Queues with Delayed Information: A Probabilistic Perspective

Sophia Novitzky
Center for Applied Mathematics
Cornell University
657 Rhodes Hall, Ithaca, NY 14853
sn574@cornell.edu

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

July 10, 2020

Abstract

Many service systems use internet or smartphone app technology to notify customers about their expected waiting times or queue lengths via delay announcements. However, in many cases, either the information might be delayed or customers might require time to travel to the queue of their choice, thus causing a lag in information. The previous literature has analyzed these systems where the lag in information is a fixed constant Δ; however, in this work, we generalize the previous work by allowing the delay to be a random variable with a fixed probability distribution. Using Jensen's inequality, we prove the constant delay is the most unstable distribution for a given fixed mean delay. Moreover, in the setting where the Laplace transform is unknown, we provide a Taylor series approach that approximates the Hopf curves using the central moments of the distribution. We prove for some distributions, that these central moment approximations converge uniformly and monotonically to the true Hopf curves. Finally, we show an equivalence between delay differential equations with multiple delays to distributed delay equations with a discrete distributions. Thus, we show the power of our probabilistic perspective to solve open questions in the delay differential equations literature. Although our methodology is applied to models from queueing theory, our results are of general interest to anyone interested in distributed delay equations and immediately generalize to other application domains as well.

Keywords: delay-differential equations, distributed delay equations, Hopf bifurcation, perturbations method, operations research, queueing theory, fluid limits, delay announcements, delayed information

AMS subject classifications: 34K40, 34K18, 41A10, 37G15, 34K27

1 Introduction

With more access to information, consumers are adjusting their behavior and expectations of the services they select. More and more customers are actively seeking information about competing businesses prior to choosing to receive service from a provider. Businesses that boost short waiting times for service can attract potential clients to their service, however, if the wait is too long, this can instead deter customers from joining the queue. 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). Services like the bike sharing networks, U-Haul truck rental locations, hospital emergency rooms, amusement parks like Disney World, and even restaurants, are often ready to provide such information to potential customers. The phone application CycleFinder in Figure 1, for example, provides customers a map of the bike-sharing racks and the number of available bikes at each location. This helps bicyclists not waste their time checking empty stations or stations with a low number of bikes.

Figure 1: Bike-sharing network app.

The availability of the waiting time information impacts the decision patterns of individual customers and the dynamics of the queueing system as a whole. In a multi-dimensional system where each queue corresponds to a separate geographical location of the same service (such as competing restaurants in a neighborhood, for example), customers can choose which queue to join giving a higher preference to the location with the shortest waiting time. Previous works have modeled this choice using choice models from the economics literature, see for example McFadden [22], Marshall et al. [21], Train [35], Tao and Pender [34]. However, many of these papers do not analyze the impact of delaying the information to customers. This delayed information has two forms. The first type occurs when the customer commits to a queue before joining. The second type occurs when there is a physical lag in the information, for example when the information may need time to be processed correctly.

In Figure 2, we describe a possible situation of joining several queues where the customer must commit before arriving. The customer, Mr. Green, in Figure 2, has to commute to the service location. Mr. Green's commute causes a time delay prior him securing his physical spot in the queue. 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. The waiting time information used by customers is therefore somewhat outdated and unreliable, causing ripple effects throughout the system. For example, it has been shown in Pender et al. [28, 30], Novitzky et al. [26] that when all customers experience the same constant 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 enough, the queues will stop oscillating over time and converge to an equilibrium queue length [29].

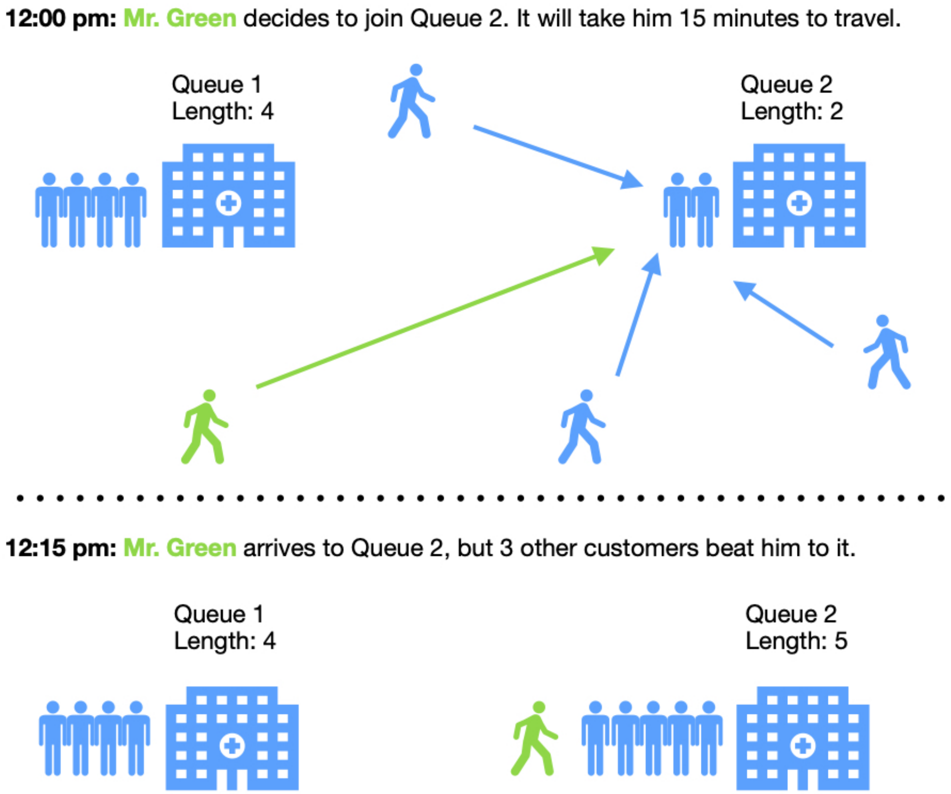
12:00 pm: Mr. Green decides to join Queue 2. It will take him 15 minutes to travel.
Queue 1
Length: 4
Queue 2
Length: 2
12:15 pm: Mr. Green arrives to Queue 2, but 3 other customers beat him to it.
Queue 1
Length: 4
Queue 2
Length: 5

Figure 2: Mr. Green tries to join the shortest queue.

Accounting for the delay is clearly an important research question that deserves much attention, however, it is still an open question of how to do so accurately. Moreover, in most physical systems, each customer takes a slightly different time to arrive to the queue of their choice as seen in Figure 3. Thus, it is important to understand how the randomness of the travel time to each queue or randomness in the delay might impact the underlying dynamics of the queue length process. In this paper, we analyze queueing systems where the individual's travel time is modeled as a random variable from a known distribution. This allows us to study the impact that the distribution of the delay has on the queueing system dynamics.

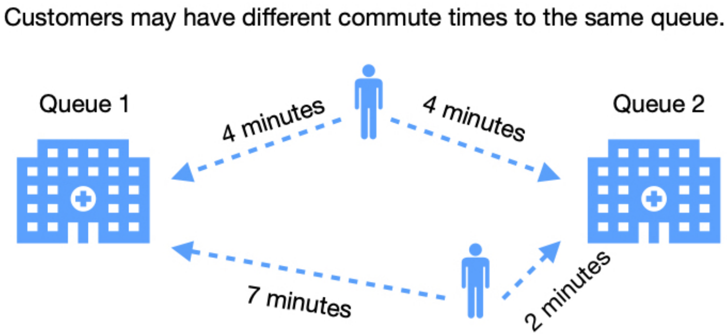
Customers may have different commute times to the same queue.
Queue 1
Queue 2
4 minutes
4 minutes
7 minutes
2 minutes

Figure 3: An individual's commute time can be modelled as a random variable.

1.1 Contributions of the Paper

- We formulate a new queueing model that captures customers' preference for shorter waiting times, as well as the delay in information caused by randomized travel time of the customers to the queues.

- We study the asymptotic behavior of the generalized queueing system. In particular, we show 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 parameter is met. The equilibrium can become unstable only if a Hopf bifurcation occurs.

- We derive the characteristic equation for this generalized model and use it to determine the stability of the queueing systems for several specific distributions of the delay.

- In the case where the delay distribution might be unknown or difficult to calculate, we develop a novel approximation technique based on 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. We also show that central moment approximations are not equivalent to raw moment approximations.

- Finally, our unique probabilistic perspective allows us to make an equivalence between multi-delay systems and distributed delay systems with discrete distributions. This equivalence is important for analysis of multi-delay systems, which has been intractable until now.

1.2 Organization of Paper

We begin by reviewing the relevant literature in Section 2. The generalized queueing model is presented in Section 3, 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 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 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.

2 Literature Review

In this section, we provide a review of the literature that is relevant to this work as the delayed information space is relatively new in the context of this work. We highlight work that is not only relevant from a queueing perspective, but also literature that has explored distributed delay differential equations. There are several papers on distributed delay differential equations and we describe how our work is different and novel.

2.1 Delayed Information and Queueing Theory

There are several papers by the authors such as Pender et al. [29, 30], Novitzky et al. [25], which have analyzed a similar model to the one presented in this paper. However, in each of those papers, the delay is a constant and is not a non-degenerate random variable like in this work. The first paper, Pender et al. [29] considered a two dimensional fluid model and derived an explicit formula for the Hopf bifurcation under this constant delay setting. The second paper Pender et al. [30] also analyzed a similar model to the first paper, however, the second paper added the complexity of time varying arrival rates. This is a significant difference since non-stationary arrival rates are much more complicated than their stationary counterparts. One must distinguish the Hopf bifurcation oscillations from the time varying dynamics of the arrival rate. We do not consider time varying arrival rates in this work. Thus, the problem with random delay distributions and non-stationary arrival rates is still an open problem for research.

A third paper Novitzky et al. [25] develops a statistical method to compute the amplitude of the oscillations generated by the Hopf bifurcations. This method is called the ”Slope Function Method” and the main idea is to use non-linear regression to learn the amplitude from numerically integrating a few ddes and using their amplitudes as the data. It also analyzes the constant delay model in the N-dimensional case, yielding a Hopf bifurcation formula for N queues. Finally, Novitzky et al. [26] develops a new delay announcement by incorporating the velocity of the queue length into the delay announcement. Novitzky et al. [26] shows that the velocity can help reduce the size of the oscillations created by the Hopf bifurcation or can eliminate them altogether. However, Novitzky et al. [26] also assumes that the delay is a constant and not a random variable as in this work.

In addition to work by the authors there are a few papers that also explore the impact of delayed information in the context of queueing systems. The first paper by Lipshutz and Williams [19] derives sufficient conditions for when oscillations will occur in reflected delay differential equations when oscillations are present in non-reflected delay differential equations. A second paper by Raina and Wischik [32] incorporates concepts from queueing theory with delay differential equations and applies them to sizing router buffers in internet infrastructure services. Raina and Wischik [32] uses a common technique named Lindstedt's method to construct estimates for the amplitude of oscillations created by Hopf bifurcations. Finally, a recent paper by Lipshutz [20] proves heavy traffic limit theorems for queues with delays in information. They show that they can eliminate oscillations by keeping track of arrivals to each queue. This work is similar in spirit to Pender et al. [28], which proves fluid and diffusion limit theorems for queues with delayed information.

Delayed information has also been studied empirically as well. In work by and Dong et al. [9], the authors show that oscillations are present when information is given to customers via mobile apps. Moreover, in work by Nirenberg et al. [24], the authors explore applications of delayed information to amusement park queues and constructs a novel model where the information itself is also rounded. Nirenberg et al. [24] shows that oscillations can result either from delayed information or the rounding of the information as it is passed to consumers. They also show that the mobile app induces oscillations in the real queue length data. However, a common theme in all of the above papers is that the delay is a constant and is not a random variable. We will show in the sequel that the delay being a random variable is a significant challenge and is a much more difficult problem.

