Dynamic Control for Nonstationary Queueing Networks

Ziyuan Qin
Cornell University
School of Operations Research and Information Engineering
zq29@cornell.edu

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

January 25, 2017

Abstract

Nonstationary queueing networks are notoriously difficult to analyze and control. One reason is that steady state analysis and techniques are not useful since the model parameters in practice are not constant and depend on time. In this work, we analyze two optimal control problems for nonstationary Jackson networks with abandonment where our main goal is to optimally control the number of servers in the network. The first control problem approximates the stochastic dynamics of the Jackson network with a deterministic fluid model and optimizes with respect to the fluid model. In this case, we prove that the optimal solution is bang-bang and prove an asymptotic optimality result, which shows the fluid model optimal solution is asymptotically optimal for the scaled stochastic control problem. Our second approach exploits a Gaussian infinite server approximation for the queue length process and optimizes with respect to the number of servers. Unlike the fluid model's bang-bang solution, the infinite server approximation yields an square root staffing formula that depends on the cost to revenue ratio at each station. This proves the optimality of the square root staffing formula in a network setting.

Keywords Healthcare; service systems; optimal control; Jackson networks, Erlang-A networks, fluid model, infinite server queue, square root staffing, functional forward equations.

1 Introduction

Motivated by large scale service systems such as healthcare delivery organizations, call centers, and telecommunication networks, we formulate and analyze optimal control problems

for appropriately staffing these large scale systems. To model the stochastic dynamics of these service systems, it is necessary to construct an appropriate queueing model. In order to construct the most realistic, yet tractable queueing model, we need to understand the individual actions and collective dynamics of customers, patients, and the agents that serve them.

One common feature of these large scale service systems is time varying arrival rates. The customer arrival process of any queueing model must be able to handle time varying arrivals as customers do not arrive at constant rates as much of the current literature assumes. However, the analysis of queues with time varying rates is often quite complicated and requires non-trivial stochastic analysis. Moreover, customers that arrive to the system act independently of other customers, which allows us to model each customer as an independent entity. The second feature that a queueing model must have is parallel service. At any time there are multiple agents that are willing to assist customers with their needs, which means that customers have access to agents in a parallel fashion. In a hospital for example, these agents may be the hospital beds, the number of x-ray machines, the number of nurses, or the number of available doctors. In a call center for example, the parallel service could be the call center agents or interactive voice response lines Khudyakov et al. [13]. A third feature of these systems is customer abandonment since customers are impatient. In both of the previous examples, arriving customers wanting to engage in service may be delayed if all the available agents are busy with other customers or patients. The customers that are delayed waiting for service may decide to leave the system if they feel that their delay in receiving quality service has been excessive. In a healthcare context, this type of abandonment is called *Left Without Being Seen* (LWBS).

If these were the only features of service systems, then it would be tempting to model the dynamics with a $M_t/M/C_t + M$ queue, or the time varying Erlang-A queue. However, in many applications customers may go through many stages of service, which suggests that our queueing model should have multiple stations for service. For example, in hospitals, it is common for patients to interact with many different parts of the hospital, especially if the patient's condition is not life-threatening. A patient with a dislocated shoulder would not only interact with a doctor, but also would have to get an x-ray to determine the severity of the dislocation. This routing to multiple parties within healthcare is common because one doctor or nurse may not be sufficient to address all the needs of the patient. Inspired by these intrinsic features of these large scale systems, leads us to model these systems with nonstationary Jackson networks with abandonment or the $(M_t/M/C_t + M)^N$ queueing network where N is the number of stations.

To understand how to staff our Jackson network optimally, we need to derive approximations for the queueing process since it is intractable in its current form. The first method we use approximates the stochastic behavior of our queueing system via a fluid model. More recently, fluid models have been receiving more attention in the queueing theory literature, see for example Bäuerle et al. [2], Pang and Day [23], Cudina and Ramanan [4], Niyirora and Pender [22]. One reason is that there is a intimate connection between fluid models and stochastic stability of the queueing network. Furthermore, fluid models are derived rigorously from functional strong law of large number limit theorems and provide insight on the average sample path dynamics of the stochastic system and how to control it optimally see for example Bäuerle and Rieder [1] and Nazarathy and Weiss [21]. Using the fluid limit as a

simplified model for the dynamics of the queueing network, we then optimize with respect to the number of servers in the network. Thus, we generalize the seminal work of Hampshire et al. [10], Hampshire [9] by analyzing the *Competing Lagrangian* problem within a Jackson network. We derive a near-optimal staffing schedule by analyzing the fluid approximation and show that fluid optimal staffing reduces the queueing dynamics to maximizing over the set of 2^N *Competing Lagrangians* where N is the number of stations in the network. Moreover, we show that the profit generated from using this fluid optimal staffing procedure is asymptotically optimal for the scaled system as we let $\eta \to \infty$, which gives us justification for using the fluid limit to perform our approximate optimal staffing analysis. However, the fluid model is purely deterministic and does not incorporate any stochastic effects from the original stochastic queueing model.

Our second method of approximation is inspired by the infinite server queue. Since the infinite server queue, even in the nonstationary setting, has a Poisson distribution, then the mean and variance of the infinite server queue are identical. Thus, we propose to approximate the queue length by a Gaussian distribution with the same mean and variance. We show that this yields a good approximation for the queue length and also preserves the fact that the mean queue length only depends on differential equations for the mean queue length. More importantly, this approximation not only includes stochastic effects that the fluid model does not capture, but also the allows us to derive a closed form square root optimal staffing formula that has an elegant managerial interpretation in terms of the cost to revenue ratio at each station. Unlike the fluid model optimization, the Gaussian approximation does not require optimization and is known in closed form. Thus, this approach allows for new analysis and managerial insights into nonstationary optimal staffing of queueing networks, especially when the objective is to maximize profitability. Throughout the paper, we provide theoretical support of our analysis and conduct numerical experiments to complement our findings. Finally, we should mention that we have written a Matlab script to compute the optimal staffing policy for an arbitrary network and this code has been made avaialble on the first author's website.

1.1 Contributions

Thus, to the best of our knowledge, our contributions in this work are the following:

- We derive the optimal staffing function for time varying Jackson networks with abandonment approximated by a fluid model.
- We prove the asymptotic profit optimality of the fluid model approximation.
- We show that a square root staffing function is optimal when we approximate the Jackson network with by an infinite server queue and show that this square root staffing function only depends on the queue length and local parameters of each station.

1.2 Organization of the Paper

The rest of the paper continues as follows. In Section 2, we present our model for large scale service systems as a dynamic rate Jackson network with abandonment. We discuss

many of the different applications and the primitive data that is used for constructing our model. We also show that an appropriately scaled version of our Jackson network model fits into the Markovian service network framework, which allows us to derive a deterministic dynamical system or fluid model to approximate the mean behavior of our stochastic network model. We also show that the Jackson network framework encompasses many special cases of queueing networks such as tandem queues, queues with abandonments and retrials, and loss networks via fast abandonment. In Section 3, we describe the optimization problem that we intend to approximate and solve. In Section 4, we present a fluid model approximation and show that the optimal staffing policy is asymptotically optimal as we scale the system up. In Section 5, we present a Gaussian infinite server approximation for the queue length derive a square root staffing optimal solution. In Section 6, we present our numerical results and finally in the Appendix we provide all necessary proofs.

1.3 Notation

The paper will use the following notation:

- $\lambda_i(t)$ is the external arrival rate to node i at time t
- $\theta_i(t)$ is the abandonment rate for node i at time t
- $\mu_i(t)$ is the service rate for node i at time t
- $\gamma_{ij}(t)$ is the abandonment routing probability from node i to node j at time t
- $\tau_{ij}(t)$ is the service routing probability from node i to node j at time t
- $\gamma_i(t)$ is the abandonment departure probability from node i at time t
- $\tau_i(t)$ is the service departure probability from node i at time t
- $c_i(t)$ is the number of servers for node i at time t
- $x \wedge y = \min(\text{x , y})$
- $(x-y)^+ = \max(0, \text{x - y})$
- $x \otimes y =$ Kronecker product of two vectors x and y
- $\mathbf{v}_i =$ jump vectors as explained in Section 2 of [18]
- $\Delta(\mu) =$ diagonalization of vector μ
- $\{x < y\}$ denotes an *indicator function* that equals one if the statement is true i.e. if $x < y$, and zero if the statement is false.
- P^s denotes the matrix of service departure routing probabilities.
- P^a denotes the matrix of abandonment routing probabilities.

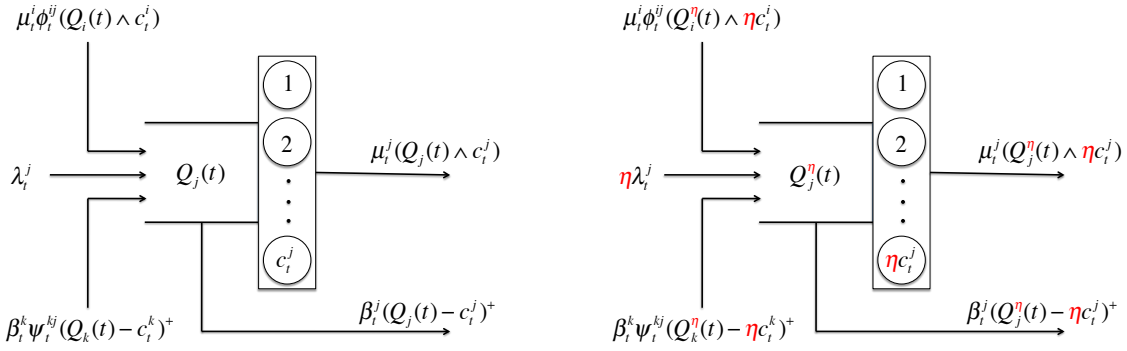

Figure 1: Jackson network with abandonment unscaled (Left) and scaled by η (Right).

