-\frac{\lambda}{N\mu}+\delta)}{(\frac{\lambda}{N\mu}-\delta)}} - \mu\Big(\frac{\lambda}{N\mu} + \epsilon\Big) && (327) \\ &= -\frac{\lambda}{N}\big(1 - e^{-\epsilon-\delta}\big) - \mu\big(\epsilon + \delta e^{-\epsilon-\delta}\big) < 0. && (328) \end{aligned}$$

Hence $\bar{q}_i$ is not an equilibrium, and so the equilibrium must be unique. □

Showing stability of the equilibrium The following proposition is used to prove the stability of the equilibrium.

Proposition 7.1. *If there is a root $r = x + iy$ of the characteristic equation*

$$r = \alpha + \beta e^{-r\Delta} \tag{329}$$

with positive real part ($x > 0$) then it is bounded by $x \leq \alpha + |\beta|$ and $|y| \leq |\beta|$.

Proof. Plug $r = x + iy$ into Equation (329) and separate real and imaginary

parts to get

$$\cos(y\Delta) = \frac{e^{x\Delta}(x-\alpha)}{\beta}, \quad \sin(y\Delta) = -\frac{e^{x\Delta}y}{\beta} \tag{330}$$

These equations give the inequalities

$$-1 \leq \frac{e^{x\Delta}(x-\alpha)}{\beta} \leq 1, \quad -1 \leq -\frac{e^{x\Delta}y}{\beta} \leq 1 \tag{331}$$

Assuming that $x > 0$ and $\Delta \geq 0$, we know that $e^{x\Delta} \geq 1$. Therefore inequalities reduce to

$$-1 \leq \frac{(x-\alpha)}{\beta} \leq 1, \quad -1 \leq -\frac{y}{\beta} \leq 1, \tag{332}$$

and give the desired bounds $x \leq \alpha + |\beta|$ and $|y| \leq |\beta|$. □

Third order Taylor expansion A third order Taylor expansion of $\dot{q}_1(t)$ and $\dot{q}_2(t)$ is used to approximate the deviation of the queues from the equilibrium. This is required both by the Lindstedt's method and by the the slow flow method. To find the expansion, we define new functions $\tilde{u}_1$ and $\tilde{u}_2$ that represent the deviation of the queues q_1 and q_2 from the equilibrium state at $\frac{\lambda}{2\mu}$

$$q_1(t) = \frac{\lambda}{2\mu} + \tilde{u}_1(t), \quad q_2(t) = \frac{\lambda}{2\mu} + \tilde{u}_2(t). \tag{333}$$

Equations (2) - (3) give expressions for $\dot{\tilde{u}}_1(t)$ and $\dot{\tilde{u}}_2(t)$, which can be approximated by with a third degree polynomial about the equilibrium point

$\tilde{u}_1(t) = \tilde{u}_2(t) = 0$. We denote the approximations by $w_1(t)$ and $w_2(t)$

$$\begin{aligned}
\dot{w}_1(t) &= -\mu w_1(t) && (334)\\
&+ \lambda\left(-\frac{w_1 - w_2}{4} + \frac{w_1^3 - 3w_2w_1^2 + 3w_1w_2^2 - w_2^3}{48}\right)(t-\Delta)\\
\dot{w}_2(t) &= -\mu w_2(t) && (335)\\
&+ \lambda\left(-\frac{w_2 - w_1}{4} + \frac{w_2^3 - 3w_1w_2^2 + 3w_2w_1^2 - w_1^3}{48}\right)(t-\Delta)
\end{aligned}$$

Reduction to one cubic delay equation The symmetry of Equations (334) - (335) allows the equations to become uncoupled. We consider sum and the difference of w_1 and w_2,

$$\tilde{v}_1(t) = w_1(t) + w_2(t), \quad \tilde{v}_2(t) = w_1(t) - w_2(t). \tag{336}$$

This change of variables leads to the differential equations

$$\begin{aligned}
\dot{\tilde{v}}_1(t) &= -\mu\big(w_1(t) + w_2(t)\big) = -\mu\tilde{v}_1(t) && (337)\\
\dot{\tilde{v}}_2(t) &= \lambda\left(-\frac{\tilde{v}_2(t-\Delta)}{2} + \frac{\tilde{v}_2^3(t-\Delta)}{24}\right) - \mu\tilde{v}_2(t), && (338)
\end{aligned}$$

which are uncoupled. Equation (337) has the solution $\tilde{v}_1(t) = Ce^{-\mu t}$ so $\tilde{v}_1(t)$ decays to 0 regardless of what the delay parameter is, making $\tilde{v}_2(t)$ the function of interest.