2.2 Distributed Delay Equations Literature

In addition to the work that has been applied on queueing theory and delayed information there is also research that has explored the impact of the delay distribution on the delay differential equations. For example in Bernard et al. [2], Braverman and Zhukovskiy [3], Kiss and Krauskopf [17], Calleja et al. [5], Cooke and Grossman [7], Campbell and Jessop [6], Breda et al. [4], Cuvas and Mondié [8], Bélair and Campbell [1], Yuan and Bélair [36], Morărescu et al. [23], Rahman et al. [31], the authors study distributed delay equations. Most of the applications are in biological settings and explore epidemic models or disease models.

The work that has the most similarity to ours is Bernard et al. [2] and Yuan and Bélair [36]. In these papers, the authors attempt to explore the impact that the probability distribution has on Hopf bifurcations and the underlying dynamics of the system. Unlike their work, our work takes a purely probabilistic perspective on the distributed delay differential equations. We also develop a novel Hopf bifurcation approximation approach based on a centered moment expansion of the Laplace transform of the delay distribution. We will show in this paper that our expansion is quite accurate and provides a way to approximate Hopf bifurcations for distributed delay equations. Finally, we also show how recast delay differential equations with multiple delays as a distributed delay equation where the probability distribution is a discrete distribution. This perspective is new and provides a new way of analyzing delay differential equations with multiple delays.

3 The Randomly Delayed Queueing Model

In this section, we describe the queueing model with randomly delayed information that we intend to analyze. In Figure 4, we provide an illustration of the structure of the N-dimensional queueing system. Customers arrive to the queue 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 [12, 18, 11, 14, 16, 33]. Although an infinite server queue has no wait in reality, we consider this model because it actually suffices to study this model for more complicated queues like the Erlang-A queueing model, which do have waiting times. For a more detailed explanation of this reduction to infinite server analysis, see Section 2 of Novitzky et al. [26]. We also think this model is relevant because in reality, the customers have no idea of how the queue is implemented and the linear departure rate is the simplest non-trivial model to consider. With these assumptions, the queue lengths can be described by the following system of N functional differential equations

$$\overset{\bullet}{q}_i(t) = \lambda p_i(q_1, \dots, q_N) - \mu q_i(t), \quad \forall i \in \{1, ..., N\}, \tag{3.1}$$

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

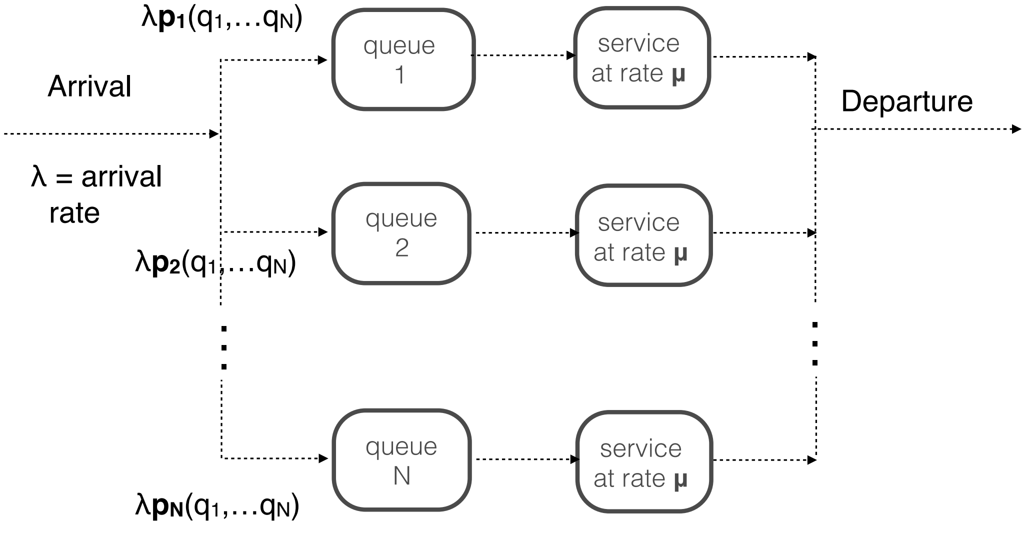

Figure 4: Customers going through a N-queue service system.

Here we assume that all queues offer identical service, but the queue length reported may differ depending on the number of customers in each 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 choice model (MNL) Train [35], which is a common choice model in a variety of application domains. A customer who appears at time $t - X$ will chose a queue, and then commuted to it for the time X, and therefore at time t has the following probability of joining the i^{th} queue.

$$\text{Probability of joining the } i^{th} \text{ queue} = \frac{\exp(-\theta q_i(t-X))}{\sum_{j=1}^{N} \exp(-\theta q_j(t-X))}. \tag{3.2}$$

As desired, the probability that they join any of the N queues is $\sum_{i=1}^{N} \frac{\exp(-\theta q_i(t-X))}{\sum_{j=1}^{N} \exp(-\theta q_j(t-X))} = 1$. 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 $f(s)$, the proportion of all customers who arrive to the i^{th} queue at time t is given by the following equation

$$p_i(q_1, \ldots, q_N) = \frac{\exp\left(-\theta \int_0^\infty q_i(t-s) f(s) ds\right)}{\sum_{j=1}^{N} \exp\left(-\theta \int_0^\infty q_j(t-s) f(s) ds\right)}. \tag{3.3}$$

Thus, our queueing model described in Equation (3.1) solves the following system of functional differential equations

$$\overset{\bullet}{q}_i(t) = \lambda \cdot \frac{\exp\left(-\theta \int_0^\infty q_i(t-s) f(s) ds\right)}{\sum_{j=1}^{N} \exp\left(-\theta \int_0^\infty q_j(t-s) f(s) ds\right)} - \mu q_i(t), \quad \forall i \in \{1, ..., N\}. \tag{3.4}$$

3.1 Equilibria and Stability

Now that we have a system of functional differential equations that describes our queueing system, many interesting questions emerge. For example, we would like to know about the equilibrium of Equation 3.4, is the equilibrium unique, and when is the queueing system stable? One important observation about Equation 3.4 is that we can analyze many of these questions without restricting our model to specific probability distribution functions $f(s)$. For example, we show in Theorem 3.1 that regardless of the delay distribution, there is a unique equilibrium state. For convenience, we will reformulate Equation (3.4) as

$$\overset{\bullet}{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta g(q_i(t))\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(q_j(t))\big)} - \mu q_i(t), \tag{3.5}$$

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

Theorem 3.1. *The unique equilibrium to the system of equations* (3.5) *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 $\overset{\bullet}{q}_i(t) = 0$ 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

$\sum_{i=1}^{N} \overset{\bullet}{q}_i = 0 = \lambda - \mu \sum_{i=1}^{N} \bar{q}_i$, 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 at least two indices k and l, $1 \le k, l \le N$, without loss of generality the following inequalities below must hold

$$\bar{q}_k < \frac{\lambda}{N\mu}, \quad \text{and} \quad \bar{q}_l > \frac{\lambda}{N\mu}. \tag{3.6}$$

We define $\bar{q}_k$ to be the minimum equilibrium value and define $\bar{q}_l$ to be the maximum equilibrium value. Because the function g is monotonically increasing, $-g(\bar{q}_k) > -g(\bar{q}_l)$ and

$$\frac{\exp\big(-\theta g(\bar{q}_l)\big)}{\exp\big(-\theta g(\bar{q}_k)\big)} < 1. \tag{3.7}$$

Further, $\overset{\bullet}{q}_k(t) = 0$ so

$$\begin{aligned} \overset{\bullet}{q}_k(t) = \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_k)\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big)} - \mu \bar{q}_k(t) = 0 && (3.8) \\ \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_k)\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big)} = \mu \bar{q}_k(t) < \frac{\lambda}{N} && (3.9) \\ \sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big) > N \exp\big(-\theta g(\bar{q}_k)\big). && (3.10) \end{aligned}$$

Finally, we use inequalities from Equation (3.7) and Equation (3.10) to show that $\overset{\bullet}{q}_l(t) \neq 0$:

$$\begin{aligned} \overset{\bullet}{q}_l(t) &= \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_l)\big)}{\sum_{j=1}^{N} \exp\big(-\theta g(\bar{q}_j)\big)} - \mu \bar{q}_l(t) && (3.11) \\ &< \lambda \cdot \frac{\exp\big(-\theta g(\bar{q}_l)\big)}{N \exp\big(-\theta g(\bar{q}_k)\big)} - \frac{\lambda}{N} < \frac{\lambda}{N} - \frac{\lambda}{N} = 0. && (3.12) \end{aligned}$$

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

3.1.1 Finding the characteristic equation

Now that we know that the queueing system has a unique equilibrium, we can use this to study the stability of this equilibrium. Our analysis about the stability of the equilibrium can help us understand the connection between the shape of the delay distribution and whether or not the queueing network will oscillate indefinitely i.e. have a Hopf bifurcation. We start with trying to determine when the queueing system is locally stable. To this end, we need to derive the characteristic equation for the linearized version of the queueing network's functional differential system of equations. In order to do this, we will find an alternative way to write the functional differential equations. Using some auxiliary variables, we can write the system of functional differential equations as the following $2N$-dimensional system

$$\begin{aligned} \overset{\bullet}{q}_i(t) &= \lambda \cdot \frac{\exp\big(-\theta m_i(t)\big)}{\sum_{j=1}^N \exp\big(-\theta m_j(t)\big)} - \mu q_i(t) && (3.13) \\ m_i(t) &= \int_0^\infty q_i(t-s) f(s) ds. && (3.14) \end{aligned}$$

Now by linearizing the above system of equations, we arrive at the following linearized system of equations

$$\begin{aligned} \overset{\bullet}{q}_i(t) &\approx -\frac{\lambda\theta}{N} m_i(t) + \frac{\lambda\theta}{N^2} \sum_{j=1}^N m_j(t) - \mu q_i(t) && (3.15) \\ \overset{\bullet}{m}_i &= \frac{d}{dt} \int_0^\infty q_i(t-s) f(s) ds, && (3.16) \end{aligned}$$

which can be expressed in a vector form as

$$\begin{aligned} \overset{\bullet}{\bar{q}}(t) &= -\frac{\lambda\theta}{N} \bar{m}(t) + \frac{\lambda\theta}{N^2} A \bar{m}(t) - \mu \bar{q}(t) && (3.17) \\ \overset{\bullet}{\bar{m}} &= \frac{d}{dt} \int_0^\infty \bar{q}(t-s) f(s) ds, && (3.18) \end{aligned}$$

where $\bar{q} = [\bar{q}_1, \ldots, \bar{q}_N]^T \in \mathbb{R}^N$, $\bar{m} = [\bar{m}_1, \ldots, \bar{m}_N]^T \in \mathbb{R}^N$, and $A \in \mathbb{R}^{N \times N}$ with $A_{ij} = 1$ for $1 \leq i, j \leq N$. The matrix can be diagonalized,

$$A = VDM, \quad \text{where } D, M, V \in \mathbb{R}^{N \times N} \quad \text{and } D_{ij} = \begin{cases} 1 & i = j = 1 \\ 0 & \text{otherwise} \end{cases}. \qquad (3.19)$$

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

$$\begin{aligned} V\overset{\bullet}{\bar{w}}(t) &= -\frac{\lambda\theta}{N} V\bar{u}(t) + \frac{\lambda\theta}{N^2} VD\bar{u}(t) - \mu V\bar{w}(t) && (3.20) \\ V\overset{\bullet}{\bar{u}}(t) &= \frac{d}{dt} \int_0^\infty V\bar{w}(t-s) f(s) ds. && (3.21) \end{aligned}$$

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

$$\begin{aligned} \overset{\bullet}{\bar{w}}(t) &= -\frac{\lambda\theta}{N} \bar{u}(t) + \frac{\lambda\theta}{N^2} D\bar{u}(t) - \mu \bar{w}(t) && (3.22) \\ \overset{\bullet}{\bar{u}}(t) &= \frac{d}{dt} \int_0^\infty \bar{w}(t-s) f(s) ds. && (3.23) \end{aligned}$$