- e_i = binary expansion vector for integer the $i-1$.

Moreover, we also require the following relations

$$\tau_i(t) + \sum_{j=1}^{N} \tau_{ij}(t) = 1 \quad \text{and} \quad \gamma_i(t) + \sum_{j=1}^{N} \gamma_{ij}(t) = 1. \tag{1.1}$$

These conditions ensure that the outflow from each node matches the number of depatures from the system and the inflow to other nodes. Lastly, we define functions related to Gaussian random variables i.e.

$$\varphi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) \equiv \int_{-\infty}^{x} \varphi(y)dy, \quad \overline{\Phi}(x) \equiv 1 - \Phi(x) = \int_{x}^{\infty} \varphi(y)dy. \tag{1.2}$$

2 Jackson Network Model

In this section, we formulate a general Jackson network model that will be used later to represent the dynamics of large scale service systems such as call centers or healthcare centers. We assume that there are $N \in \mathbb{N}$ stations or nodes of the Jackson network. We also assume that each node $j \in [1, 2..., N]$ has $c_j(t)$ statistically identical servers at time t. A simple schematic of our nonstationary Jackson network model is given in Figure 1.

2.1 Model Parameters

We now provide a detailed explanation of the parameters and flows of the Jackson network queueing model.

2.1.1 Arrivals

In our model we assume that the customer arrival process for each node follows an independent non-homogeneous Poisson process. We denote $\lambda_i(t)$ as the arrival rate function for a non-homogeneous Poisson process of node i. Probabilistically, the integral $\frac{1}{t-s}\int_s^t \lambda_i(u)du$ represents the average number of arriving customers to node i during the time interval (s,t]. A non-homogeneous Poisson process is a natural model for arrivals if you assume that the

customers arriving to each are statistically independent in disjoint time intervals and if they are arriving as single entities (single jumps). These are natural assumptions when modeling call centers and healthcare centers since customers act independently of other customers.

2.1.2 Service Completions

For the service completions, we assume that each node of the Jackson network has its own service completion rate of μ_i. This means that the $\frac{1}{\mu_i}$ is the average service time of a customer of receiving service from node i. We assume as in the formal theory of Markovian service networks, that these inter-service times are exponentially distributed.

2.1.3 Abandonments

For abandonments, we assume that each node of the Jackson network has its own abandonment rate of θ_i. Probabilistically, this means that the $\frac{1}{\theta_i}$ is the average time that a customer who has not received service is willing to wait in order to start to receive service at node i. The abandonment times are also exponentially distributed.

2.1.4 Routing

After a service completion or an abandonment from a node, each customer can leave the system entirely, be routed to another node of the network, or even stay at the same node for another service opportunity.

2.2 Jackson Network Model

Now that we understand the parameters of the Jackson network model, we now give a mathematical description of the queueing process in terms of Poisson random measures. The corresponding mathematical expression for our Jackson network $Q \equiv \{Q(t)|t \geq 0\}$ is defined by the following implicit equation

$$\begin{aligned} Q(t) \;=\; & Q(0) + \sum_{i=1}^{N}\sum_{j=1}^{N} \Pi^b_{ij}\left(\int_0^t (Q_i(s) - c_i(s))^+ \theta_i(s)\gamma_{ij}(s)ds\right)(\mathbf{v}_j - \mathbf{v}_i) \\ & + \sum_{i=1}^{N}\sum_{j=1}^{N} \Pi^c_{ij}\left(\int_0^t (Q_i(s) \wedge c_i(s))\mu_i(s)\tau_{ij}(s)ds\right)(\mathbf{v}_j - \mathbf{v}_i) + \sum_{i=1}^{N} \Pi^a_i\left(\int_0^t \lambda_i(s)ds\right)\mathbf{v}_i \\ & - \Pi^b_i\left(\int_0^t (Q_i(s) - c_i(s))^+ \theta_i(s)\gamma_i(s)ds\right)\mathbf{v}_i - \Pi^c_i\left(\int_0^t (Q_i(s) \wedge c_i(s))\mu_i(s)\tau_i(s)ds\right)\mathbf{v}_i \end{aligned} \tag{2.3}$$

where each of the Π_i's are independent unit rate Poisson processes.

2.3 Applications in Healthcare

Jackson networks are also used in healthcare for staffing nurses and beds as seen in De Véricourt and Jennings [5], Véricourt and Jennings [31] and Yom-Tov and Mandelbaum [32]. Figure 2 provides an illustration of a hospital system as a Jackson network. In this model, patients

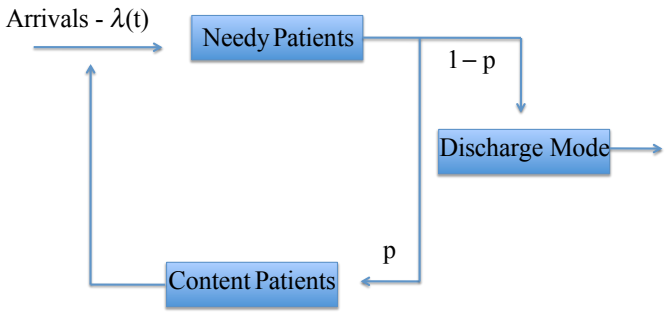

Figure 2: Erlang-R Model of De Véricourt and Jennings [5], Yom-Tov and Mandelbaum [32].

arrive to the hospital according to a non-homogenous Poisson process with arrival rate $\lambda(t)$. In the initial state patients are needy and are in need of a nurse. Once the nurse or doctor has made the patient stable or taken care of their needs, the patient transitions either to home (discharge) or they transition to a state where they do not need the assistance of a nurse. Yom-Tov and Mandelbaum [32] has coined this model the *Erlang R* model since patients are allowed to re-enter the needy state if necessary. The schematic of Figure 2 does not show any abandonments, however, this can be easily added to make the model more realistic.

2.4 Applications to Call Centers with Interactive Voice Response

Figure 3 shows how service systems such as call centers with IVR's can be modeled using Jackson networks. In this model, the arrival process is a non-homogenous Poisson process with arrival rate $\lambda_1(t)$. In this setting, the number of interactive voice processors is $c_1(t)$ and the number of human agents is $c_2(t)$. A customer is first served by an IVR processor at rate θ when one becomes available. Next, the customer may leave the system with probability $1 - \phi(t)$ if they were able to be serviced properly from the IVR or they may proceed to a human agent with probability $\phi(t)$. Although not explicitly shown in Figure 3, our model is general enough to consider abandonment at all states of the IVR. However, this model is not just restricted to call center applications.

From the variety of applications, it is obvious that Jackson networks are important modeling tools for large scale service systems. As a manager of these systems, it is vital to understand how to use these models or simplify these models to manage systems effectively and optimally. As the equation of our queueing system is expressed in terms of Poisson random measures, we cannot use the standard calculus of variations or stochastic control tools to optimally staff our queueing system. In order to take advantage of standard dynamic optimization tools, it is necessary to obtain approximate queueing models that capture a great deal of information about the original queueing system. In the next section via *uniform acceleration* we will show how the fluid limit of a scaled version of the Jackson network can be used as an approximate model to staff our service systems near optimality.

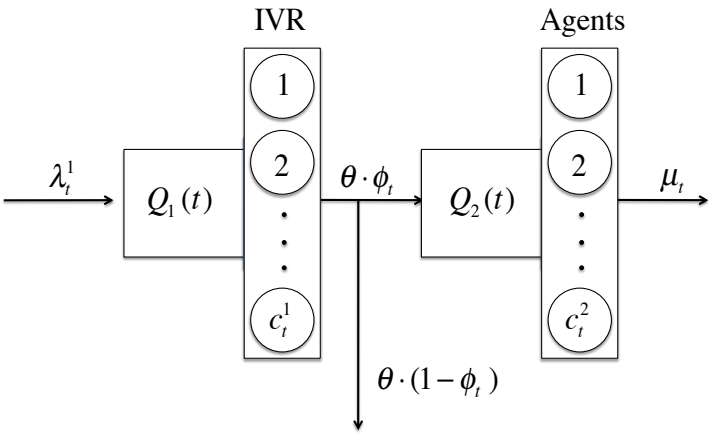

Figure 3: Schematic of IVR

2.5 Fluid Model

As our implicit equation for the queue length dynamics is not amenable for standard control methods, it is necessary to simplify the dynamics of our model in order to use standard optimization tools. With this in mind, we use the functional strong law of large numbers to construct a deterministic approximation for the sample path (and mean) behavior of our Jackson network. To this end, we first construct the uniformly accelerated version (see Mandelbaum et al. [18]) of our Jackson network model (3.1) with scale factor $\eta > 0$. This uniformly accelerated version of our model scales the arrival rate ($\lambda_i(t) \to \eta \cdot \lambda_i(t)$) and the number servers ($c_i(t) \to \eta \cdot c_i(t)$) by a scale factor η. In healthcare applications, this scaling is appropriate when there are a large number of patients and nurses in medium and large sized hospitals. Furthermore, for modeling call centers, this scaling is also natural as the number of customers making phone calls and the number of agents are very large. The right of Figure 1 highlights the new dynamics of our Jackson network under the uniform acceleration scaling. Figure 1 also illustrates that uniform acceleration does not change the service rate or the abandonment rate, which is appropriate for call centers and hospitals because managers do not have the ability to change how fast agents answer phone calls or the time that doctors spend with their patients. However, managers do have the ability to hire more agents, which will be used as our control parameter in the sequel.