Limit cycle stability via floquet exponents Theorem 4.14 shows that the Hopf bifurcations are supercritical by perturbing the system about the

point of bifurcation. However, the stability of limit cycles can also be determined by projecting the infinite-dimensional DDE on a center manifold, and then finding the characteristic floquet exponent of the resulting system of ODE's. This approach is explained in detail by Hassard et al. [19]. In our case, the DDE is given by

$$\dot{v} = -\mu v(t) - \frac{\lambda}{2} v(t-\Delta) + f\big(v(t), v(t-\Delta)\big), \tag{339}$$

where $f\big(v(t), v(t-\Delta)\big)$ contains all the nonlinear terms. To project Equation (339) onto a center manifold, we follow Chapter 14.3 in [37] precisely. First, we get rid of delay in our equation by defining $v_t(\theta) = v(t+\theta)$ for $\theta \in [-\Delta, 0]$ and the operators

$$Av_t(\theta) = \begin{cases} \frac{\partial x_t(\theta)}{\partial \theta} & \text{for} \quad \theta \in [-\Delta, 0) \\ -\mu v_t(0) - \frac{\lambda}{2} v_t(-\Delta) & \text{for} \quad \theta = 0 \end{cases} \tag{340}$$

$$Fv_t(\theta) = \begin{cases} 0 & \text{for} \quad \theta \in [-\Delta, 0) \\ f\big(v_t(0), v_t(-\Delta)\big) & \text{for} \quad \theta = 0 \end{cases} \tag{341}$$

so that the DDE (339) can be written as

$$\frac{d}{d\theta} v_t(\theta) = Av_t(\theta) + Fv_t(\theta). \tag{342}$$

We assume that $\Delta = \Delta_{cr}$, so there is a pair of purely imaginary roots $\Lambda = \pm i\omega_{cr}$ with the corresponding eigenfunctions $s_1(\theta)$ and $s_2(\theta)$ such that

$$A\big(s_1(\theta) + is_2(\theta)\big) = i\omega_{cr}\big(s_1(\theta) + is_2(\theta)\big). \tag{343}$$

The solution v_t of Equation (342) can then be expressed as a sum of points lying in the center subspace spanned by $s_1(\theta)$ and $s_2(\theta)$, and the points that don't lie in the center subspace, which is the rest of the solution and we denote it by w:

$$v_t(\theta) = y_1(t)s_1(\theta) + y_2(t)s_2(\theta) + w(t,\theta). \tag{344}$$

The idea of the center manifold reduction is to approximate w as a function of y_1 and y_2 (the center manifold), therefore replacing the infinite dimensional system with a two dimensional approximation. After some algebra we can determine y_1 and y_2 to be

$$\dot{y}_1 = \omega y_2 - \frac{\big(4\omega y_2 + 4(\mu + \Delta\mu^2 + \Delta\omega^2)y_1\big)^3}{24\lambda^2\big((1+\Delta\mu)^2 + \Delta^2\omega^2\big)^3} + O\big(y_i^5\big) \tag{345}$$

$$\dot{y}_2 = -\omega y_1 + O\big(y_i^5\big). \tag{346}$$

Now we will follow the technique in Chapter 1 of Hassard et al. [19] to analyze the stability of y_1 and y_2. The system of ODE's (345) - (346) can be

equivalently written as

$$\dot{z} = \omega z + \sum_{2 \leq i+j \leq L} g_{ij} \frac{z^i \bar{z}^j}{i!j!} + O(|z|^{L+1}), \tag{347}$$

where z is a complex function $z = y_2 + iy_1$, and $\bar{z}$ is its complex conjugate. The coefficients g_{ij} can be determined from Equations (345) - (347), and we find that $g_{20} = g_{02} = g_{11} = 0$ and

$$g_{21} = -\frac{2(\mu - i\omega_{cr})(\mu + i\omega_{cr})^2}{\lambda^2(1 + \Delta_{cr}(\mu - i\omega_{cr}))(1 + \Delta_{cr}(\mu + i\omega_{cr}))^2}. \tag{348}$$