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

$$\begin{aligned} \overset{\bullet}{w}_1(t) &= -\mu w_1(t) && (3.24) \\ \overset{\bullet}{w}_i(t) &= -\frac{\lambda\theta}{N} u_i(t) - \mu w_i(t), \quad i = 2, 3, \ldots, N && (3.25) \\ \overset{\bullet}{u}_i(t) &= \frac{d}{dt} \int_0^\infty w_i(t-s) f(s) ds, \quad i = 1, 2, \ldots, N && (3.26) \end{aligned}$$

All w_i functions are now uncoupled, and w_1 from Equation (3.24) 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 (3.25) yields the following characteristic equation

$$\Phi(r, \Delta) = r + \mu + \frac{\lambda\theta}{N} F(r) = 0. \tag{3.27}$$

It is important to recognize that our linear algebra of reducing the N-dimensional system to one eigenvalue equation is quite important. This reduction in dimensions will significantly simplify our analysis going forward since it reduces an N-dimensional problem to a one dimensional one. Moreover, understanding the roots of Equation 3.27 is critical to our understanding of the stability of the functional differential system. It is clear that when the Laplace transform of the probability distribution $F(r)$ changes, the stability of the system will change as well. Thus, it is important to understand the impact of the Laplace transform $F(r)$ on the stability of the queueing system. It turns out that 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 3.2. *If $\lambda\theta < N\mu$, then the equilibrium from Theorem* (3.1) *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 $a, b \in \mathbb{R}$.

$$\Phi(r, \Delta) = a + ib + \mu + \frac{\lambda\theta}{N} \int_0^\infty e^{-as}\big(\cos(bs) - i\sin(bs)\big) f(s)ds = 0. \tag{3.28}$$

Separating the real and imaginary parts we arrive at two equations

$$a + \mu + \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \cos(bs) f(s)ds = 0 \tag{3.29}$$

$$b - \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \sin(bs) f(s)ds = 0. \tag{3.30}$$

To reach a contradiction, let us suppose that $\lambda < N\mu/\theta$ and there exists $a \geq 0$ that satisfies Equations (3.29) - (3.30). Since $\int_0^\infty f(s)ds = 1$, then

$$\left|\int_0^\infty e^{-as} \cos(bs) f(s)ds\right| \leq \left|\int_0^\infty e^{-as} f(s)ds\right| \leq \left|\int_0^\infty f(s)ds\right| = 1. \tag{3.31}$$

From Equation (3.29) it then follows that

$$a = -\mu - \frac{\lambda\theta}{N} \int_0^\infty e^{-as} \cos(bs) f(s)ds \tag{3.32}$$

$$\leq -\mu + \frac{\lambda\theta}{N} < -\mu + \frac{N\mu\theta}{N\theta} = 0, \tag{3.33}$$

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 (3.1) 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 3.3. *Any real eigenvalue of the characteristic equation given in Equation* (3.27) *is negative.*

Proof. Suppose $r \in \mathbb{R}$. Then $F(r) = \int_0^\infty e^{-rs} f(s) ds \geq 0$, and from Equation (3.27) it follows that r must be negative

$$r = -\mu - \frac{\lambda\theta}{N} F(r) \leq -\mu < 0. \tag{3.34}$$

□

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 3.4. *If one pair of eigenvalues is purely imaginary, a Hopf bifurcation occurs as λ increases or as μ decreases.*

Proof. The infinite-dimensional version of the Hopf Theorem of Hale and Lunel [15] 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, let us assume that $b(\lambda^*, \mu^*)$ is positive. The equations (3.29)-(3.30) at (λ^*, μ^*) become

$$\mu^* + \frac{\lambda^*\theta}{N} \int_0^\infty \cos(bs) f(s) ds = 0 \tag{3.35}$$

$$b - \frac{\lambda^*\theta}{N} \int_0^\infty \sin(bs) f(s) ds = 0. \tag{3.36}$$

To show that the third condition of the Hopf Theorem of [15] is fulfilled, we differentiate Equation (3.29) with respect to λ:

$$\frac{da}{d\lambda} + \frac{\theta}{N}\int_0^\infty e^{-as}\cos(bs)f(s)ds + \frac{\lambda\theta}{N}\cdot\frac{d}{d\lambda}\int_0^\infty e^{-as}\cos(bs)f(s)ds = 0. \tag{3.37}$$

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

$$\frac{da}{d\lambda} + \frac{\theta}{N}\int_0^\infty e^{-as}\cos(bs)f(s)ds + \frac{\lambda\theta}{N}\int_0^\infty (-s)e^{-as}\cos(bs)f(s)ds\cdot\frac{da}{d\lambda} = 0. \tag{3.38}$$

When $(\lambda,\mu) = (\lambda^*,\mu^*)$ we recall that $a = 0$ and by Equation (3.35) we find $\int_0^\infty \cos(bs)f(s)ds = -\frac{N\mu^*}{\lambda^*\theta}$. Substituting this into (3.38)

$$\frac{d}{d\lambda}a(\lambda^*,\mu^*) - \frac{\mu^*}{\lambda^*} - \frac{\lambda^*\theta}{N}\int_0^\infty s\cos(bs)f(s)ds\cdot\frac{d}{d\lambda}a(\lambda^*,\mu^*) = 0 \tag{3.39}$$

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

$$\frac{d}{d\lambda}\text{Re}[r(\lambda^*,\mu^*)] = \frac{d}{d\lambda}a(\lambda^*,\mu^*) \neq 0, \tag{3.40}$$

so the third condition of the Hopf Theorem holds with respect to the parameter λ. Analogously, it can be shown that $\frac{d}{d\mu}\text{Re}[r(\lambda^*,\mu^*)] \neq 0$ as well. □

For a fixed delay distribution, as the arrival rate of customers changes, the queueing system can qualitatively change its behavior. When the condition ($\lambda\theta < N\mu$) holds, the queue will stabilize over time and all queues will converge to the equilibrium queue length of $\frac{\lambda}{N\mu}$. However, when the condition does not hold i.e. $\lambda\theta \geq N\mu$, 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 probability distributions and discuss how the queueing system behaves in each case. In the cases that are possible, we describe exactly when a Hopf bifurcation will occur in terms of the model parameters.

4 Explicit Analysis for Common Delay Distributions

Queueing systems may exhibit different behavior depending on the distribution of the delay. Based on the results from Section 3.1, any queueing system of the form given by Equation (3.4) is guaranteed to be stable unless a Hopf bifurcation occurs. Our goal for this section is to do an extensive analysis on some common distributions for the time delay and determine under what conditions (if ever) the resulting queueing systems become unstable. We will begin by reviewing the simplest possible distribution, and then will gradually introduce more complexity to get a better sense of how the distribution of the delay affects the underlying dynamics of the queueing system.

4.1 Constant Delay

For our first model, we will suppose that there is no variation in the commute time of the customers. In this example, they all travel from the same population center to the queue of their choice, where the queues are equidistant from the customers' initial location. Figure 5 provides a geographical representation of such a queueing system. We will refer to this system as the *constant delay* model since each customer's delay is is given by a constant Δ.

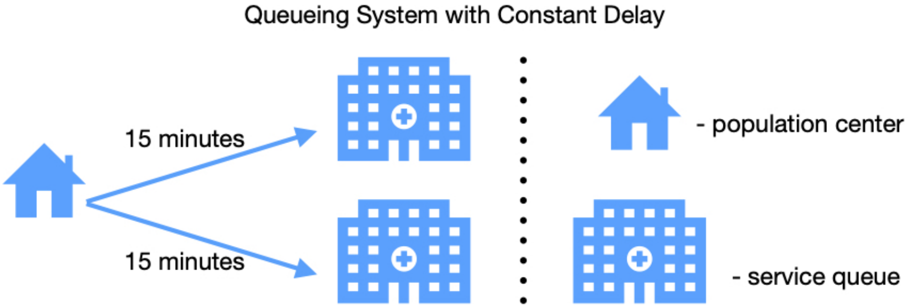

Figure 5: All customers have the same commute time.

The system of differential equations from Equation (3.4) simplifies to

$$\overset{\bullet}{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta q_i(t-\Delta)\big)}{\sum_{j=1}^{N}\exp\big(-\theta q_j(t-\Delta)\big)} - \mu q_i(t) \qquad \forall i = 1,2,\ldots,N, \tag{4.41}$$

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 given by Equation (3.27) becomes

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N}\exp(-r\Delta) = 0. \tag{4.42}$$

The queueing system in Equation (4.41) undergoes a Hopf bifurcation and the point of bifurcation can be found analytically and in closed form. Like in Novitzky et al. [25], we set the eigenvalue to be purely imaginary i.e. $r = ib$ and separate the real and imaginary parts of the characteristic equation to find

$$\mu + \frac{\lambda\theta}{N}\cos(b\Delta) = 0, \quad b - \frac{\lambda\theta}{N}\sin(b\Delta) = 0. \tag{4.43}$$

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

$$b = \sqrt{\frac{\lambda^2\theta^2}{N^2} - \mu^2}, \qquad \Delta_{cr} = \frac{N\arccos(-N\mu/\ (\lambda\theta))}{\sqrt{\lambda^2\theta^2 - N^2\mu^2}}. \tag{4.44}$$

By Theorem (3.4), the queues undergo a Hopf bifurcation when the equations (4.44) hold. In the parameter space λ and Δ, the Hopf curve is given by Figure 6. The queues are stable in the region to the left of the Hopf curve, and unstable in the region to the right. However, this model has been analyzed before in Novitzky et al. [25]. We now will try to analyze a more complex model that captures two delays.

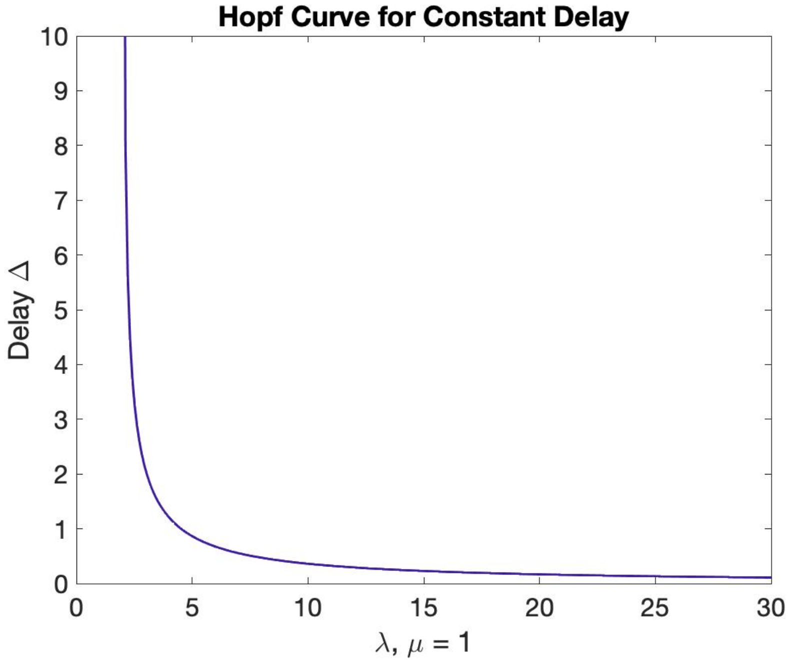

Figure 6: Hopf curve given a constant delay.

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 now 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 7 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.

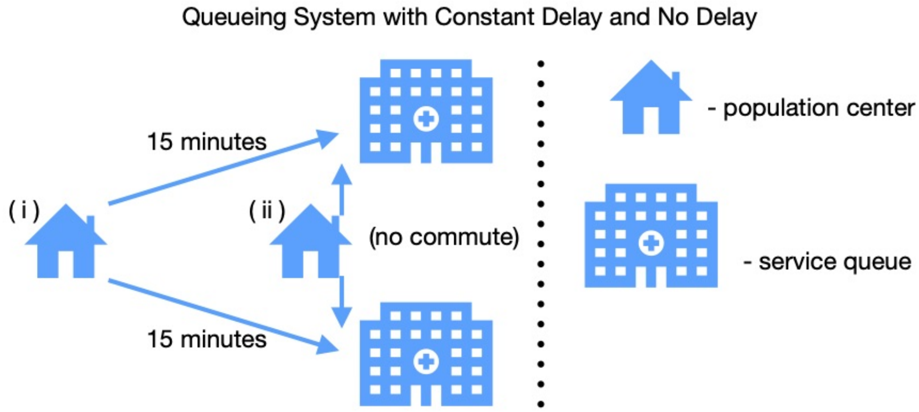

Figure 7: Customers experience either a constant delay or no delay, depending on their initial location.