The uniformly accelerated version of the queue length dynamics is given by the following implicit equation with scale factor $\eta > 0$

$$
\begin{aligned}
Q_i^\eta(t) &= Q_i^\eta(0) + \Pi_i^a\left(\int_0^t \eta\cdot\lambda_i(s)ds\right) - \sum_{j=1}^N \Pi_{ij}^b\left(\int_0^t (Q_i^\eta(s)-\eta\cdot c_i(s))^+\cdot\theta_i(s)\cdot\gamma_{ij}(s)ds\right)\\
&+ \sum_{j=1}^N \Pi_{ji}^b\left(\int_0^t (Q_j^\eta(s)-\eta\cdot c_j(s))^+\cdot\theta_j(s)\cdot\gamma_{ji}(s)ds\right)\\
&- \sum_{j=1}^N \Pi_{ij}^c\left(\int_0^t (Q_i^\eta(s)\wedge(\eta\cdot c_i(s)))\cdot\mu_i(s)\cdot\tau_{ij}(s)ds\right)\\
&+ \sum_{j=1}^N \Pi_{ji}^c\left(\int_0^t (Q_j^\eta(s)\wedge(\eta\cdot c_j(s)))\cdot\mu_j(s)\cdot\tau_{ji}(s)ds\right)\\
&- \Pi_i^b\left(\int_0^t (Q_i^\eta(s)-\eta\cdot c_i(s))^+\cdot\theta_i(s)\cdot\gamma_i(s)ds\right)\\
&- \Pi_i^c\left(\int_0^t (Q_i^\eta(s)\wedge(\eta\cdot c_i(s)))\cdot\mu_i(s)\cdot\tau_i(s)ds\right). \qquad (2.4)
\end{aligned}
$$

Although our equation for the queue length is accelerated with scale factor η the queue length dynamics are still expressed in terms of time changed unit rate Poisson processes. In order to simplify our dynamics, we must take the pointwise limit as $\eta\to\infty$ to construct our deterministic process $q(t)$, which satisfies simpler dynamics. This yields the following proposition.

Proposition 2.1. *Suppose that* $\lim_{\eta\to\infty} Q^\eta(0)/\eta = q(0)$, *then*

$$
\lim_{\eta\to\infty}\sup_{0\le t\le T}\left|\frac{1}{\eta}Q^\eta(t)-q(t)\right| = 0 \quad a.s\ u.o.c \qquad (2.5)
$$

where $q(t)\equiv\{q_1(t),q_2(t),....,q_n(t)|t\ge 0\}$ *is the solution to the following dynamical system*

$$
\begin{aligned}
\dot{q}_i &= \lambda_i - \mu_i\cdot(q_i\wedge c_i) - \theta_i\cdot(q_i-c_i)^+ \qquad (2.6)\\
&+ \sum_{j=1}^N (q_j-c_j)^+\cdot\theta_j\cdot\gamma_{ji} + \sum_{j=1}^N (q_j\wedge c_j)\cdot\mu_j\cdot\tau_{ji}, \quad 1\le i\le N.
\end{aligned}
$$

Proof. See Mandelbaum et al. [18]. □

The above proposition rigorously shows via strong approximations that for large η the dynamical system $q(t)\equiv\{q_1(t),....,q_n(t)|t\ge 0\}$ is a good approximation for the mean sample path behavior of the original queueing system. Throughout the rest of the paper we will refer to q as the fluid model for our queue length dynamics. Although our dynamical system approximation is true for all Jackson networks, there are many special cases of Jackson networks that of independent interest that may be used to model specific aspects of large scale service systems.

2.6 Functional Kolmogorov Forward Equations

The functional Kolmogorov forward equations for the mean queue length and covariance matrix for the Jackson network with abandonment satisfy the following differential equations in matrix form

Proposition 2.2. *Let $Q = \{Q_1, Q_2, ...Q_n\}$, then the forward equations for the mean vector of queue lengths and the covariance matrix are given by the following dynamical systems*

$$\begin{aligned}
\dot{E[Q]} &= \frac{d}{dt}E[Q] \\
&= \boldsymbol{\lambda} + \boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\boldsymbol{P}^s - \mathbb{I}) + \boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\boldsymbol{P}^a - \mathbb{I}) \\
&= \boldsymbol{\lambda} - \boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\mathbb{I} - \boldsymbol{P}^s) - \boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\mathbb{I} - \boldsymbol{P}^a) \\
\dot{\mathrm{Cov}}[Q, Q] &= \frac{d}{dt}\left(E[Q \otimes Q] - E[Q] \otimes E[Q]\right) \\
&= -\mathrm{Cov}[Q, Q \wedge \boldsymbol{c}] \cdot \mathbb{I} \otimes (\Delta(\boldsymbol{\mu}) \cdot (\mathbb{I} - \boldsymbol{P}^s)) - \mathrm{Cov}[Q \wedge \boldsymbol{c}, Q] \cdot (\Delta(\boldsymbol{\mu}) \cdot (\mathbb{I} - \boldsymbol{P}^s)) \otimes \mathbb{I} \\
&\quad -\mathrm{Cov}[Q, (Q - \boldsymbol{c})^+] \cdot \mathbb{I} \otimes (\Delta(\boldsymbol{\beta}) \cdot (\mathbb{I} - \boldsymbol{P}^a)) - \mathrm{Cov}[(Q - \boldsymbol{c})^+, Q] \cdot (\Delta(\boldsymbol{\beta}) \cdot (\mathbb{I} - \boldsymbol{P}^a)) \otimes \mathbb{I} \\
&\quad +\Delta\left(\boldsymbol{\lambda} + \boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\mathbb{I} - \boldsymbol{P}^s) + \boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\mathbb{I} - \boldsymbol{P}^a)\right) \\
&\quad -\Delta\left(\boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\boldsymbol{P}^s \oplus \boldsymbol{P}^s)\right) - \Delta\left(\boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\boldsymbol{P}^a \oplus \boldsymbol{P}^a)\right).
\end{aligned}$$

Proof. See Pender and Massey [28]. □

Corollary 2.3. *The mean queue length for station i of the Jackson network with abandonment can be written as the following differential equation*

$$\begin{aligned}
\dot{E}[Q_i] &= \lambda_i - \mu_i \cdot E\left[(Q_i \wedge c_i)\right] - \theta_i \cdot E\left[(Q_i - c_i)^+\right] \\
&\quad + \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot E\left[(Q_j \wedge c_j)\right] + \sum_{j=1}^{N} \theta_j \cdot \tau_{ji} \cdot E\left[(Q_j - c_j)^+\right].
\end{aligned} \tag{2.7}$$

Now that we have a good understanding of the model dynamics and various approximations, we now return to the main question of the paper. What is the proper level of staffing for our network given managerial and customer constraints?

3 Optimal Control of the Jackson Network

Now that we have constructed a queueing model to represent the dynamics of a service system network, we now turn our attention to understanding the issues that are important for the operator of the network to acheive profitability while maintaining a high quality service for customers. It is clear from the functional forward equations that $\{Q_j(t)|t \geq 0\}$ represents the total number of customers at station j (in the queue or in service) at time t. The term $Q_j(t) \wedge c_j(t)$ represents the number of customers in currently being served by c_j agents and service while term $(Q_j(t) - c_j(t))^+$ is used to represent the number of customers at station j that are waiting for service.

Thus, the fundamental question that we attempt to answer in this paper is how to find the optimal number of customer representatives or agents $c_j(t)$ to maximize profitability of

the service system network and to meet some given quality standard related to customer satisfaction. In clinical and healthcare settings, it is quite natural to minimize the waiting times of patients. By minimizing the waiting times of customers, we first preclude the abandonment phenomenon that is common in most queueing systems; in the context of healthcare this abandonment is called Left Without Being Seen (LWBS). In our model we choose to represent the customer or customer satisfaction constraint at station j by the following inequality:

$$\int_0^T \theta_j \cdot E[(Q_j(t) - c_j(t))^+]dt \leq \mathcal{E}_j \tag{3.8}$$

where

$$\mathcal{E}_j \equiv \epsilon_j \int_0^T \lambda_j(t)dt \tag{3.9}$$

and ϵ_j is the threshold of abandonment probability. This constraint enforces that during the time interval $[0, T]$, the number of customers that abandon, $\int_0^T \theta_j \cdot E[(Q_j(t) - c_j(t))^+]dt$, must be less or equal to the maximum allowable fraction of the total number patient arrivals $\mathcal{E}_j$. The objective function seeks maximize the profit of the queueing network by optimizing with respect to the number of servers $c(t) = \{c_1(t), c_2(t), ..., c_N(t)\}$. This may be formulated as the following dynamic optimization problem where $\mathcal{L}(c(t))$ is called the Lagrangian.

$$\mathcal{L}(c(t)) \quad = \quad \max_{\{c(t)\geq 0:\, 0\leq t\leq T\}} \int_0^T \left(\sum_{j=1}^N r_j \cdot \mu_j \cdot E[(Q_j(t) \wedge c_j(t)] - w_j \cdot c_j(t) \right) dt. \tag{3.10}$$

This objective function indicates that optimal number of servers $c(t) = \{c_1(t), c_2(t), ..., c_N(t)\}$ must be found to maximize the operating net profit of the network. The operating net profit of the network can be obtained by substracting the operating cost $w_j \cdot c_j(t)$ of each station j from the operating revenue $r_j \cdot \mu_j \cdot E[(Q_j(t) \wedge c_j(t)]$ obtained at each station j. Summing over all of the stations in the network yields the main optimal control problem that we intend to solve in this paper.