Further, if the floquet exponent is negative, then the bifurcating periodic solutions of Equation (347) are asymptotically, orbitally stable with asymptotic phase. When Δ is sufficiently close to Δ_{cr}, the floquet exponent is of the same sign as β_2 that is given by Equation (5.9) in Chapter 1 of Hassard et al. [19]:

$$\beta_2 = 2 \text{ Re } \Big[\frac{i}{2\omega_{cr}}\big(g_{20}g_{11} - 2|g_{11}|^2 - \frac{1}{3}|g_{02}|^2\big) + \frac{1}{2}g_{21}\Big]. \tag{349}$$

Hence,

$$\beta_2 = -\frac{2(\mu^2 + \omega_{cr}^2)(\mu + \Delta_{cr}\mu^2 + \Delta_{cr}\omega_{cr}^2)}{\lambda^2((1 + \Delta_{cr}\mu)^2 + \Delta_{cr}^2\omega_{cr}^2)^2} < 0, \tag{350}$$

so the floquet exponent is negative and the limit cycle is stable in its center manifold.

7.2 Approximating Amplitude of Oscillations for Queueing System with Uniformly Distributed Delay

Showing the existence and uniqueness of equilibrium

Proof of Theorem 3.3. Suppose the queues are in equilibrium. Then $q_1(t) = q_1^*$, $q_2(t) = q_2^*$, $m_1(t) = \frac{1}{\Delta}\int_{t-\Delta}^{t} q_1(s)ds = q_1^*$, and $m_2(t) = \frac{1}{\Delta}\int_{t-\Delta}^{t} q_2(s)ds = q_2^*$. By summing Equations (237) - (47) we find

$$\lambda - \mu(q_1^* + q_2^*) = 0, \quad q_1^* = \frac{\lambda}{\mu} - q_2^*. \tag{351}$$

Eliminating q_1^* from Equations (237) - (47) and subtracting one equation from the other, we find that for $x = 2q_2^* - \frac{\lambda}{\mu}$

$$x = \frac{\lambda}{\mu}\left(\frac{1 - e^x}{1 + e^x}\right). \tag{352}$$

Since $\frac{\lambda}{\mu} > 0$, when $x > 0$ the right-hand side of Equation (352) is negative so $x \leq 0$. Similarly, when $x < 0$ then the right hand side of the equation is positive, which means that $x = 0$ is the only solution. Hence $q_2^* = \frac{\lambda}{2\mu}$ and $q_1^* = \frac{\lambda}{\mu} - q_2^* = \frac{\lambda}{2\mu}$ is the only equilibrium point of $q_1(t)$ and $q_2(t)$, which implies that $m_1(t) = m_2(t) = \frac{\lambda}{2\mu}$. □

Showing stability of the equilibrium The equilibrium is stable whenever all eigenvalues of the characteristic equation (55) have negative real parts. The following propositions help to establish that.

Proposition 7.2. *Any real eigenvalue of the characteristic equation* (55) *is negative.*

Proof. Under the assumption $\Lambda \neq 0$ and $\Lambda \in \mathbb{R}$, the characteristic equation can be rewritten as

$$1 + \frac{2\Delta}{\lambda} \cdot \Lambda(\Lambda + \mu) = e^{-\Lambda\Delta}. \tag{353}$$

The left hand side (LHS) and the right hand side (RHS) intersect at $\Lambda = 0$, and for $\Lambda > 0$ the LHS is monotonically increasing while the RHS is monotonically decreasing. Hence when $\Lambda \in \mathbb{R}$, this equality can only hold for $\Lambda < 0$. □

Proposition 7.3. *If* $\Delta \geq \frac{\lambda}{\mu^2}$, *then any complex eigenvalue of the Equation* (55) *has a negative real part.*