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

$$\overset{\bullet}{q}_i(t) = \lambda \cdot \frac{\exp\big(-\theta(1-p)q_i(t) - \theta p q_i(t-\Delta)\big)}{\sum_{j=1}^{N} \exp\big(-\theta(1-p)q_j(t) - \theta p q_j(t-\Delta)\big)} - \mu q_i(t), \quad i = 1, 2, \ldots, N. \quad (4.45)$$

Here the delay distribution is a linear combination of Dirac delta functions, $f(s) = p\delta(s - \Delta) + (1-p)\delta(s)$. The Laplace transform of the delay distribution is therefore $F(r) = p\exp(-r\Delta) + (1-p)$, and the characteristic equation (3.27) becomes

$$\Phi(r) = r + \mu + \frac{\lambda\theta p}{N}\exp(-r\Delta) + \frac{\lambda\theta}{N}\cdot(1-p) = 0. \quad (4.46)$$

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 4.1. *When the queueing model has the dynamics of Equation 4.45, then we have the following expression for the Hopf bifurcation critical delay value for the real-time delay model*

$$\Delta_{cr} = \frac{N\mu \arccos\left(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\right)}{\lambda\theta p\sqrt{1 - \left(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\right)^2}}. \quad (4.47)$$

Moreover, the model is always stable when

$$p \;\leq\; \frac{1}{2} + \frac{N\mu}{2\lambda\theta}. \quad (4.48)$$

Proof. The proof follows from the same reasoning in Novitzky et al. [25], however, we must proof the later statement in Equation 4.48.

$$\Delta = \frac{\arccos\Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)}{b}, \quad b = \frac{\lambda\theta p}{N\mu} \cdot \sqrt{1 - \Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)^2}. \tag{4.49}$$

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 (4.47) must be real and positive, which imposes the condition i.e.

$$\frac{\lambda\theta p}{N\mu} \cdot \sqrt{1 - \Big(-\frac{N\mu}{\lambda\theta p} - \frac{1}{p} + 1\Big)^2} > 0 \tag{4.50}$$

$$\Big(\frac{N\mu}{\lambda\theta p} + \frac{1}{p}\Big)\Big(2 - \frac{N\mu}{\lambda\theta p} - \frac{1}{p}\Big) > 0 \tag{4.51}$$

$$p > \frac{N\mu + \lambda\theta}{2\lambda\theta} = \frac{1}{2} + \frac{N\mu}{2\lambda\theta}. \tag{4.52}$$

This completes the proof. □

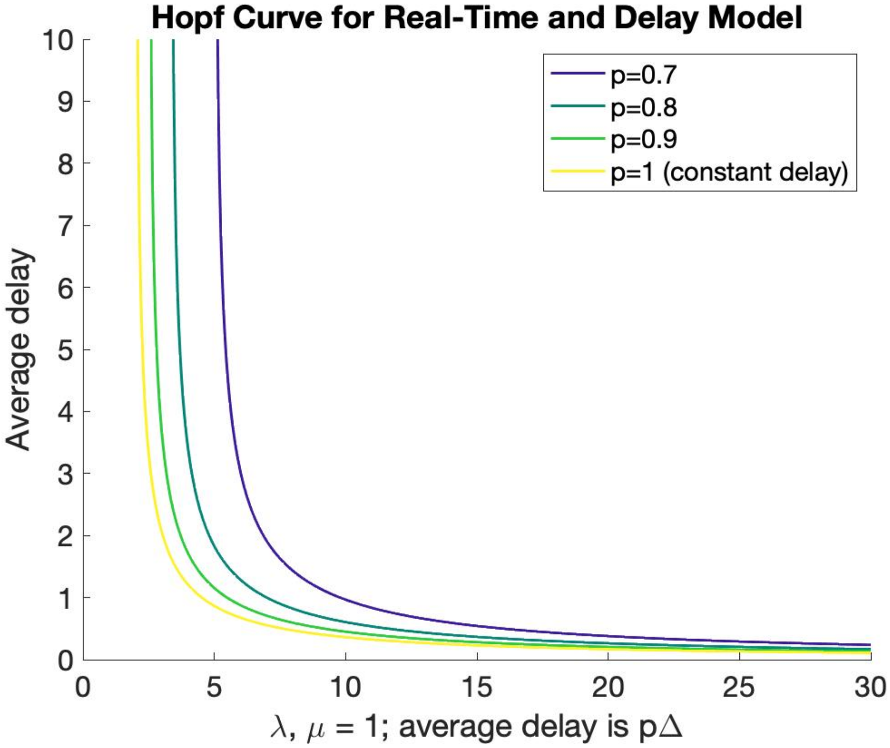

Figure 8: Hopf curve for different values of p.

Theorem 4.1 provides some important insights. The first insight implies that if at least half of the customers experience no delay, then the queueing system will remain stable despite all other factors. This condition is even stronger in the sense that strictly more than one half of all customers must be delayed in order to have a Hopf bifurcation. From a manager perspective, this is encouraging as giving some bit of real-time information dramatically improves the system and one can achieve local stability without everyone having real time information. The second insight is that the queueing system can become unstable if the condition (4.52) holds, meaning that a significant proportion of the population experienced the constant delay Δ. In Figure 8, we plot the Hopf bifurcation curves for different proportions of customers that experience the constant delay Δ. When the delay affects all customers ($p = 1$), the Hopf curve becomes exactly the constant delay Hopf curve from Figure 5. This implies that the formula for the Hopf bifurcation converges to the constant delay curve when $p \to 1$. 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. One important thing to note that when $p = .5$, this is not equivalent to reducing the delay Δ to $\Delta/2$. The fact that the system is completely stable when $p = .5$ implies that the real-time information has a stronger affect on the stability than the delayed information does.

4.3 M Discrete Delays

To further generalize the delay distribution, we will consider a queueing system where M different constant 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 probability mass function (pmf) and is specified by the probabilities $p_1, p_2, ..., p_m$ and values $\Delta_1, \Delta_2, ..., \Delta_m$. Thus, we have that

$$\mathbb{P}\left(X = \Delta_k\right) = p_k, \tag{4.53}$$

and the discrete random variable has Laplace transform given by

$$F(r) = \sum_{k=1}^{m} p_k e^{-r\Delta_k}. \tag{4.54}$$

The eigenvalue equation then follows directly from Equation (3.27),

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N} \sum_{k=1}^{m} p_k e^{-r\Delta_k} = 0. \tag{4.55}$$

Although this representation of the eigenvalue equation given in Equation 4.55 does not give us any information on how to solve it at first. However, it does show us that delay differential equations with multiple delays are equivalent to distributed delay differential equations where the delay is random variable given by a discrete probability distribution. Thus, if we can develop a framework for analyzing distributed delay equations, then it will immediately provide us with a methodology for understanding ddes with multiple constant delays. To the authors' knowledge this observation appears to be a new connection between distributed ddes and multi-delay ddes that has not been observed in the previous literature.

Thus, getting approximate solutions to the Hopf curves for discretely distributed delays will also yield approximate solutions to Hopf curves for multi delay systems.

Now that the delay distribution is more complicated, we cannot easily extract useful closed-form expressions to determine where the system of queues becomes unstable as we did before. In the next section, we will describe how to develop approximate analytical formulas for the Hopf curves. Although we cannot develop exact Hopf curves for these models, it is possible to solve (4.55) 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 (9), 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 have 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, we set $\mu = 0.5$ and on the right plot we set $\mu = 1$.

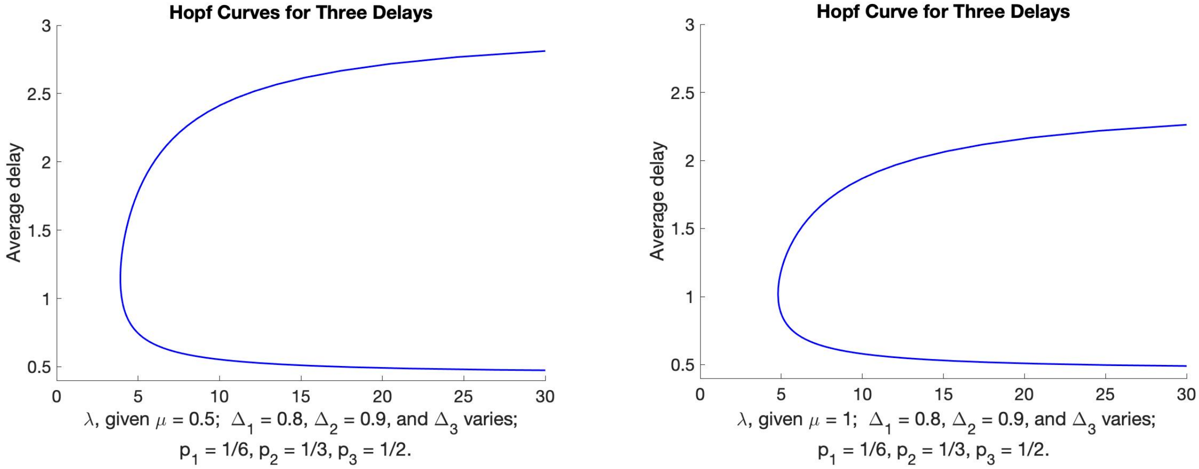

Figure 9: Hopf curve for three discrete delays.

An interesting 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. This phenomenon has been observed in delay differential equations with moving averages before Novitzky et al. [25]. The intuition for the moving average case is that for small delays, it is clear that the moving average model is stable. It then becomes unstable as one increases the delay, however, the model becomes stable again. One explanation for the stability for large Δ is that the moving average becomes closer and closer to a constant as the delay becomes larger. This is because for large values of Δ the moving average does not change much since it is averaging over a large interval. However, in the case of multiple delays, it is not clear why the same behavior occurs.

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 equal probability i.e.

$$X = \frac{2\Delta k}{M} \quad \text{w.p. } \frac{1}{M+1} \quad \text{where } k \in \{0, 1, 2, ..., M\}. \tag{4.56}$$

Lemma 4.2. *The Laplace transform for the random variable X above is equal to*

$$F_X^M(r) = \mathbb{E}[e^{-rX}] = \frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})}. \tag{4.57}$$

Proof.

$$\begin{aligned} F_X^M(r) &= \mathbb{E}[e^{-rX}] && (4.58)\\ &= \frac{1}{M+1}\sum_{k=0}^{M} e^{\frac{-2r\Delta k}{M}} && (4.59)\\ &= \frac{1}{M+1}\sum_{k=0}^{M} \left(e^{\frac{-2r\Delta}{M}}\right)^k && (4.60)\\ &= \frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})} \quad \text{(by the truncated geometric sum).} && (4.61) \end{aligned}$$

□

By Lemma 4.2, the characteristic equation becomes

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N} \cdot \frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})} = 0. \tag{4.62}$$

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 4.3.

Lemma 4.3. *As the number of points M tends to infinity, the Laplace transform for the random variable X has the following expression*

$$\lim_{M\to\infty} F_X^M(r) = \frac{1 - e^{-2r\Delta}}{2r\Delta} \tag{4.63}$$

Proof.

$$\begin{aligned} \lim_{M\to\infty} F_X^M(r) &= \lim_{M\to\infty} \frac{1 - e^{-2r\Delta\frac{M+1}{M}}}{(M+1)\cdot(1 - e^{-\frac{2r\Delta}{M}})} && (4.64)\\ &= \frac{1 - e^{-2r\Delta}}{2r\Delta}. && (4.65) \end{aligned}$$

□

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 4.5. Figure 10 shows the convergence as $M \to \infty$ of the Hopf curves from Equation 4.62 to the uniform distribution Hopf curve. One should also note that we do not display the two delay case. In this case, the queueing system is completely stable sinde it is identical to the real-time delay model where the fraction of customers with real-time information is one half of the population. Thus, the most important situation is studying when the number of delays is greater than two but not large.