Problem 3.1 (Network Optimal Staffing Control Problem)**.**

$$\mathcal{L}(c(t)) \quad = \quad \max_{\{c(t)\geq 0:\, 0\leq t\leq T\}} \int_0^T \left(\sum_{j=1}^N r_j \cdot \mu_j \cdot E[(Q_j(t) \wedge c_j(t)] - w_j \cdot c_j(t) \right) dt \tag{3.11}$$

subject to

$$\dot{E}[Q_j(t)] = \lambda_j(t) - \mu_j \cdot E[Q_j(t) \wedge c_j(t)] - \theta_j \cdot E[(Q_j(t) - c_j(t))^+] \tag{3.12}$$
$$+ \sum_{i=1}^N \mu_i \cdot \gamma_{ij} \cdot E\left[(Q_i \wedge c_i)\right] + \sum_{i=1}^N \theta_i \cdot \tau_{ij} \cdot E\left[(Q_i - c_i)^+\right], \quad 1 \leq j \leq N.$$
$$\int_0^T \theta_j \cdot E[(Q_j(t) - c_j(t))^+]dt \leq \epsilon_j \int_0^T \lambda_j(t) = \mathcal{E}_j, \quad 1 \leq j \leq N. \tag{3.13}$$

Although the optimal control problem is not glaringly difficult, many issues arise when trying to solve this control problem. One of the main issues is obtaining the solution of the forward equations for the mean queue length and its max and min functionals. This is difficult since the forward equations are not a closed system and are not autonomous. This is because the min and max functions are not explicit functions of the mean queue length process. For more insight on this problem and a more rigorous explanation of this problem, see Pender [25], Engblom and Pender [7]. We should also mention that this problem is not isolated to just the state dynamics, but is also an issue for the objective function and the abandonment constraint since the queue length distributions are not known in closed form. Thus, in order to make the optimal control problem more tractable, we must develop approximations for the queue length processes in order to estimate and compute the expectation terms that appear in the optimal control problem. We describe two such methods of approximation in the sequel.

4 Optimal Control via Fluid Models

In order to approximate the solution to the optimal control problem presented in Problem 3.1, we will use the fluid approximation or Equation 2.6 for the queueing network dynamics to approximate the solution. Thus, we replace the expectations of functions of the queue length in Problem 3.1 with their deterministic counterparts from Equation 2.6 and we have that

$$\begin{aligned}\mathcal{L}(c,p,q) &= \int_0^T \left(\sum_{j=1}^N r_j \cdot \mu_j \cdot (q_j \wedge c_j) - w_j \cdot c_j\right) dt \qquad (4.14)\\ &+ \int_0^T \left(\sum_{j=1}^N p_j \cdot \left(\dot{q}_j - f_j(t,q,c)\right)\right) dt + \int_0^T \left(\sum_{j=1}^N x_j \cdot \left(\dot{z}_j - g_j(t,q,c)\right)\right) dt\end{aligned}$$

where

$$\begin{aligned}f_j(t,q,c) &= \lambda_j - \mu_j \cdot (q_j \wedge c_j) - \theta_j \cdot (q_j - c_j)^+ \qquad (4.15)\\ &+ \sum_{i=1}^N \theta_i \cdot \gamma_{ij} \cdot (q_i - c_i)^+ + \sum_{i=1}^N \mu_i \cdot \tau_{ij} \cdot (q_i \wedge c_i), \quad 1 \le j \le N.\\ g_j(t,q,c) &= -\theta_j \cdot (q_j - c_j)^+, \quad 1 \le j \le N. \qquad (4.16)\end{aligned}$$

Here p_j is the Lagrange multiplier of the state variable q_j and x_j is the co-state variable of some auxiliary variable z_j where

$$\begin{aligned}z_j &= -\int_0^t \theta_j \cdot (q_j - c_j)^+ \, dt \qquad (4.17)\\ \dot{z}_j &= -\theta_j \cdot (q_j - c_j)^+ \qquad (4.18)\end{aligned}$$

and

$$z_j(T) \geq -\mathcal{E}_j\,. \tag{4.19}$$

Since z_j does not appear in Equation 4.14, then $\dot{x}_j = -\partial\mathcal{H}/\partial z_j = 0$, meaning that x_j is a constant that satisfies the following complementary of slackness equation:

$$x_j \cdot \left[\mathcal{E}_j - \int_0^T \theta_j \cdot (q_j - c_j)^+ \, dt\right] = 0. \tag{4.20}$$

Thus, $x_j = 0$ when $\mathcal{E}_j - \int_0^T \theta_j \cdot (q_j - c_j)^+ \, dt > 0$, else $x_j > 0$. Moreover, the parameters of the Lagrangian have the following economic interpretation

- r_j is the (reward) per customer that leaves from queue j due to a completion of service.
- w_j is the (wage) or cost of each server used at the j^{th} queue.

We can also form the Hamiltonian of the optimal control problem as

$$\begin{aligned}\mathcal{H}(c,p,q) &= \int_0^T \left(\sum_{j=1}^N r_j \cdot \mu_j \cdot (q_j \wedge c_j) - w_j \cdot c_j\right) dt \\ &+ \int_0^T \left(\sum_{j=1}^N p_j \cdot f_j(t,q,c)\right) dt + \int_0^T \left(\sum_{j=1}^N x_j \cdot g_j(t,q,c)\right) dt.\end{aligned} \tag{4.21}$$

With the Hamiltonian we are now ready to perform our optimal control analysis. The first insight is that since the wages of the healthcare providers are linear, we can show that our profit maximization problem [4.14] reduces to the analysis of 2^N *Competing Lagrangians* $\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_{2^N}$ where $e_k \in \mathbb{R}^N$ is the binary expansion vector for the integer $i \in \{1, 2, ..., N\}$, and

$$\mathcal{L}_k \equiv \mathcal{L}(q, p, e_k \circ q).$$

To give the reader more intuition about this result, we provide the exact dynamics for the case where $N = 2$. In this case, we have that

$$e_1 = \begin{bmatrix}0\\0\end{bmatrix}, e_2 = \begin{bmatrix}1\\0\end{bmatrix}, e_3 = \begin{bmatrix}0\\1\end{bmatrix}, e_4 = \begin{bmatrix}1\\1\end{bmatrix}. \tag{4.22}$$

This yields the following optimal staffing vectors respectively for each Lagrangian

$$c^* = e_1 \circ q = \begin{bmatrix}0\\0\end{bmatrix}, c^* = e_2 \circ q = \begin{bmatrix}q_1\\0\end{bmatrix}, c^* = e_3 \circ q = \begin{bmatrix}0\\q_2\end{bmatrix}, c^* = e_4 \circ q = \begin{bmatrix}q_1\\q_2\end{bmatrix}. \tag{4.23}$$

In the case that the optimal solution is $c^* = \{0, 0\}$, then we have the following dynamics for the time derivatives of p and q

$$\dot{p}_1 = -(-p_1 \cdot \theta_1 + p_1 \cdot \theta_1 \cdot \gamma_{11} + p_2 \cdot \theta_1 \cdot \gamma_{12} - x_1 \cdot \theta_1) \tag{4.24}$$

$$\dot{p}_2 = -(p_1 \cdot \theta_2 \cdot \gamma_{21} - p_2 \cdot \theta_2 + p_2 \cdot \theta_2 \cdot \gamma_{22} - x_2 \cdot \theta_2) \tag{4.25}$$

$$\dot{q}_1 = \lambda_1 - \theta_1 \cdot q_1 + \theta_1 \cdot \gamma_{11} \cdot q_1 + \theta_2 \cdot \gamma_{21} \cdot q_2 \tag{4.26}$$

$$\dot{q}_2 = \lambda_2 - \theta_2 \cdot q_2 + \theta_1 \cdot \gamma_{12} \cdot q_1 + \theta_2 \cdot \gamma_{22} \cdot q_2, \tag{4.27}$$

and in the case that the optimal solution is $c^* = \{q_1, 0\}$, then we have the following dynamics for the time derivatives of p and q

$$\begin{aligned}
\dot{p}_1 &= -(r_1 \cdot \mu_1 - w_1 - p_1 \cdot \mu_1 + p_1 \cdot \mu_1 \cdot \tau_{11} + p_2 \cdot \mu_1 \cdot \tau_{12}) && (4.28)\\
\dot{p}_2 &= -(p_1 \cdot \theta_2 \cdot \gamma_{21} - p_2 \cdot \theta_2 + p_2 \cdot \theta_2 \cdot \gamma_{22} - x_2 \cdot \theta_2) && (4.29)\\
\dot{q}_1 &= \lambda_1 - \mu_1 \cdot q_1 + \mu_1 \cdot \tau_{11} \cdot q_1 + \theta_2 \cdot \gamma_{21} \cdot q_2 && (4.30)\\
\dot{q}_2 &= \lambda_2 - \theta_2 \cdot q_2 + \mu_1 \cdot \tau_{12} \cdot q_1 + \theta_2 \cdot \gamma_{22} \cdot q_2, && (4.31)
\end{aligned}$$