Proof. We will argue by contradiction. Assume that $\Delta \geq \frac{\lambda}{\mu^2}$, $a \geq 0$, and $b \neq 0$, for some $\Lambda = a + ib$ where $a, b \in \mathbb{R}$. We substitute Λ into Equation (55) and separate the real and imaginary parts:

$$\cos(b\Delta)e^{-a\Delta}\lambda = 2a^2\Delta - 2b^2\Delta + \lambda + 2a\mu\Delta \tag{354}$$

$$\sin(b\Delta)e^{-a\Delta}\lambda = -2b\Delta(2a + \mu). \tag{355}$$

Summing the squares of the two equations, we get

$$e^{-2a\Delta}\lambda^2 = \big(2a^2\Delta - 2b^2\Delta + \lambda + 2a\mu\Delta\big)^2 + \big(2b\Delta(2a + \mu)\big)^2, \tag{356}$$

and after some algebra we find

$$b^2 \leq \frac{1}{\Delta}(\lambda + 2a\mu\Delta + 2a^2\Delta) - (2a+\mu)^2 \tag{357}$$

$$= \frac{\lambda}{\Delta} - \mu^2 - 2a(a+\mu) \leq -2a(a+\mu) \leq 0, \tag{358}$$

so b must be 0, which contradicts our assumption. Therefore $\mathrm{Re}[\Lambda] = a < 0$ for any complex eigenvalue when $\Delta > \frac{\lambda}{\mu^2}$. □

It is now left to establish the stability of the equilibrium.

Proof of Theorem 3.3. We will show that for the specified range of Δ, all eigenvalues of the characteristic equation (55) have negative real parts. Recall that prior to deriving the characteristic equation, we considered the case with the trivial eigenvalue separately, so to analyze the stability we now only need to look at the non-trivial eigenvalues. Proposition 7.2 shows that any nontrivial real eigenvalue must be negative. Hence, it remains to show that the complex eigenvalues have negative real parts.

Case 1. Suppose the characteristic equation (68) does not have positive roots Δ_{cr}. This implies that a complex eigenvalue Λ never reaches the imaginary axis as Δ varies. Since Λ is continuous as a function of Δ, then $\mathrm{Re}[\Lambda]$ must be of the same sign for all $\Delta > 0$. Proposition 3.2 shows that for sufficiently small Δ, all complex eigenvalues have negative real parts, which is therefore true for all $\Delta > 0$.

Case 2. Suppose Equation (68) has at least one positive root Δ_{cr}. By the continuity of Λ with respect to Δ, $\mathrm{Re}[\Lambda]$ must be of the same sign on the

interval where Δ is less than the smallest positive root of Equation (68), and by Proposition 3.2 the sign is negative. Same holds when Δ is greater than the largest root Δ_{cr} of Equation (68). Any root Δ_{cr} is less than $\frac{\lambda}{\mu^2}$ by the condition $0 \neq \omega_{cr} \in \mathbb{R}$, and for Δ exceeding $\frac{\lambda}{\mu^2}$ all complex eigenvalues have negative real parts by Proposition 7.3. Therefore, the continuity of Λ implies that $\mathrm{Re}[\Lambda] < 0$ whenever Δ exceeds the largest root Δ_{cr}.

We showed that for the specified ranges of Δ, all eigenvalues have negative real parts and therefore the equilibrium is stable. □

Third order polynomial expansion We will perform a Taylor series expansion for the deviations about the equilibrium (52) of equations (46) - (49) and keep terms up to the third order. To start, we find the perturbations of our functions from the equilibrium,

$$q_1(t) = \frac{\lambda}{2\mu} + \tilde{u}_1(t), \quad q_2(t) = \frac{\lambda}{2\mu} + \tilde{u}_2(t), \tag{359}$$

$$m_1(t) = \frac{\lambda}{2\mu} + \tilde{u}_3(t), \quad m_2(t) = \frac{\lambda}{2\mu} + \tilde{u}_4(t), \tag{360}$$

and from Equations (46) - (49) we find their derivatives. A third order polynomial expansion of $\dot{\tilde{u}}_i(t)$ is given by $\dot{w}_i(t)$, where

$$\begin{aligned}
\dot{w}_1(t) &= \lambda \cdot \left(- \frac{w_3(t) - w_4(t)}{4} \right) - \mu w_1(t) \\
&+ \lambda \cdot \left(\frac{w_3^3(t) - 3w_4(t)w_3^2(t) + 3w_3(t)w_4^2(t) - w_4^3(t)}{48} \right) \qquad (361) \\
\dot{w}_2(t) &= \lambda \cdot \left(- \frac{w_4(t) - w_3(t)}{4} \right) - \mu w_2(t) \\
&+ \lambda \cdot \left(\frac{w_4^3(t) - 3w_3(t)w_4^2(t) + 3w_4(t)w_3^2(t) - w_3^3(t)}{48} \right) \qquad (362)
\end{aligned}$$

$$\begin{aligned}
\dot{w}_3(t) &= \frac{1}{\Delta}\Big(w_1(t) - w_1(t-\Delta) \Big) \qquad (363) \\
\dot{w}_4(t) &= \frac{1}{\Delta}\Big(w_2(t) - w_2(t-\Delta) \Big). \qquad (364)
\end{aligned}$$