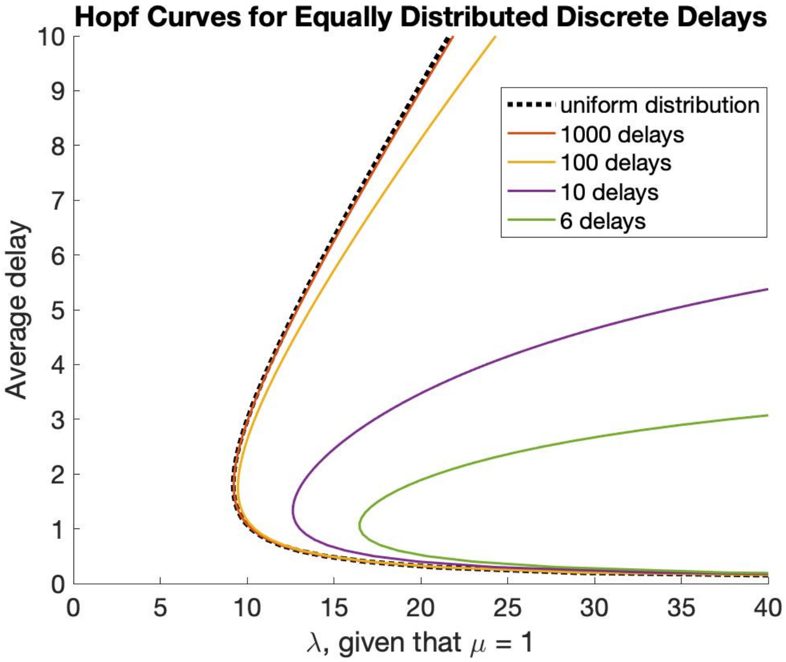

Figure 10: Hopf curves for a discrete delay system converge to the Hopf curve of a uniformly distributed delay. Uniform distribution yields the least stable system.

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 11 shows a queueing system where customers' commute to the queues may be approximated by a uniform distribution. This model is the *moving average* model.

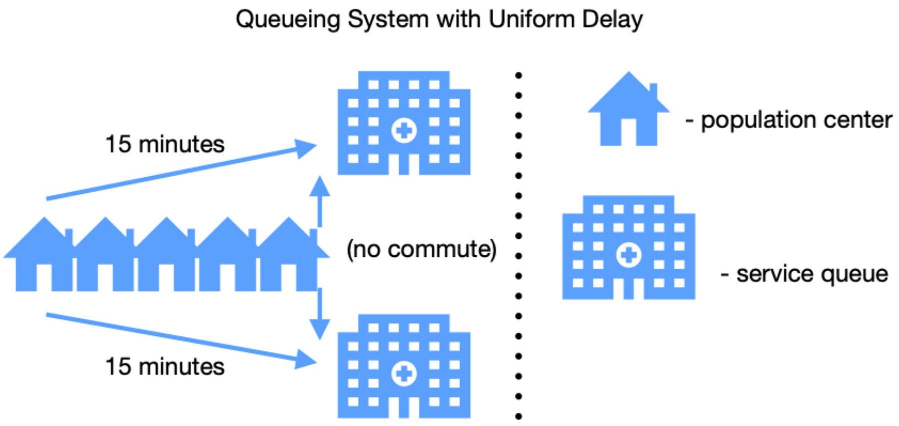

Figure 11: Customers going through a N-queue service system.

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

$$F(r) = \int_0^{2\Delta} e^{-rt} \cdot \frac{1}{2\Delta} dt = \frac{1}{2\Delta r}\left(1 - e^{-2r\Delta}\right). \tag{4.66}$$

The characteristic equation (3.27) for the queueing system is therefore

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N \cdot 2\Delta r} - \frac{\lambda\theta}{N \cdot 2\Delta r} \cdot e^{-2r\Delta} = 0. \tag{4.67}$$

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

$$\sin(2b\Delta) = -\frac{4\Delta\mu b}{\lambda\theta}, \quad \cos(2b\Delta) = 1 - \frac{4\Delta b^2}{\lambda\theta}. \tag{4.68}$$

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

$$b = \sqrt{\frac{\lambda\theta}{2\Delta} - \mu^2}. \tag{4.69}$$

When b is substituted into Equation (4.68), we get a transcendental equation for Δ,

$$\sin\left(2\Delta \cdot \sqrt{\frac{\lambda\theta}{2\Delta} - \mu^2}\right) + \frac{4\mu\Delta}{\lambda\theta} \cdot \sqrt{\frac{\lambda\theta}{2\Delta} - \mu^2} = 0. \tag{4.70}$$

When Equation (4.70) is solved numerically, the Hopf curve from Figure 12 can be found. As with the three discrete delays from Figure 9, the Hopf curve is not necessarily uniquely determined for a fixed customer arrival rate λ. As seen from plot on the right in Figure 12, when $\lambda = 15$, the queueing system is stable when the average delay Δ is below 0.5, unstable roughly when the average delay is in the range $[1, 6]$, and the stable again when $\Delta > 7$.

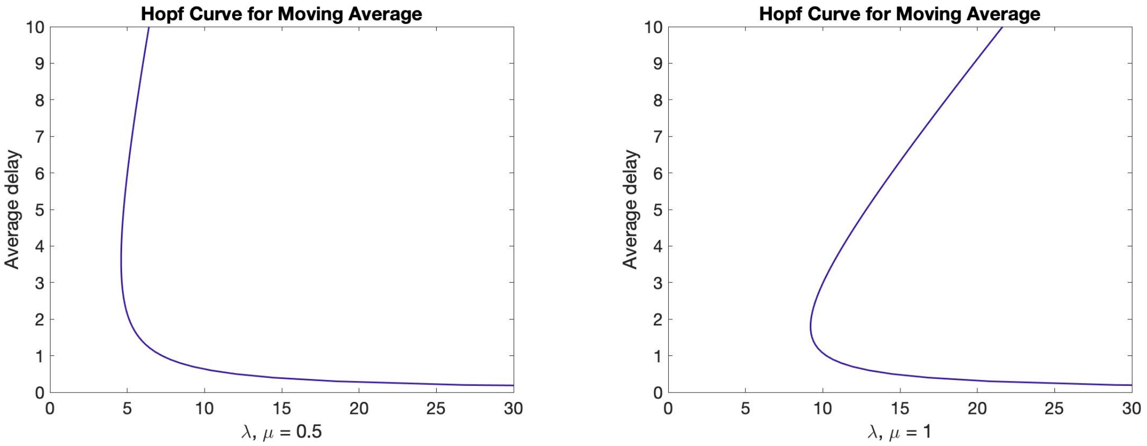

Figure 12: Hopf curve for uniformly distributed delay on $[0, 2\Delta]$. The average delay is Δ.

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

$$\begin{aligned} \Phi(r) &= r + \mu + \frac{\lambda\theta}{2arN} \cdot e^{-r\Delta}(e^{ra} - e^{-ra}) && (4.71) \\ &= r + \mu + \frac{\lambda\theta}{arN} \cdot e^{-r\Delta} \sinh(ra) && (4.72) \\ &= 0. && (4.73) \end{aligned}$$

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.

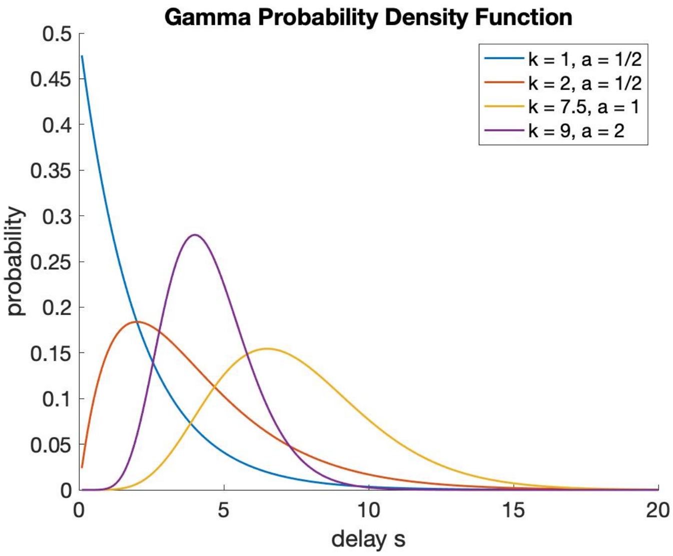

Figure 13: Customers going through a N-queue service system.

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

$$f(s) = \frac{a^k}{\Gamma(k)} s^{k-1} e^{-as}, \tag{4.74}$$

with the Laplace transform $F(r) = \frac{a^k}{(r+a)^k}$. The eigenvalue equation follows from (3.27),

$$\Phi(r) = r + \mu + \frac{\lambda\theta a^k}{N(r+a)^k} = 0. \tag{4.75}$$

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 $\tan(\phi) = \frac{b}{a}$:

$$(ib + \mu)\Big(\frac{ib}{a} + 1\Big)^k + \frac{\lambda\theta}{N} = (ib + \mu)\Big(i\tan(\phi) + 1\Big)^k + \frac{\lambda\theta}{N} = 0. \tag{4.76}$$

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

$$(ib + \mu)\Big(i\sin(\phi) + \cos(\phi)\Big)^k = -\frac{\lambda\theta}{N}\cos^k(\phi) \tag{4.77}$$

$$(ib + \mu)\Big(i\sin(k\phi) + \cos(k\phi)\Big) = -\frac{\lambda\theta}{N}\cos^k(\phi). \tag{4.78}$$

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

$$-b\sin(k\phi) + \mu\cos(k\phi) = -\frac{\lambda\theta}{N}\cos^k(\phi) \tag{4.79}$$

$$\mu\sin(k\phi) = -b\cos(k\phi), \tag{4.80}$$

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

$$\mu = -b\cot(k\phi) \tag{4.81}$$

$$\frac{\lambda\theta}{N} = (b\sin(k\phi) + b\cot(k\phi)\cos(k\phi))/\cos^k(\phi). \tag{4.82}$$

Based on these equalities, the Hopf curve can be found numerically. Figure 14, for example, is a result of setting $k = 2$ and varying the other parameter a. This in turn changes the average delay, which is given by $E[f(s)] = \frac{k}{a}$.

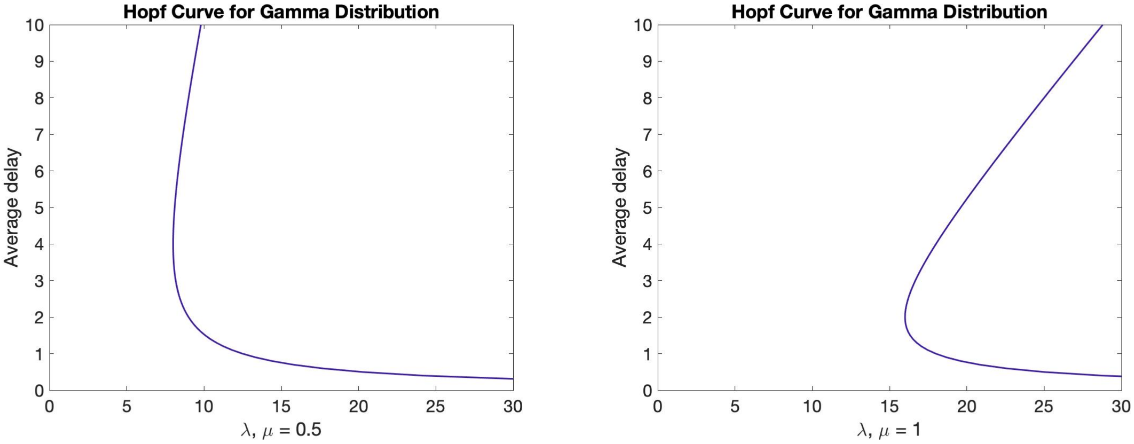

Figure 14: Hopf curve for gamma distribution, given that $k = 2$.