and in the case that the optimal solution is $c^* = \{0, q_2\}$, then we have the following dynamics for the time derivatives of p and q

$$\begin{aligned}
\dot{p}_1 &= -(-p_1 \cdot \theta_1 + p_1 \cdot \theta_1 \cdot \gamma_{11} + p_2 \cdot \theta_1 \cdot \gamma_{12} - x_1 \cdot \theta_1) && (4.32)\\
\dot{p}_2 &= -(r_2 \cdot \mu_2 - w_2 + p_1 \cdot \mu_2 \cdot \tau_{21} - p_2 \cdot \mu_2 + p_2 \cdot \mu_2 \cdot \tau_{22}) && (4.33)\\
\dot{q}_1 &= \lambda_1 - \theta_1 \cdot q_1 + \theta_1 \cdot \gamma_{11} \cdot q_1 + \mu_2 \cdot \tau_{21} \cdot q_2 && (4.34)\\
\dot{q}_2 &= \lambda_2 - \mu_2 \cdot q_2 + \theta_1 \cdot \gamma_{12} \cdot q_1 + \mu_2 \cdot \tau_{22} \cdot q_2, && (4.35)
\end{aligned}$$

and finally in the case that the optimal solution is $c^* = \{q_1, q_2\}$, then we have the following dynamics for the time derivatives of p and q

$$\begin{aligned}
\dot{p}_1 &= -(r_1 \cdot \mu_1 - w_1 - p_1 \cdot \mu_1 + p_1 \cdot \mu_1 \cdot \tau_{11} + p_2 \cdot \mu_1 \cdot \tau_{12}) && (4.36)\\
\dot{p}_2 &= -(r_2 \cdot \mu_2 - w_2 + p_1 \cdot \mu_2 \cdot \tau_{21} - p_2 \cdot \mu_2 + p_2 \cdot \mu_2 \cdot \tau_{22}) && (4.37)\\
\dot{q}_1 &= \lambda_1 - \mu_1 \cdot q_1 + \mu_1 \cdot \tau_{11} \cdot q_1 + \mu_2 \cdot \tau_{21} \cdot q_2 && (4.38)\\
\dot{q}_2 &= \lambda_2 - \mu_2 \cdot q_2 + \mu_1 \cdot \tau_{12} \cdot q_1 + \mu_2 \cdot \tau_{22} \cdot q_2, && (4.39)
\end{aligned}$$

Proposition 4.1. *Since the staffing costs are a linear function of the number of servers, then we have*

$$\max_{c \geq 0} \mathcal{L}(p, q, c) = \max_{\{k: 0 \leq k \leq 2^N\}} \mathcal{L}_k$$

where

$$\mathcal{L}_k \equiv \mathcal{L}(q, p, e_k \circ q)$$

and where e_k is the binary expansion vector for the integer $i \in \{1, 2, ..., N\}$. This also implies that

$$\max_{c \geq 0} \int_0^T \mathcal{L}(p, q, c) = \int_0^T \max_{c \geq 0} \mathcal{L}(p, q, c)$$

Proof. The proof follows from a similar argument given in Hampshire et al. [10], thus we omit it. □

At a given time t, we define the largest of these 2^N Lagrangians to be the one that is dominant. Using this dominant Lagrangian approach we are able to derive the following optimal staffing policy for our value function.

Theorem 4.2. *Given $\{p, q\}$, the optimal staffing procedure $c^*(t)$ is the following:*

$$c^*(t) = e_k \circ q \quad \text{if } \mathcal{L}_k \text{ is dominant}$$

Proof. The proof follows from the fact that the Lagrangian is piecewise linear. By invoking the maximum modulus property of piecewise linear functions and the boundedness of p and q, then we have that the maximum must occurs at one of the vertices given by the competing Lagrangians. Now one invokes the Pontryagin maximum principle and the proof is complete. □

Summarizing the Lagrangian analysis, we see that our fluid model has 2^N operational modes of behavior. In mode $k \in \{1, 2, 3...2^N\}$ we serve all customers that have non-zero entries in the corresponding binary expansion vector e_k.

Remark 4.3. *We should also mention that it is possible to also consider control sets such as $[c_{min}, c_{max}]$ where $0 \leq c_{min} \leq c_{max}$. These might be more suitable in a healthcare environment where there are government regulations on the minimum number of nurses at each hospital and a maximum is given the capacity of the hospital.*

Thus, with the Euler-Lagrange equations in all of the different optimal solution scenarios, we have a complete characterization of the optimal solution and the state dynamics of the Jackson network under our fluid model approximation.

4.1 Asymptotic Profit Rate Optimality

We have shown that using the fluid model as an approximation for the original stochastic network reduces our optimization to finding the maximal *Competing Lagrangian*. Although this gives us an optimal staffing schedule we now want to understand how close our approximate staffing schedule is to the optimal staffing schedule for the original problem. Here we show that the profit generated by the fluid limit control problem is asymptotically optimal for our original problem as we let the arrival rate and the number of servers tend to ∞. However, before we give the main result, we need to prove a few lemmas that are fundamental to the proof.

Lemma 4.4. *The maximum and minimum functions are Lipschitz continuous.*

Proof. Let f and g be any functions, then for the maximum function we have that

$$\begin{aligned} |f(x) \vee g(x) - f(y) \vee g(y)| &= |f(x) \vee g(x) - f(x) \vee g(y) + f(x) \vee g(y) - f(y) \vee g(y)| \\ &\leq |f(x) \vee g(x) - f(x) \vee g(y)| + |f(x) \vee g(y) - f(y) \vee g(y)| \\ &\leq (||f||_\infty + ||g||_\infty) \cdot |x - y|. \end{aligned}$$

Lastly, for the minimum function, we have that

$$\begin{aligned} |f(x) \wedge g(x) - f(y) \wedge g(y)| &= |f(x) \wedge g(x) - f(x) \wedge g(y) + f(x) \wedge g(y) - f(y) \wedge g(y)| \\ &\leq |f(x) \wedge g(x) - f(x) \wedge g(y)| + |f(x) \wedge g(y) - f(y) \wedge g(y)| \\ &\leq (||f||_\infty + ||g||_\infty) \cdot |x - y|. \end{aligned}$$

□

Lemma 4.5. *Let f be a Lipschitz continuous function and let*

$$\lim_{\eta\to\infty}\frac{1}{\eta}Q^\eta(t)\to q(t)\quad a.s. \tag{4.40}$$

Then, we have that

$$\lim_{\eta\to\infty} f\left(\sup_{0\le t\le T}\frac{1}{\eta}Q^\eta(t)\right)=f\left(\sup_{0\le t\le T}q(t)\right)\quad a.s. \tag{4.41}$$

and

$$\lim_{\eta\to\infty}\int_0^T f\left(\frac{1}{\eta}Q^\eta(t)\right)=\int_0^T f\left(q(t)\right)dt\quad a.s. \tag{4.42}$$

Proof. This follows from the continuous mapping theorem and the Lipschitz continuity of the function f. □

Mathematically, we define the profit rate $\mathcal{R}(x,C)$ to be:

$$\mathcal{R}(x,c) \;=\; \sum_{j=1}^{N} r_j\cdot\mu_j\cdot(x_j\wedge c_j)-\sum_{j=1}^{N}w_j\cdot c_j \tag{4.43}$$

Theorem 4.6. *If we let $c^\eta\equiv\{c^\eta(t)|0\le t\le T\}$ be the optimal staffing function associated with the scaled queueing process $\{Q^\eta(t)/\eta|0\le t\le T\}$ and we let $c^\eta\to c$ almost surely, then we have*

$$\lim_{\eta\to\infty}\left|\sup_{c^\eta}\int_0^T\mathcal{R}\left(\frac{Q^\eta(s)}{\eta},c^\eta(t)\right)ds-\sup_{c}\int_0^T\mathcal{R}\left(q(s),c(s)\right)ds\right|=0\quad a.s \tag{4.44}$$

Proof. First we exchange the supremum from outside to inside the integral, which gives us

$$\lim_{\eta\to\infty}\left|\sup_{c^\eta}\int_0^T\mathcal{R}\left(\frac{Q^\eta(s)}{\eta},c^\eta(s)\right)ds-\sup_{c}\int_0^T\mathcal{R}\left(q(s),c(s)\right)ds\right| \tag{4.45}$$

$$=\lim_{\eta\to\infty}\left|\int_0^T\sup_{c^\eta}\mathcal{R}\left(\frac{Q^\eta(s)}{\eta},c^\eta(s)\right)ds-\int_0^T\sup_{c}\mathcal{R}\left(q(s),c(s)\right)ds\right| \tag{4.46}$$

In addition, we know that

$$\sup_c\mathcal{R}\left(x,c\right)ds=\max_{\{i:1\le i\le 2^n\}}\mathcal{R}_i(x,e_i\circ q) \tag{4.47}$$

Now as the profit rate function is Lipschitz continuous, we can apply Lemma 4.4 to get that

$$\left|\sup_{c^\eta}\mathcal{R}\left(\frac{Q^\eta(t)}{\eta},c^\eta(t)\right)-\sup_c\mathcal{R}\left(q(t),c(t)\right)\right|\le M\cdot\left(\left|\frac{Q^\eta(t)}{\eta}-q(t)\right|+|c^\eta(t)-c(t)|\right) \tag{4.48}$$

where M is the following constant

$$M = \sum_{j=1}^{N} r_j \cdot \mu_j + w_j. \tag{4.49}$$

Now making using Lemma 4.5, the proof is completed by using the assumptions and the fluid limit from Equation 2.6. □

This theorem highlights the effectiveness of the fluid model. By optimizing the fluid model, we obtain a staffing schedule that is nearly optimal for our original stochastic Jackson network queueing model in the uniform acceleration regime. Thus, for large scale systems like call centers and hospitals where η is large, the profit that we make is near the optimal profit that we would make under the stochastic setting.