Reduction to two cubic delay equations We will utilize the symmetry of the Equations (361) - (364) to simplify our problem by uncoupling the four equations. To do so we introduce a change of variables

$$\tilde{v}_1 = w_1 + w_2, \quad \tilde{v}_2 = w_1 - w_2, \quad \tilde{v}_3 = w_3 + w_4, \quad \tilde{v}_4 = w_3 - w_4. \qquad (365)$$

The expressions for variables $\tilde{v}_1$ and $\tilde{v}_3$ are uncoupled from $\tilde{v}_2$ and $\tilde{v}_4$

$$\dot{\tilde{v}}_1 = -\mu \tilde{v}_1(t), \qquad \dot{\tilde{v}}_3 = \frac{1}{\Delta}\Big(\tilde{v}_1(t) - \tilde{v}_1(t-\Delta) \Big), \qquad (366)$$

$$\dot{\tilde{v}}_2 = \lambda \left(- \frac{\tilde{v}_4(t)}{2} + \frac{\tilde{v}_4(t)^3}{24} \right) - \mu \tilde{v}_2(t), \qquad \dot{\tilde{v}}_4 = \frac{1}{\Delta}\Big(\tilde{v}_2(t) - \tilde{v}_2(t-\Delta) \Big). \qquad (367)$$

Furthermore, $\tilde{v}_1(t)$ and $\tilde{v}_3(t)$ can be solved directly and they converge to zero as $t \to \infty$. Hence we are left with only two functions of further interest, $\tilde{v}_2$ and $\tilde{v}_4$.

7.3 Limiting the Oscillations in Queues Through a Novel Type of Delay Announcement

Proof of Theorem 4.2. To check that $q_i(t) = \frac{\lambda}{N\mu}$ is an equilibrium, plug into Equation (95) to get

$$\dot{q}_i(t) = \lambda \cdot \frac{\exp(-\frac{\lambda\theta}{N\mu} - 0)}{\sum_{j=1}^{N} \exp(-\frac{\lambda\theta}{N\mu} - 0)} - \mu \cdot \frac{\lambda}{N\mu} = \frac{\lambda}{N} - \frac{\lambda}{N} = 0. \tag{368}$$

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

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

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

$$\bar{q}_s = \frac{\lambda}{\mu N} - \gamma, \qquad \bar{q}_b = \frac{\lambda}{N\mu} + \epsilon, \qquad \gamma, \epsilon > 0. \tag{369}$$

This leads to a contradiction:

$$\dot{\bar{q}}_s(t) = \lambda \frac{\exp\big(-\theta \bar{q}_s\big)}{\sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i\big)} - \mu \bar{q}_s = 0 \tag{370}$$

$$\implies \sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i\big) = \frac{\lambda}{\mu} \cdot \frac{\exp\big(-\frac{\theta\lambda}{N\mu} + \theta\gamma\big)}{(\frac{\lambda}{N\mu} - \gamma)}, \tag{371}$$

$$\dot{\bar{q}}_g(t) = \lambda \frac{\exp\big(-\theta \bar{q}_g\big)}{\sum_{i=1}^{N} \exp\big(-\theta \bar{q}_i\big)} - \mu \bar{q}_g(t) \tag{372}$$

$$= \lambda \frac{\exp\big(-\frac{\theta\lambda}{N\mu} - \theta\epsilon\big)}{\frac{\lambda}{\mu} \cdot \frac{\exp(-\frac{\theta\lambda}{N\mu} + \theta\gamma)}{(\frac{\lambda}{N\mu} - \gamma)}} - \mu\Big(\frac{\lambda}{N\mu} + \epsilon\Big) \tag{373}$$

$$= -\frac{\lambda}{N}\big(1 - e^{-\theta(\epsilon+\gamma)}\big) - \mu\big(\epsilon + \gamma e^{-\theta(\epsilon+\gamma)}\big) < 0. \tag{374}$$

Since $\dot{\bar{q}}_g(t) \neq 0$, then $\bar{q}_i(t)$ is not an equilibrium, and the equilibrium (98) is unique. □

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 Section 4.5, 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 determing the second-order approximation are outlined below.