4.6.1 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 4.4. *When the delay distribution is given by an exponential distribution, the queueing system given in Equation* (3.4) *is always stable.*

Proof. The characteristic equation given in Equation (4.75) simplifies to a quadratic equation with respect to the eigenvalue r, namely,

$$r^2 + r(\mu + a) + \left(\mu a + \frac{\lambda\theta}{N} a\right) = 0, \tag{4.83}$$

$$r = \frac{1}{2}\left(-(\mu + a) \pm \sqrt{(\mu + a)^2 - 4a\left(\mu + \frac{\lambda\theta}{N}\right)}\right). \tag{4.84}$$

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

$$(\mu + a)^2 - 4a\left(\mu + \frac{\lambda\theta}{N}\right) = (\mu - a)^2 - \frac{4a\lambda\theta}{N} < (\mu - a)^2, \tag{4.85}$$

which reveals that the eigenvalue must be real and negative:

$$r = \frac{1}{2}\left(-(\mu + a) \pm \sqrt{(\mu + a)^2 - 4a\left(\mu + \frac{\lambda\theta}{N}\right)}\right) \tag{4.86}$$

$$< \frac{1}{2}\Big(- (\mu + a) + |\mu - a|\Big) < \max[-\mu, -a] < 0. \tag{4.87}$$

□

Hence when the delay is exponentially distributed, the queues are always stable. Figure 15 shows how the queueing model becomes more stable as k goes to 1.

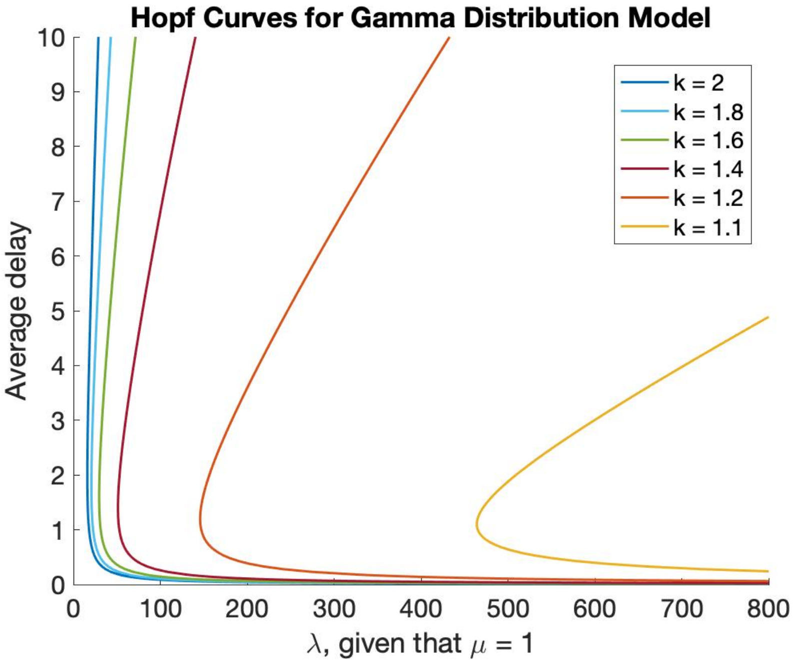

Figure 15: As $k \to 1$, gamma distribution model becomes more stable.

4.6.2 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 (4.75) with k then becomes

$$\lim_{k\to\infty} \frac{a^k}{(r+a)^k} = \lim_{k\to\infty} \left(\frac{k}{r\Delta + k}\right)^k = \lim_{k\to\infty} \exp\left(k \ln\left(\frac{k}{r\Delta + k}\right)\right). \tag{4.88}$$

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

$$\lim_{k\to\infty} k \ln\left(\frac{k}{r\Delta + k}\right) = \lim_{k\to\infty} \frac{\ln\left(\frac{k}{r\Delta+k}\right)}{\frac{1}{k}} = \lim_{k\to\infty} \frac{\frac{r\Delta+k}{k} \cdot \frac{r\Delta}{(r\Delta+k)^2}}{-k^{-2}} \tag{4.89}$$

$$= \lim_{k\to\infty} -\frac{r\Delta k}{r\Delta + k} = -r\Delta. \tag{4.90}$$

Combining (4.90) with (4.88), we find that

$$\lim_{k\to\infty} \frac{a^k}{(r+a)^k} = e^{-r\Delta}, \tag{4.91}$$

which confirms that in the limit as $k \to \infty$ the characteristic equation of the gamma distribution model given in Equation (4.75) converges to the characteristic equation of the constant delay model (4.42):

$$r + \mu + \frac{\lambda \theta a^k}{N(r+a)^k} \to r + \mu + \frac{\lambda \theta}{N} e^{-r\Delta}. \tag{4.92}$$

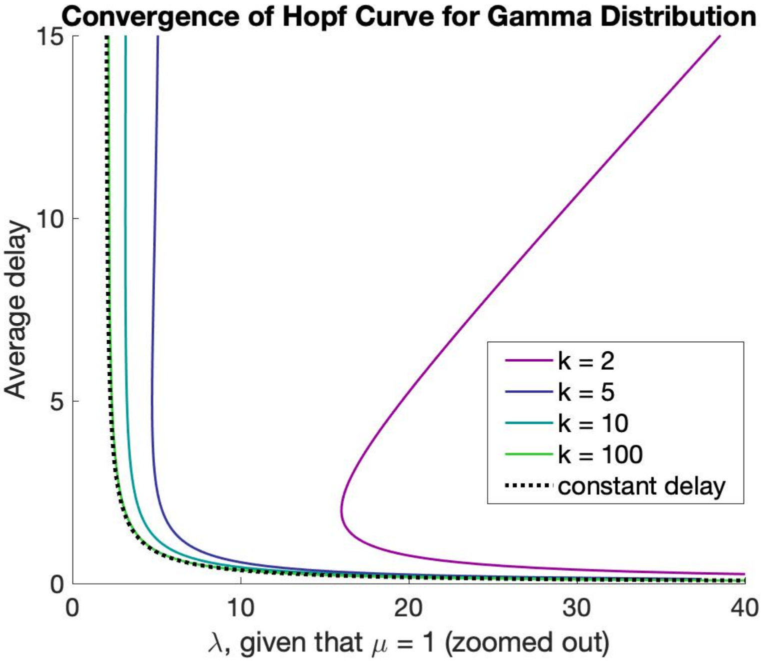

Figure 16: As $k \to \infty$, gamma distribution model converges to constant delay model.

4.7 Hyperexponential Distribution

Another continuous distribution of interest is the hyperexponential distribution. The hyperexponential distribution is a special case of a phase type distribution. Unlike the Erlang distribution it allows for more variability and has more variance than the exponential distribution. Usually the hyperexponential distribution is determined by two M-dimensional vectors of parameters $(p_1, p_2, ..., p_M)$ and $(a_1, a_2, ..., a_M)$. The vector $(p_1, p_2, ..., p_M)$ represents the probabilistic weights of each exponential distribution and the vector $(a_1, a_2, ..., a_M)$ is the associated rate parameters of each exponential. Thus, the hyperexponential is a convex combination of exponential distribution with potentially different rate parameters. The density is specified by $f(s)$ for $s \geq 0$ as

$$f(s) = \sum_{j=1}^{M} p_j a_j e^{-a_j s}, \tag{4.93}$$

with the Laplace transform

$$F(r) = \sum_{j=1}^{M} \frac{p_j a_j}{r + a_j}. \tag{4.94}$$

The eigenvalue equation follows from (3.27),

$$\Phi(r) = r + \mu + \sum_{j=1}^{M} \frac{\lambda\theta p_j a_j}{N(r + a_j)} = 0. \tag{4.95}$$

Here we specialize to the case of $M = 2$ since the case of $M = 1$ is an exponential distribution that we already analyzed. When $M = 2$, we have that

$$\begin{aligned}
\Phi(r) &= r + \mu + \frac{\lambda\theta p a_1}{N(r + a_1)} + \frac{\lambda\theta(1-p)a_2}{N(r + a_2)} && (4.96)\\
&= N(r + \mu)(r + a_1)(r + a_2) + \lambda\theta p a_1(r + a_2) + \lambda\theta(1-p)a_2(r + a_1) && (4.97)\\
&= N(r + \mu)(r + a_1)(r + a_2) + \lambda\theta\left(p a_1 + (1-p)a_2\right) r + \lambda\theta a_1 a_2 && (4.98)\\
&= N(r^3 + (a_1 + a_2 + \mu)r^2 + (a_1 a_2 + a_1\mu + a_2\mu)r + a_1 a_2 \mu) && (4.99)\\
&\quad +\lambda\theta\left(p a_1 + (1-p)a_2\right) r + \lambda\theta a_1 a_2
\end{aligned}$$

Theorem 4.5. *Suppose that $M = 2$ and we assume the probability distribution is given by the hyperexponential distribution, then for any set of model parameters the queueing model given in Equation 3.4 is always stable.*

Proof. Before we begin the proof, we want to note that it is impossible for the eigenvalue equation to have a positive root since all of the coefficients are positive. Moreover, since the equation is cubic, must have at least one negative root. Thus, it remains for us to show that if the polynomial has complex roots, then the roots have negative real parts. We will exploit the fact that the polynomial coefficients are positive and use the Routh-Hurwitz criterion for the remainder of the proof. Since the polynomial is a cubic polynomial and the coefficients are positive, it remains to show that

$$(a_1 + a_2 + \mu) \cdot \left(a_1 a_2 + \mu(a_1 + a_2) + \frac{\lambda\theta p a_1}{N} + \frac{\lambda\theta(1-p)a_2}{N}\right) > \left(\frac{\lambda\theta}{N} + \mu\right) a_1 a_2. \tag{4.100}$$

We will verify this condition below. We have that

$$\begin{aligned}
&(a_1 + a_2 + \mu) \cdot \left(a_1 a_2 + \mu(a_1 + a_2) + \frac{\lambda\theta p a_1}{N} + \frac{\lambda\theta(1-p)a_2}{N}\right) && (4.101)\\
&\geq (a_1 + a_2 + \mu) \cdot \left(a_1 a_2 + \frac{\lambda\theta p a_1}{N} + \frac{\lambda\theta(1-p)a_2}{N}\right) && (4.102)\\
&\geq (a_1 + a_2 + \mu) \cdot \left(a_1 a_2 + \frac{\lambda\theta p a_1}{N} + \frac{\lambda\theta(1-p)a_2}{N}\right) && (4.103)\\
&\geq (a_2 + \mu) \cdot \left(a_1 a_2 + \frac{\lambda\theta p a_1}{N} + \frac{\lambda\theta(1-p)a_2}{N}\right) && (4.104)\\
&\geq (a_2 + \mu) \cdot \left(a_1 a_2 + \frac{\lambda\theta a_1}{N}\right) && (4.105)\\
&= \mu a_1 a_2 + a_1 a_2^2 + \frac{\lambda\mu\theta a_1}{N} + \frac{\lambda\theta}{N} a_1 a_2 && (4.106)\\
&> \left(\frac{\lambda\theta}{N} + \mu\right) a_1 a_2. && (4.107)
\end{aligned}$$

This completes the proof. □

We have shown that the hyperexponential distribution for $M = 2$ inherits the stability property of the exponential distribution. We suspect that this is also true for $M > 2$, however, we do not have a proof for this result as of now. The Routh-Hurwitz approach should work for the case where $M = 3$, however, for higher dimensions it is not clear that it will work since it involves analyzing the roots of higher order polynomials.

5 Data-Driven Central Moment Approximations

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. [2]. So if $\int_0^\infty sf(s)ds = \Delta$ and $\lambda\theta > N\mu$, the queues are asymptotically stable when

$$\Delta < \frac{\arccos(-N\mu/(\lambda\theta))}{\sqrt{\lambda^2\theta^2/N^2 - \mu^2}}. \tag{5.108}$$

  The condition for Δ is derived in Equation (4.44), 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 4.6.1.