5 Optimal Control via Infinite Server Approximation

In our pursuit of a refined approximation for the queue length process we choose to use the infinite server queue as our motivation. The infinite server queue is quite natural for modeling multiserver systems that are lightly loaded or provide a high quality of service. Perhaps the most important advantage of studying the infinite server queue is that the M/G/∞ queue is very tractable, even when the arrival process is nonstationary. In the nonstationary M_t/G/∞ queue, we know from Eick et al. [6] that the queue length process has a Poission distribution with time varying rate $q^{\infty}(t)$. The exact analysis of the infinite server queue is often useful since it represents the dynamics of the queueing process as if there were an unlimited amount of resources to satisfy the demand process. As observed in Eick et al. [6], the mean of the queue length process $q^{\infty}(t)$ has the following representation

$$q^{\infty}(t) \equiv E[Q^{\infty}(t)] = \int_{-\infty}^{t} \overline{G}(t-u)\lambda(u)du = E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t-S_e)] \cdot E[S]$$

where S represents a service time with distribution G, $\overline{G} = 1 - G(t) = \mathbb{P}(S > t)$, and S_e is a random variable with distribution that follows the stationary excess of residual-lifetime cdf G_e, defined by

$$G_e(t) \equiv \mathbb{P}(S_e < t) = \frac{1}{E[S]} \int_0^t \overline{G}(u)du, \quad t \geq 0.$$

It turns out the Poisson distribution is also characterized by the fact that all of its cumulant moments are equal to its mean. Thus, we have that the mean and variance of the M_t/G/∞ queue are equal to one another when initialized with a Poisson distribution or at zero. This cumulant moment property of the M_t/G/∞ queue motivates our next approximation for the queue length process in order to incorporate stochastic effects of the queueing process. To allow for general solutions and to relax the assumption on increasing and concave operating costs, we propose approximating the queue length at station j, $Q_j(t)$, using a Gaussian random variable with equal mean and variance such that

$$Q_j(t) \approx q_j(t) + X_j \cdot \sqrt{q_j(t)} \tag{5.50}$$

Here X_j is a standard Gaussian random variable with mean 0 and variance 1 and is independent of all other X_i random variables where $i \neq j$. Then following from the methods of optimal control theory Sethi and Thompson [29], after omitting t to simplify notation, the Gaussian approximated Hamiltonian function related to Problem 3.1 is given by:

$$\begin{aligned}
\mathcal{H}(c,p,q,x) &= \sum_{i=1}^{N} \left(r_i \cdot \mu_i \cdot E[Q_i \wedge c_i] - c_i \cdot w_i\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\lambda_i - \mu_i \cdot E[Q_i] - (\theta_i - \mu_i) \cdot E[(Q_i - c_i)^+]\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot E\left[Q_j\right] + \sum_{j=1}^{N} (\theta_j \cdot \tau_{ji} - \mu_j \cdot \gamma_{ji}) \cdot E\left[(Q_j - c_j)^+\right]\right) \\
&- \sum_{i=1}^{N} x_i \cdot \theta_i \cdot E[(Q_i - c_i)^+].
\end{aligned} \tag{5.51}$$

Here p_j is the momentum variable (co-state of q_j) and x_j is the co-state variable of some auxiliary variable Z_j where

$$Z_j = -\int_0^t \theta_j \cdot E[((q_j + X_j \cdot \sqrt{q_j}) - c_j)^+]\, dt \tag{5.52}$$

$$\dot{Z}_j = -\theta_j \cdot E[((q_j + X_j \cdot \sqrt{q_j}) - c_j)^+] \tag{5.53}$$

where $Z_j(T) \geq -\mathcal{E}_j$. Since Z_j does not appear in equation 5.51 , then $\dot{x}_j = -\partial \mathcal{H} / \partial Z_j = 0$, meaning that x_j is a constant that satisfies the following complementary of slackness equation:

$$x_j \cdot \left[\mathcal{E}_j - \int_0^T \theta_j \cdot E[((q_j + X_j \cdot \sqrt{q_j}) - c_j)^+]\, dt\right] = 0 \tag{5.54}$$

Accordingly, $x_j = 0$ when $\mathcal{E}_j - \int_0^T \theta_j \cdot E[((q_j + X_j \cdot \sqrt{q_j}) - c_j)^+]\, dt > 0$, else $x_j > 0$. Now by exploiting Stein's Lemma 8.2, we obtain the following expression for the Hamiltonian in Equation 5.51.

Proposition 5.1. *By substituing $Q_j = q_j + \sqrt{q_j} \cdot X_j$ into the Hamiltonian of Equation 5.51 and taking expectations, we obtain the following Hamiltonian*

$$\begin{aligned}
\mathcal{H}(c,p,q,x) &= \sum_{i=1}^{N} \left(r_i \cdot \mu_i \cdot \left(q_i + \sqrt{q_i} \cdot \left(\chi_i \cdot \overline{\Phi}(\chi_i) - \varphi(\chi_i)\right)\right) - c_i \cdot w_i\right) \\
&+ \sum_{i=1}^{N} p_i \cdot (\lambda_i - \mu_i \cdot q_i - (\mu_i - \theta_i) \cdot \sqrt{q_i} \cdot \left(\chi_i \cdot \overline{\Phi}(\chi_i) - \varphi(\chi_i)\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot q_j + \sum_{j=1}^{N} (\theta_j \cdot \tau_{ji} - \mu_j \cdot \gamma_{ji}) \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right) \cdot \sqrt{q_j}\right) \\
&- \sum_{i=1}^{N} x_i \cdot \theta_i \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right) \cdot \sqrt{q_i}
\end{aligned} \tag{5.55}$$

Proof. It is sufficient to show that

$$E[(Q_j - c_j)^+] = \sqrt{q_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right) \tag{5.56}$$

$$E[Q_j \wedge c_j] = q_j - \sqrt{q_j} \cdot \left(\varphi(\chi_j) + \chi_j \cdot \overline{\Phi}(\chi_j)\right) \tag{5.57}$$

where

$$\chi_j \equiv \frac{c_j - q_j}{\sqrt{q_j}}. \tag{5.58}$$

This is demonstrated in Appendix 8.2, by using Stein's Lemma 8.2 for Gaussian random variables. □

Now that we have the Hamiltonian for our Gaussian approximated control problem, we can use it to compute the necessary conditions and optimal staffing procedures for the optimal control problem. In the next theorem, we present the necessary conditions of our optimal control problem, which are essential for the numerical work that we present later.

Theorem 5.2 (Necessary conditions)**.** *If for all t, $Q_j(t)$ is Gaussian with equal mean and variance with Hamiltonian given by Equation 5.55, then the queueing dynamics $q_j^*(t)$ and the corresponding momentum variable $p_j^*(t)$ satisfy the following system of differential equations for $1 \leq j \leq N$:*

$$\begin{aligned} \dot{q^*}_j &= \lambda_j - \mu_j \cdot q_j^* - (\mu_j - \theta_j) \cdot \left(\chi_j^* \cdot \overline{\Phi}(\chi_j^*) - \varphi(\chi_j^*)\right) \cdot \sqrt{q_j^*} \\ &+ \sum_{i=1}^{N} \mu_i \cdot \tau_{ij} \cdot q_i^* + \sum_{i=1}^{N} (\theta_i \cdot \gamma_{ij} - \mu_i \cdot \tau_{ij}) \cdot \left(\chi_i^* \cdot \overline{\Phi}(\chi_i^*) - \varphi(\chi_i^*)\right) \cdot \sqrt{q_i^*} \end{aligned} \tag{5.59}$$

$$\dot{p^*}_j = \left(\mu_j \cdot (r_j - p_j^*) + \theta_j \cdot (p_j^* + x_j)\right) \cdot \left(\overline{\Phi}(\chi_j^*) + \frac{\varphi(\chi_j^*)}{2\sqrt{q_j^*}}\right) - \mu_j \cdot (r_j - p_j^*) \tag{5.60}$$

Proof: The proof is given in the Appendix 8.3.

With the necessary conditions for our optimal control problem, we know what trajectory the state variables and the multiplier variables must take. In fact they solve differential equations, which depend on the optimal staffing levels or the c_j functions. We now show exactly what the optimal staffing functions are under our Gaussian assumptions for the queue length distributions.