This process is very closely related to the staps taken in Theorem 4.14.

We begin with Equation (162), and expand the time $\tau = \omega t$. Then expand our functions of interest in ϵ to the second order:

$$x(\tau) = x_0(\tau) + \epsilon x_1(\tau) + \epsilon^2 x_2(\tau),$$
$$\Delta = \Delta_0 + \epsilon\Delta_1 + \epsilon^2\Delta_2,$$
$$\omega = \omega_0 + \epsilon\omega_1 + \epsilon^2\omega_2,$$

where Δ_0 and ω_0 are the delay and frequency at bifurcation, so $\Delta_0 = \Delta_{cr}$ 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 Equation (193). 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 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

$$x(\tau) \approx x_0(\tau) + \epsilon x_1(\tau) \qquad (375)$$
$$= A\cos(\tau) + \epsilon\big(a_1\sin(\tau) + a_2\cos(\tau) + a_3\sin(3\tau) + a_4\cos(3\tau)\big), \qquad (376)$$

where the coefficients given below:

$$A = \sqrt{\frac{4\Delta_1(\lambda^2\theta^2 - 4\mu^2)(4 - \delta^2\lambda^2\theta^2)^2}{\theta^2(1-\delta^2\mu^2)(16\mu + \lambda^2\theta^2(4\Delta_0 - 4\delta + \delta^3\lambda^2\theta^2 - 4\delta^2\mu - 4\delta^2\Delta_0\mu^2))}}$$

$$\omega_1 = \frac{4\Delta_1\theta^2\lambda^2\left(\delta^2\mu^2 - 1\right)\sqrt{\theta^2\lambda^2 - 4\mu^2}}{\sqrt{4 - \delta^2\theta^2\lambda^2}\left(\theta^2\lambda^2\left(\delta\left(\delta^2\theta^2\lambda^2 - 4\delta\mu(\Delta 0\mu + 1) - 4\right) + 4\Delta 0\right) + 16\mu\right)}$$

$$\begin{aligned} a_1 = -3a_3 = -\Big(2A^3\theta^2\omega_0^3\left(\theta^2\lambda^2\mu\left(\delta^2\omega_0^2 + 1\right)^3 - 4\delta^3\left(\mu^2 + \omega_0^2\right)^3\right)\Big) \\ /\Big(\theta^4\lambda^4\left(\delta^2\omega_0^2 + 1\right)^3\left(\mu^2 + 9\omega_0^2\right) + 16\left(9\delta^2\omega_0^2 + 1\right)\left(\mu^2 + \omega_0^2\right)^3 \\ +8\theta^2\lambda^2\big(-9\delta^4\omega_0^8 - 6\mu^2\omega_0^2\left(\delta^2\mu^2 + 1\right) + 2\delta^2\omega_0^6(\delta\mu(9\delta\mu - 32) + 9) \\ +3\omega_0^4\left(\delta^4\mu^4 - 12\delta^2\mu^2 + 1\right) - \mu^4\big)\Big) \end{aligned}$$

$$\begin{aligned} a_4 = -\frac{1}{12}\Big(A^3\theta^2\Big(\theta^2\lambda^2\left(\delta^2\omega_0^2 + 1\right)^3\left(\mu^4 + 6\mu^2\omega_0^2 - 3\omega_0^4\right) \\ +4\left(3\delta^4\omega_0^4 - 6\delta^2\omega_0^2 - 1\right)\left(\mu^2 + \omega_0^2\right)^3\Big)\Big) \end{aligned}$$

$$\begin{aligned} \cdot\Big(\theta^4\lambda^4\left(\delta^2\omega_0^2 + 1\right)^3\left(\mu^2 + 9\omega_0^2\right) + 16\left(9\delta^2\omega_0^2 + 1\right)\left(\mu^2 + \omega_0^2\right)^3 \\ +8\theta^2\lambda^2\big(-9\delta^4\omega_0^8 - 6\mu^2\omega_0^2\left(\delta^2\mu^2 + 1\right) + 2\delta^2\omega_0^6(\delta\mu(9\delta\mu - 32) + 9) \\ +3\omega_0^4\left(\delta^4\mu^4 - 12\delta^2\mu^2 + 1\right) - \mu^4\big)\Big)^{-1} \end{aligned}$$