- Even if the distribution is known to be bounded, an upper bound on stability is still not guaranteed. One can consider the distribution from Section 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 (3.27). This allows us to avoid dealing with the probability density function $f(s)$, and instead use its moments. Lemma 5.1 below establishes the connection between the Laplace transform and the moments of a given probability distribution.

Lemma 5.1. *The Laplace transform of the non-negative random variable X can be expanded in the following Taylor series around a generic point a*

$$E[e^{-rX}] \;=\; F(r) = e^{-ra}\left(\sum_{j=0}^{\infty}\frac{(-r)^j\cdot E[(X-a)^j]}{j!}\right). \tag{5.109}$$

Moreover, when $a=0$ we have

$$E[e^{-rX}] \;=\; F(r) = \sum_{n=0}^{\infty}\frac{(-r)^n}{n!}\cdot E[X^n]. \tag{5.110}$$

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

$$E[e^{-rX}] \;=\; F(r) = e^{-rE[X]}\left(1+\sum_{j=2}^{\infty}\frac{(-r)^j\cdot E[(X-E[X])^j]}{j!}\right). \tag{5.111}$$

Proof. This follows immediately from standard Taylor expansions. □

The random variable X from Lemma 5.1 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 $\Delta = E[X] = \int_0^\infty sf(s)ds$, and the j-th centered moment can be expressed as

$$E[(X-\Delta)^j] = \int_0^\infty (s-\Delta)^j f(s)ds, \quad \text{where} \quad \Delta = \int_0^\infty sf(s)ds. \tag{5.112}$$

It follows that the characteristic equation (3.27) can be express as

$$\Phi(r) = r + \mu + \frac{\lambda\theta}{N}\cdot e^{-r\Delta}\left(1+\sum_{j=2}^{\infty}\frac{(-r)^j\cdot E[(X-\Delta)^j]}{j!}\right) = 0. \tag{5.113}$$

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

$$\Phi(r) \approx r + \mu + \frac{\lambda\theta}{N}\cdot e^{-r\Delta}\left(\sum_{j=0}^{K}\frac{(-r)^j}{j!}\int_0^\infty (s-\Delta)^j f(s)ds\right), \tag{5.114}$$

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

Moreover, when r is not complex, we can use Jensen's inequality to show that

$$F(r) \;=\; E[e^{-rX}] \geq e^{-rE[X]}. \tag{5.115}$$

This simple application of Jensen's inequality provides much insight on why the constant delay is the most unstable distribution for a fixed mean value of the delay. This result also suggests that randomness helps the system be more stable. However, we still do not understand what type of randomness leads to more stable distributions since in our examples. It is clear from our previous examples that variance is not the lone quantity

We find it extremely important to note to readers that the central moments are critical as opposed to the standard moments of the delay distribution. Using the central moments allows us to use the constant delay model as a base model and expand around that. Expanding around a base distribution is common in the probability literature when using orthogonal polynomial expansions, see for example Dufresne [10], Pender [27], Engblom and Pender [13]. It is also important to note that using the moments directly, one does not obtain sine or cosine functions, which are used to compute the critical delay and frequency. Using a centering point that is not equal to the mean will yield sine and cosine functions, however, from our numerical experiments, it does not perform well unless it is near the mean. Thus, using the central moments is vital to the analysis and our bounds that we derive in the sequel.

When the distribution of the delay is symmetric, Equation (5.114) 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

$$F(r) = E[e^{-rX}] = e^{-r\Delta}\Big(1 + \sum_{j=1}^{\infty} \frac{(-r)^j}{j!} E[(X-\Delta)^j]\Big) \tag{5.116}$$

$$= e^{-r\Delta}\Big(1 + \sum_{j=1}^{\infty} \frac{r^{2j}}{2j!} E[(X-\Delta)^{2j}]\Big). \tag{5.117}$$

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

$$F(r) = \Big(\cos(b\Delta) - i\sin(b\Delta)\Big) \cdot \sum_{j=0}^{\infty} \Big((-1)^j \cdot \frac{b^{2j}}{2j!} \cdot E\big[(X-\Delta)^{2j}\big]\Big) \tag{5.118}$$

$$= \Big(\cos(b\Delta) - i\sin(b\Delta)\Big) \cdot \sum_{j=0}^{\infty} (-1)^j k_j, \quad \text{where} \quad k_j = \frac{b^{2j}}{2j!} \cdot E\big[(X-\Delta)^{2j}\big]. \tag{5.119}$$

The characteristic equation at the Hopf takes the form

$$\Phi(ib) = ib + \mu + \frac{\lambda\theta}{N} \cdot \Big(\cos(b\Delta) - i\sin(b\Delta)\Big) \cdot \sum_{j=0}^{\infty} (-1)^j k_j = 0. \tag{5.120}$$

Therefore, an approximation based on K central moments to where the Hopf bifurcation occurs is given by the solution of the system of equations

$$\begin{cases} \mu + \frac{\lambda\theta}{N} \cdot \cos(b\Delta) \cdot \sum_{j=0}^{K} (-1)^j k_j = 0 \\ b - \frac{\lambda\theta}{N} \cdot \sin(b\Delta) \cdot \sum_{j=0}^{K} (-1)^j k_j = 0. \end{cases} \tag{5.121}$$

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 5.2. *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 $r = ib$:*

$$-\sin(ba)\sum_{j=0}^{2N}(-1)^j k_j \leq \operatorname{Im}\big(F(r)\big) \leq -\sin(ba)\sum_{j=0}^{2N+1}(-1)^j k_j \tag{5.122}$$

$$\cos(ba)\sum_{j=0}^{2N}(-1)^j k_j \leq \operatorname{Re}\big(F(r)\big) \leq \cos(ba)\sum_{j=0}^{2N+1}(-1)^j k_j, \tag{5.123}$$

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

Proof. By Equation (5.119),

$$\lim_{n\to\infty}\Big(\cos(ba) - i\sin(ba)\Big) \cdot S_n = F(ib), \tag{5.124}$$

$$\text{where} \quad S_n = \sum_{j=0}^{n}(-1)^j k_j. \tag{5.125}$$

For any n, S_n is a real quantity, so by separating the real and imaginary parts of (5.124), we see that

$$\lim_{n\to\infty}\cos(ba) \cdot S_n = \operatorname{Re}[F(ib)] \tag{5.126}$$

$$\lim_{n\to\infty} -\sin(ba) \cdot S_n = \operatorname{Im}[F(ib)] \tag{5.127}$$

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

$$S_{2(n+1)} = \sum_{j=0}^{2(n+1)}(-1)^j k_j = \sum_{j=0}^{2n}(-1)^j k_j + \big(-k_{2n+1} + k_{2n+2}\big) \tag{5.128}$$

$$\leq \sum_{j=0}^{2n}(-1)^j k_j = S_{2n}. \tag{5.129}$$

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

$$S_{2(n+1)+1} = \sum_{j=0}^{2(n+1)+1}(-1)^j k_j = \sum_{j=0}^{2n+1}(-1)^j k_j + \big(k_{2n+2} - k_{2n+3}\big) \tag{5.130}$$

$$\geq \sum_{j=0}^{2n+1}(-1)^j k_j = S_{2n+1}. \tag{5.131}$$

Lastly, we note that $b > 0$ must satisfy the characteristic equation (3.27), meaning that

$$\begin{cases} \mu + \mathrm{Re}\big[\frac{\lambda\theta}{N} F(ib)\big] & = 0 \\ b + \mathrm{Im}\big[\frac{\lambda\theta}{N} F(ib)\big] & = 0, \end{cases} \tag{5.132}$$

$$\begin{cases} \mu + \frac{\lambda\theta}{N} \cdot \cos(ba) \sum_{j=0}^{\infty} (-1)^j k_j & = 0 \\ b - \frac{\lambda\theta}{N} \sin(ba) \sum_{j=0}^{\infty} (-1)^j k_j & = 0. \end{cases} \tag{5.133}$$

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

$$\cos(ba) < 0 \quad \text{and} \quad \sin(ba) > 0. \tag{5.134}$$

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

$$\cos(ba) S_{2n} \geq \cos(ba) S_{2(n+1)} \geq \mathrm{Re}[F(ib)] \tag{5.135}$$

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

$$\cos(ba) S_{2n+1} \leq \cos(ba) S_{2(n+1)+1} \leq \mathrm{Re}[F(ib)]. \tag{5.136}$$

These inequalities state an upper and lower bound on the real part of the Laplace transform, as given in Equation (5.123). Similarly, because $\sin(ba) > 0$, we get the upper and lower bounds on the imaginary part of the Laplace transform as in Equation (5.122):

$$-\sin(ba) S_{2n} \leq -\sin(ba) S_{2(n+1)} \leq \mathrm{Im}[F(ib)] \tag{5.137}$$

$$-\sin(ba) S_{2n+1} \geq -\sin(ba) S_{2(n+1)+1} \geq \mathrm{Im}[F(ib)]. \tag{5.138}$$

Note that Equations (5.135)-(5.138) also demonstrate that choosing a larger n provides tighter bounds on both the real and the imaginary parts of $F(r)$. □

Since the exact point of the Hopf bifurcation is given by

$$\mu + \frac{\lambda\theta}{N} \mathrm{Re}(F(ib)) = 0 \tag{5.139}$$

$$b - \frac{\lambda\theta}{N} \mathrm{Im}(F(ib)) = 0, \tag{5.140}$$

or when the arrival rate of the customers satisfies the equation

$$\lambda = -\frac{\mu N}{\theta \, \mathrm{Re}(F(ib))} > 0, \tag{5.141}$$

then by Theorem 5.2 for every $n \geq 0$ we have the following upper and lower bounds on λ:

$$\lambda \geq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2(n+1)} (-1)^j k_j} \geq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2n} (-1)^j k_j}, \tag{5.142}$$

$$\lambda \leq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2n+3} (-1)^j k_j} \leq -\frac{\mu N}{\theta \cos(ba) \sum_{j=0}^{2n+1} (-1)^j k_j}. \tag{5.143}$$

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.

5.1 Examples Using the Approximation Method

We demonstrate numerically the performance of the approximation from Equations (5.121) based on several delay distributions.

Discrete uniform delays on a bounded interval Recall the distribution from Section 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 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

$$E[(X-\Delta)^{2n}] = \begin{cases} \frac{2}{M+1} \cdot \sum_{j=1}^{M/2} \left(\frac{2\Delta j}{M}\right)^{2n}, & M \text{ is even} \\ \frac{2}{M+1} \cdot \sum_{j=1}^{(M-1)/2} \left(\frac{2\Delta j}{M} - \frac{\Delta}{M}\right)^{2n}, & M \text{ is odd.} \end{cases} \tag{5.144}$$

In Figure 17, 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 (5.121). 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.

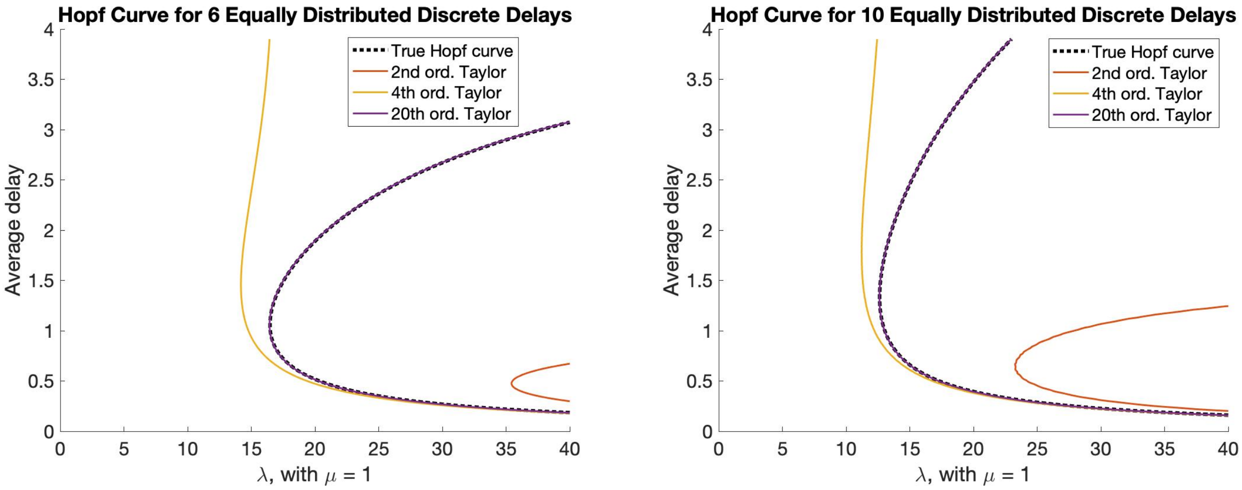

Figure 17: The approximation method applied to a system with 6 delays (on the left) and 10 delays (on the right).