Theorem 5.3 (Optimal Staffing Policy)**.** *The optimal control policy for the Gaussian approximated Problem 3.1 at station j is given by*

$$c_j^* = q_j^* + \Phi^{-1}(1 - \varrho_j) \cdot \sqrt{q_j^*} \tag{5.61}$$

where

$$\varrho_j = \frac{w_j}{\mu_j \cdot (r_j - p_j^*) + \theta_j \cdot (p_j^* + x_j)} \tag{5.62}$$

and where * signifies optimality.

Proof: The proof is given in the Appendix 8.4.
Observing the optimal control policy, we notice that the optimal policy is a closed form function of the state and multiplier variables. Moreover, it also depends on the parameters of the optimal control problem, which lends itself to an economic and managerial interpretation. We present the managerial insights that can be obtained from this problem in the sequel.

Remark 5.4. *The only case where the Jackson network is autonomous is when $P^s = P^a$ and $\mu_i = \theta_i$, when the queueing network has the same dynamics as an infinite server queueing network. In this case the dynamics are much simplier and the work of Massey and Whitt [20] provides more insights for this case. In fact, we can derive an optimal control policy that is exact in this particular case.*

Theorem 5.5 (Exact Optimal Staffing)**.** *The exact optimal control policy when $P^s = P^a$ and $\mu_i = \theta_i$ is given by*

$$c_j^* = \Gamma^{-1}(q_j^*, 1 - \varrho_j)$$

where $\Gamma^{-1}(q_j^*, 1-\varrho_j)$ is the inverse incomplete Gamma function with parameters $(q_j^*, 1-\varrho_j)$. Moreover, we have that

$$\begin{aligned} \varrho_j &= \frac{w_j}{\mu_j \cdot (r_j - p_j^*) + \theta_j \cdot (p_j^* + x_j)} \\ &= \frac{w_j}{\mu_j \cdot (r_j + x_j)} \end{aligned}$$

Proof. The solution when $P^s = P^a$ and $\mu_i = \theta_i$ is exact since the queue length distribution is known to be Poisson in this case. The readers are encouraged to read Massey and Whitt [20] for a proof of this fact. In order to derive the optimal staffing policy in this special case, we use the Chen-Stein Lemma of [7]. The proof of this is derived in the Appendix 8.4. □

Remark 5.6. *We should mention that it is quite standard to obtain numerical solutions for Theorems 5.2 and 5.3. One can use the forward-backward algorithm presented in Lenhart and Workman [16]. For the convenience of the reader we provide the most important steps of the forward-backward algorithm and how to implement them in Appendix 8.6.*

5.1 The dynamics of the optimal control $c_j^*(t)$

The optimal staffing policy at station j, $c_j^*(t)$, presented in Equation 5.61 of Theorem 5.3 is the recommended optimal staffing levels in the service system network. While the formula for $c_j^*(t)$ takes the form of traditional SRS models like that of Jennings et al. [12], the intuitions and insights generated from our model are quite different. First, in traditional SRS models, ϱ_j in Equation 5.62 would generally be the probability of delay, but for our model this is the cost-to-revenue ratio at station j given by

$$\varrho_j = \frac{w_j}{R_j(t)} \tag{5.63}$$

where $R_j(t) \equiv \mu_j \cdot (r_j - p_j(t)) + \theta_j \cdot (p_j(t) + x_j)$. As noted earlier $\mu_j \cdot (r_j - p_j(t))$ is the revenue rate from serviced patients at station j while $\theta_j \cdot (p_j(t) + x_j)$ is the revenue rate from abandoned patients at station j.

It should be noted that as the abandonment threshold $\epsilon_j \to 1$, the penalty costs $x_j \to 0$, which implies that $R_j(t) = \mu_j \cdot r_j + (\theta_j - \mu_j) \cdot p_j(t)$. Accordingly, a special case of $\theta_j = \mu_j$ eliminates the dependence of the optimal solution on the shadow price $p_j(t)$ from staffing policy decisions. Moreover, Φ^{-1}, is often interpreted as the quality service grade in traditional SRS models Mandelbaum and Zeltyn [17]. For our model, Φ^{-1} can be viewed more as the a profitability-and-quality service grade since it's not only dictated by the abandonment threshold ϵ, but also by the objective of maximizing profitability.

5.2 Profitability Analysis of the optimal control

A Hamiltonian function in optimal control problems can be considered a profit rate Sethi and Thompson [29] and Chiang [3]. For our control problem, this is Equation 5.51, after applying the Gaussian refinement. The first term of our Hamiltonian function, $r_j \cdot \mu_j \cdot \left(q_j + \sqrt{q_j} \cdot \left(\chi_j \cdot \overline{\Phi}(\chi_j) - \varphi(\chi_j)\right)\right) - w_j \cdot c_j$, represents the realized profit or simply the net operating income from patient care reimbursements. Here $r_j \cdot \mu_j \cdot \left(q_j + \sqrt{q_j} \cdot \left(\chi_j \cdot \overline{\Phi}(\chi_j) - \varphi(\chi_j)\right)\right)$ is the approximate operating revenue and $w_j \cdot c_j$ is the operating cost. For the center to be profitable, it's imperative that $r_j > c_j$. To measure the degree of profitability, we follow Gapenski [8] and define the operating margin for station j, $\mathcal{O}_j$, as follows:

$$\mathcal{O}_j = \frac{\int_0^T r_j \cdot \mu_j \cdot \left(q_j + \sqrt{q_j} \cdot \left(\chi_j \cdot \overline{\Phi}(\chi_j) - \varphi(\chi_j)\right)\right) - w_j \cdot c_j}{\int_0^T r_j \cdot \mu_j \cdot \left(q_j + \sqrt{q_j} \cdot \left(\chi_j \cdot \overline{\Phi}(\chi_j) - \varphi(\chi_j)\right)\right)} \tag{5.64}$$

The closer to 1 the operating margin $\mathcal{O}_j$ is, the more profitable that station j is.

5.3 Probability of Waiting for Service

The probability of delay or waiting for service at each individual station can be calculated by the probability that the queue length at that station is larger than the number of servers at time t. Using the infinite server approximation again, we have the following approximation for this quantity in terms of the parameters of our optimal control problem

$$\begin{aligned}
\mathbb{P}(Q_j(t) \geq c_j(t)) &\approx \mathbb{P}(Q_j^{\infty}(t) \geq c_j(t)) && (5.65)\\
&\approx \mathbb{P}\left(q_j + \sqrt{q_j} \cdot X_j \geq q_j + \sqrt{q_j} \cdot \Phi^{-1}(1 - \varrho_j)\right) && (5.66)\\
&= \mathbb{P}\left(X_j \geq \Phi^{-1}(1 - \varrho_j)\right) && (5.67)\\
&= 1 - \Phi\left(\Phi^{-1}(1 - \varrho_j)\right) && (5.68)\\
&= \varrho_j && (5.69)
\end{aligned}$$

where ϱ_j is the cost-to-revenue ratio at station j. Thus, we can give a performance measure interpretation of the delay experienced by customers at station j in terms of the cost to revenue ratio at that station. Thus, as the cost to revenue ratio increases, we should experience more delay since it is more expensive for the service provider to provide better

service. Similarly, when the cost to revenue ratio goes down we can staff the system with more servers at a low cost, which causes the delay probability to decrease. This result is important as it gives us a profitability interpretation of the delay costs.

6 Numerical Results

In this section, we present numerical results for both the fluid control and the Gaussian control problems.

6.1 Fluid Model

In Figure 4, we show the results from the fluid optimal control method where we dynamically control the number of servers in two queues. Going from the top to the bottom of the top graph in 4 we plot for the first queue, the arrival rate, the queue length, the optimal staffing function, the adjoint variable p, and the probability of delay. In the bottom graph of 4, we plot the same thing, however, for the second queue. We see that the optimal solution is to staff no one for all of the time. This can be quite problematic in healthcare staffing since this could risk patients lives, however, it does indicate that the hospital might not be as profitable as it could and perhaps it should be shut down to not lose money. In Figure **??**, we use the same parameters, however, we scale up the arrival rate. We see once again that the optimal procedure is to do nothing and to shut down service. Although, we do not add this plot, by increasing the revenue generated from the service of each customer, we can construct a solution that staffs only at the mean level of the queue length.

6.2 Infinite Server Approximation

In Figure 6, we show the results from the Gaussian optimal control method where we dynamically control the number of servers in two queues. Going from the top to the bottom of the top graph in 6 we plot for the first queue, the arrival rate, the queue length, the optimal staffing function, the adjoint variable p, and the probability of delay. In the bottom graph of 6, we plot the same thing, however, for the second queue. We see that the optimal solution is to staff according to our square root staffing policy. We use the same parameters from the fluid model plots, however, in contrast to the fluid control, you should not staff zero servers. The mean number of customers is a baseline and you staff around the baseline according to the cost to revenue ratio at each station. From a managerial perspective, this is a great option since it clear relates the ratio of profitability to the staffing number in a elegant way. In Figure 7, we use the same parameters, however, we scale up the arrival rate. We see once again as our theoretical results predict, the optimal solution is to do square root staffing. It is also important to note that the staffing in the Gaussian approximation somewhat stabilizes the probability of delay as the revenue to cost ratio. This is a nice result since the delay that customers experience is directly related to the revenue and costs that the system incurs. We will once again mention that the code for optimal staffing is available on the first author's website.