$$a_2 = \frac{1}{12}\Big(A^5\theta^4\left(\delta^2\omega_0^2+1\right)^2\left(\delta^2\mu\omega_0^2+\mu^2\left(\delta\left(\delta\Delta_0\omega_0^2-1\right)+\Delta_0\right)\right.$$
$$\left.+\omega_0^2\left(\delta\left(\delta\Delta_0\omega_0^2-1\right)+\Delta_0\right)+\mu\right.$$

$$-12A^3\theta^2\omega_0\big(\omega_1\left(\delta\left(3\delta^3\Delta_0\omega_0^4+\delta\omega_0^2(\delta(\delta\mu(2\Delta_0\mu+3)-3)+4\Delta_0)\right.\right.$$
$$\left.\left.+\delta\mu(-2\delta\mu+2\Delta_0\mu+3)-1+\Delta_0-\Delta_1\omega_0\left(\delta^2\mu^2-1\right)\left(\delta^2\omega_0^2+1\right)\right)\right.$$

$$+12A^2\theta^2\big(a_1\omega_0\left(\delta^2\mu^2-1\right)\left(\delta^2\omega_0^2+1\right)$$
$$+a_3\omega_0\left(\delta^2\left(\omega_0^2(\delta\mu(-3\delta\mu+8\Delta_0\mu+8)-5)+8\delta\Delta_0\omega_0^4+\mu^2\right)-1\right)$$
$$+a_4(3\delta^4\Delta_0\omega_0^6+3\delta^2\omega_0^4(\delta(\delta\mu(\Delta_0\mu+1)-1)-2\Delta_0)$$
$$+\omega_0^2(\delta(\delta\mu(5\delta\mu-6\Delta_0\mu-6)+1)-\Delta_0)+\mu(\delta\mu-\Delta_0\mu-1)))\big)$$
$$-96A\big(2\Delta_1\omega_0\omega_1(\mu\left(\mu\left(2\delta^2-2\delta\Delta_0+\Delta_0^2\right)-\delta+\Delta_0\right)+\delta^2\Delta_0^2\omega_0^4$$
$$+\Delta_0\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-2)+\Delta_0)-1)$$
$$+\Delta_0\omega_1^2(\delta^2\Delta_0^2\omega_0^4+\Delta_0\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-3)+\Delta_0)$$
$$+\mu(2\delta-\Delta_0)(\delta\mu-\Delta_0\mu-1))+\Delta_1^2\omega_0^2(\delta^2\Delta_0\omega_0^4+\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-1)+\Delta_0)$$
$$+\mu(-\delta\mu+\Delta_0\mu+1)))\big)-192a_1\big(\omega_1(\delta^2\Delta_0^2\omega_0^4+\Delta_0\omega_0^2(\delta(\delta\mu(\Delta_0\mu+2)-2)+\Delta_0)$$
$$+(-\delta\mu+\Delta_0\mu+1)^2)+\Delta_1\omega_0\left(\delta^2\Delta_0\omega_0^4+\mu(-\delta\mu+\Delta_0\mu+1)\right.$$
$$\left.+\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-1)+\Delta_0))\big)\right)$$

$$/\Big(3A^2\theta^2\left(\delta^2\omega_0^2+1\right)\left(\delta^2\Delta_0\omega_0^4+\omega_0^2(\delta(\delta\mu(\Delta_0\mu+1)-1)+\Delta_0)\right)$$
$$+3A^2\theta^2\left(\delta^2\omega_0^2+1\right)(\mu(-\delta\mu+\Delta_0\mu+1))+16\Delta_1\omega_0^2\left(\delta^2\mu^2-1\right)\Big).$$

To reproduce our numerical results from Section 4.5 - (4.5), set $\epsilon = 1$ and $\Delta_1 = \frac{1}{\epsilon}(\Delta - \Delta_0)$, with Δ_0 given by Equation (139). Note that in the equations above there is no presence of Δ_2, because we have 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.