Uniform distribution The uniform distribution on interval $[0, 2\Delta]$ has even central moments given by

$$E[(X-\Delta)^{2n}] = \frac{\Delta^{2n}}{2n+1}. \tag{5.145}$$

Figure 18 shows the Hopf curve approximations when the two, four, six, and twenty central moments are included in the system of equations (5.121). 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.

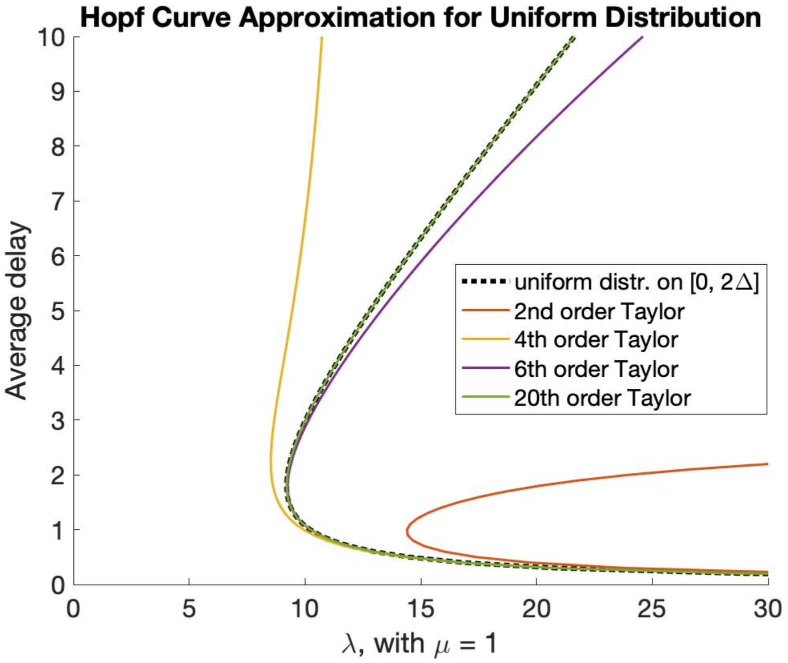

Figure 18: An approximation to a system with uniformly distributed delay on $[0, 2\Delta]$.

If the delay is uniformly distributed on the interval $[\Delta - a, \Delta + a]$ where $0 < a \leq \Delta$, the even central moments are

$$E[(X-\Delta)^{2n}] = \frac{a^{2n}}{2n+1}. \tag{5.146}$$

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 (5.146)). The left plot in Figure 19 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.

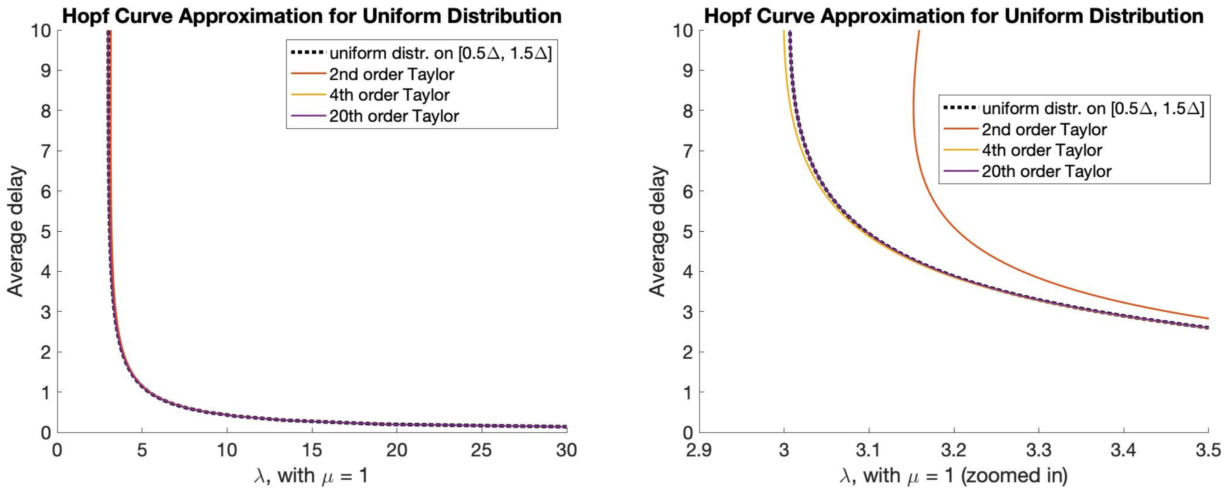

Figure 19: An approximation to a system with uniformly distributed delay on $[.5\Delta, 1.5\Delta]$.

6 Conclusion and Future Research

This paper formulates a fluid-like model for a system of N queues given in Equation (3.4), where the incoming customers choose which queue to join based on the current queue lengths. Our model accounts for the customers then travelling to the queue of their choosing thus causing a random delay in information, where the individual's travel time is represented by a random variable drawn from a fixed distribution.

Despite the randomness in the delay distribution, we show that the queueing system has a unique equilibrium state, which is guaranteed to be locally stable when the parameter relationship $\lambda\theta < N\mu$ is satisfied. The equilibrium can become unstable only if a Hopf bifurcation occurs, which we have shown depends significantly on the distribution of the customer delay. Under certain distributions, such as the exponential distribution, hyperexponential, or real-time and delay distribution, the queues may remain locally and asymptotically stable regardless of the size of the delay and any other model parameters. For other delay distributions, however, given that $\lambda\theta \geq N\mu$, a Hopf bifurcation may occur and the equilibrium may become unstable. Common delay distributions are considered in Section 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.

A natural extension of this paper is to conduct further study the of queueing models with specific distributions. For example, one can ask how does the delay distribution affect the amplitude and frequency of the oscillations in queues that result from a Hopf bifurcation. Moreover, when the queueing system is locally stable, what is the rate of convergence of the queue lengths to the equilibrium? Conversely, when the system is unstable, how quickly do the oscillations converge to the equilibrium amplitude?

It would also be great to develop guarantees on the accuracy of our central moment method for non-symmetric distributions. We noticed that the central moment 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 the Taylor series around the average delay produce by far the most accurate results, and we would like to learn why that is the case.

Finally, we suggest another interesting topic of future research is to explore delay differential equations with a countable number of constant delays. We have yet to find an application for the countably infinite setting, however, it is an interesting mathematical question. Using the probabilistic perspective, we can use countable discrete distributions like the Poisson, negative binomial, and the geometric as examples to explore the stability of a dde system with a countable number of delays. In particular the Poisson distribution is intriguing since it is described solely by its mean parameter and its cumulant moments are all equal to its mean.

References

[1] Jacques Bélair and Sue Ann Campbell. Stability and bifurcations of equilibria in a multiple-delayed differential equation. *SIAM Journal on Applied Mathematics*, 54(5): 1402–1424, 1994.

[2] 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.

[3] Elena Braverman and Sergey Zhukovskiy. Absolute and delay-dependent stability of equations with a distributed delay. *Discrete & Continuous Dynamical Systems-A*, 32 (6):2041, 2012.

[4] D Breda, Stefano Maset, and R Vermiglio. Computing the characteristic roots for delay differential equations. *IMA Journal of Numerical Analysis*, 24(1):1–19, 2004.

[5] Renato C Calleja, AR Humphries, and Bernd Krauskopf. Resonance phenomena in a scalar delay differential equation with two state-dependent delays. *SIAM Journal on Applied Dynamical Systems*, 16(3):1474–1513, 2017.

[6] SA Campbell and R Jessop. Approximating the stability region for a differential equation with a distributed delay. *Mathematical Modelling of Natural Phenomena*, 4(2): 1–27, 2009.

[7] Kenneth L Cooke and Zvi Grossman. Discrete delay, distributed delay and stability switches. *Journal of mathematical analysis and applications*, 86(2):592–627, 1982.

[8] Carlos Cuvas and Sabine Mondié. Necessary stability conditions for delay systems with multiple pointwise and distributed delays. *IEEE Transactions on Automatic Control*, 61(7):1987–1994, 2015.

[9] Jing Dong, Elad Yom-Tov, and Galit B. Yom-Tov. The impact of delay announcements on hospital network coordination and waiting times. *Management Science*, 2018.

[10] Daniel Dufresne. Laguerre series for asian and other options. *Mathematical Finance*, 10 (4):407–428, 2000.

[11] Stephen G Eick, William A Massey, and Ward Whitt. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252, 1993.

[12] Stephen G Eick, William A Massey, and Ward Whitt. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742, 1993.

[13] Stefan Engblom and Jamol Pender. Approximations for the moments of nonstationary and state dependent birth-death queues. *arXiv preprint arXiv:1406.6164*, 2014.

[14] 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.

[15] Jack Hale and Verduyn Lunel. *Introduction to Functional Differential Equations*. Springer Science, 1993.

[16] Donald L. Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, 2(2):429–441, 1965.

[17] Gábor Kiss and Bernd Krauskopf. Stability implications of delay distribution for first-order and second-order systems. *Discrete & Continuous Dynamical Systems-B*, 13(2): 327–345, 2010.

[18] 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.

[19] 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.

[20] David Lipshutz. Open problem—load balancing using delayed information. *Stochastic Systems*, 9(3):305–306, 2019.

[21] Adele H Marshall, Barry Shaw, and Sally I McClean. Estimating the costs for a group of geriatric patients using the coxian phase-type distribution. *Statistics in medicine*, 26 (13):2716–2729, 2007.

[22] 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`.

[23] Constantin-Irinel Morărescu, Silviu-Iulian Niculescu, and Keqin Gu. Stability crossing curves of shifted gamma-distributed delay systems. *SIAM Journal on Applied Dynamical Systems*, 6(2):475–493, 2007.

[24] 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.

[25] Sophia Novitzky, Jamol Pender, Richard H Rand, and Elizabeth Wesson. Nonlinear dynamics in queueing theory: Determining size of oscillations in queues with delay. *SIAM J. Appl. Dyn. Syst.*, 18:279–311, 2018.

[26] Sophia Novitzky, Jamol Pender, Richard Rand, and Elizabeth Wesson. Limiting the oscillations in queues with delayed information through a novel type of delay announcement. *arXiv preprint arXiv:1902.07617*, 2019.

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

[28] Jamol Pender, Richard Rand, and Elizabeth Wesson. A stochastic analysis of queues with customer choice and delayed information. *Mathematics of Operations Research to Appear*.

[29] Jamol Pender, Richard H. Rand, and Elizabeth Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 27(4), 2017.

[30] 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.

[31] B Rahman, Konstantin B Blyuss, and Yuliya N Kyrychko. Dynamics of neural systems with discrete and distributed time delays. *SIAM Journal on Applied Dynamical Systems*, 14(4):2069–2095, 2015.

[32] Gaurav Raina and Damon Wischik. Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. *Next Generation Internet Networks, 2005*, IEEE, 2005.

[33] 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.

[34] Shuang Tao and Jamol Pender. A stochastic analysis of bike sharing systems. *Probability in the Engineering and Informational Sciences to Appear*, 2020.

[35] Kenneth Train. *Discrete Choice Methods with Simulation.* Cambridge University Press, 2009.

[36] Yuan Yuan and Jacques Bélair. Stability and hopf bifurcation analysis for functional differential equation with distributed delay. *SIAM Journal on Applied Dynamical Systems*, 10(2):551–581, 2011.