7 Conclusion and Additional Research

Dynamic control of nonstationary queues is a very difficult problem. In this work, we have proposed two different solutions to tackle this problem. Our first method is to use a fluid limit approximation to approximate the queueing control problem as a dynamical systems control problem using the fluid limit of the queueing process. In this setting we proved that the fluid control problem is asypmptotically optimal for the original queueing process.

Our second method exploits the infinite server queueing process and its tractability. By approximating the queue length process with a Gaussian distribution where the mean is equal to the variance, we proved the optimal control policy is the celebrated square root staffing policy. Moreover, the square root staffing policy for each station only depends on the local parameters of each station and not the other parameters. Unlike in most traditional square root staffing formulas, the main parameter in our formula was not the probability of delay but rather a cost-to-revenue ratio that depends on the shadow price of each station. We also showed that the probability of delay can be expressed in terms of this cost-to-revenue ratio. One main insight is that as the cost-to-revenue ratio increased, customers experienced more delay since it was more expensive for the center to increase the number of servers. Thus, we show a unique relationship between profitability of the queueing network and the probability of delay in the network.

There are several avenues that are still open for further research. One main idea is to use other queueing approximation techniques such as spectral and orthogonal polynomial approximations as in Massey and Pender [19], Engblom and Pender [7], Pender and Massey [28] to approximate the queue length process. Also an extension to more complicated networks of queues like priority queues for example would be an interesting extension, especially in the healthcare literature. Another area for further research is to consider optimal staffing of queues with non-Markovian dynamics like in the work of Ko and Pender [15, 14], Pender and Ko [27]. Lastly, it would be interesting to use risk measures like in the work of Pender [26] and generate optimal control policies for nonstationary queues using risk measures in the objective function or constraints.

8 Appendix

8.1 Stein's Lemma

One important property that will be useful for calculating the optimal control solution is the following derivative property of min and max functions.

Lemma 8.1. *Let Q_j be any random variable and $c_j(t)$ be a deterministic function of time, then we have that*

$$\frac{\partial}{\partial c_j} E[Q_j \wedge c_j] = -\frac{\partial}{\partial c_j} E[(Q_j - c_j)^+]. \tag{8.70}$$

In addition to understanding the derivative of the Hamiltonian, we first need to compute the expectations of Problem 3.1. To tackle this problem, we exploit Stein's Lemma for Gaussian random variables. Stein's Lemma states the following:

Lemma 8.2 (Stein [30]). *X is a standard Gaussian random variable mean 0 and variance 1 if and only if*

$$E[X \cdot f(X)] = E\left[f'(X)\right]$$

for all generalized functions that satisfy $E\left[f'(X)\right] < \infty$.

In this section, we present many of the proofs and derivations needed to compute our optimal staffing schedule. We begin with the derivation of the Hamiltonian.

8.2 Derivation of the Hamiltonian in the Equation 5.51

In this section of the Appendix, we derive the Hamiltonian equation that is presented in Equation 5.51. We use Stein's lemma to derive the Hamiltonian and our derivation will allow us to calculate our optimal staffing schedule. From our optimal control problem formulation, we know that the Hamiltonian is given by

$$\begin{aligned}
\mathcal{H}(c,p,q,x) &= \sum_{i=1}^{N} \left(r_i \cdot \mu_i \cdot E[Q_i \wedge c_i] - c_i \cdot w_i\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\lambda_i - \mu_i \cdot E[Q_i] - (\theta_i - \mu_i) \cdot E[(Q_i - c_i)^+]\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot E\left[Q_j\right] + \sum_{j=1}^{N} (\theta_j \cdot \tau_{ji} - \mu_j \cdot \gamma_{ji}) \cdot E\left[(Q_j - c_j)^+\right]\right) \\
&- \sum_{i=1}^{N} x_i \cdot \theta_i \cdot E[(Q_i - c_i)^+]. \\
&\approx \sum_{i=1}^{N} \left(r_i \cdot \mu_i \cdot E[(q_i + \sqrt{q_i} \cdot X_i) \wedge c_i] - c_i \cdot w_i\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\lambda_i - \mu_i \cdot E[(q_i + \sqrt{q_i} \cdot X_i)] - (\theta_i - \mu_i) \cdot E[((q_i + \sqrt{q_i} \cdot X_i) - c_i)^+]\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot E\left[(q_j + \sqrt{q_j} \cdot X_j)\right]\right) \\
&+ \sum_{i=1}^{N} p_i \cdot \left(\sum_{j=1}^{N} (\theta_j \cdot \tau_{ji} - \mu_j \cdot \gamma_{ji}) \cdot E\left[((q_j + \sqrt{q_j} \cdot X_j) - c_j)^+\right]\right) \\
&- \sum_{i=1}^{N} x_i \cdot \theta_i \cdot E[((q_i + \sqrt{q_i} \cdot X_i) - c_i)^+].
\end{aligned}$$

Using Stein's Lemma 8.2, we derive the following expression for the maximum function

for a standard Gaussian random variable

$$\begin{aligned} E\left[(X-\chi_j)^+\right] &= E\left[(X-\chi_j)\cdot\{X\geq\chi_j\}\right] \\ &= E\left[X\cdot\{X\geq\chi_j\}\right]-\chi_j\cdot P\{X\geq\chi_j\} \\ &= \int_{-\infty}^{\infty}\delta_{\chi_j}(y)\cdot\varphi(y)dy-\chi_j\cdot\overline{\Phi}(\chi_j) \\ &= \varphi(\chi_j)-\chi_j\cdot\overline{\Phi}(\chi_j). \end{aligned} \tag{8.71}$$

Moreover, since the max and min functions satisfy the following equality

$$E\left[X\wedge\chi_j\right] = E\left[X\right]-E\left[(X-\chi_j)^+\right],$$

we have that the min function of a standard Gaussian random variable satisfies the following expression

$$E\left[X\wedge\chi_j\right] = E\left[X-(X-\chi_j)^+\right]=-E\left[(X-\chi_j)^+\right]=\chi_j\cdot\overline{\Phi}(\chi_j)-\varphi(\chi_j). \tag{8.72}$$

After substituting the approximate expectations 8.71 and 8.72 of the queue lengths into Equation 5.51, our Hamiltonian function becomes:

$$\begin{aligned} \mathcal{H}(c,p,q,x) &= \sum_{i=1}^{N}\left(r_i\cdot\mu_i\cdot\left(q_i+\sqrt{q_i}\cdot\left(\chi_i\cdot\overline{\Phi}(\chi_i)-\varphi(\chi_i)\right)\right)-c_i\cdot w_i\right) \\ &+ \sum_{i=1}^{N}p_i\cdot\left(\lambda_i-\mu_i\cdot q_i-(\mu_i-\theta_i)\cdot\sqrt{q_i}\cdot\left(\chi_i\cdot\overline{\Phi}(\chi_i)-\varphi(\chi_i)\right)\right) \\ &+ \sum_{i=1}^{N}p_i\cdot\left(\sum_{j=1}^{N}\mu_j\cdot\gamma_{ji}\cdot q_j+\sum_{j=1}^{N}(\theta_j\cdot\tau_{ji}-\mu_j\cdot\gamma_{ji})\cdot\left(\varphi(\chi_j)-\chi_j\cdot\overline{\Phi}(\chi_j)\right)\cdot\sqrt{q_j}\right) \\ &- \sum_{i=1}^{N}x_i\cdot\theta_i\cdot\left(\varphi(\chi_i)-\chi_i\cdot\overline{\Phi}(\chi_i)\right)\cdot\sqrt{q_i} \end{aligned} \tag{8.73}$$

8.3 Proof of Theorem 5.2

In order to prove that our optimal staffing schedule is optimal, we need the state variables and the Lagrange multipliers to satisfy the necessary conditions of the Pontryagin maximum principle. In our case, it suffices to calculate the partial derivatives of the Hamiltonian with respect to the state and multiplier variables. In this section we compute the partial derivatives of the Hamiltonian and use them to construct the differential equations that

describe the dynamics of the queue lengths and the opportunity costs or shadow prices. From our earlier analysis, we know the Hamiltonian can be written as

$$\begin{aligned}
\mathcal{H}(c,p,q,x) &= \sum_{i=1}^{N}\Big(r_i\cdot\mu_i\cdot\big(q_i+\sqrt{q_i}\cdot\big(\chi_i\cdot\overline{\Phi}(\chi_i)-\varphi(\chi_i)\big)\big)-c_i\cdot w_i\Big) \qquad (8.74)\\
&+ \sum_{i=1}^{N} p_i\cdot\big(\lambda_i-\mu_i\cdot q_i-(\mu_i-\theta_i)\cdot\sqrt{q_i}\cdot\big(\chi_i\cdot\overline{\Phi}(\chi_i)-\varphi(\chi_i)\big)\\
&+ \sum_{i=1}^{N} p_i\cdot\left(\sum_{j=1}^{N}\mu_j\cdot\gamma_{ji}\cdot q_j+\sum_{j=1}^{N}(\theta_j\cdot\tau_{ji}-\mu_j\cdot\gamma_{ji})\cdot\big(\varphi(\chi_j)-\chi_j\cdot\overline{\Phi}(\chi_j)\big)\cdot\sqrt{q_j}\right)\\
&- \sum_{i=1}^{N} x_i\cdot\theta_i\cdot\big(\varphi(\chi_i)-\chi_i\cdot\overline{\Phi}(\chi_i)\big)\cdot\sqrt{q_i}
\end{aligned}$$

Thus, the partial derivatives of the Hamiltonian can be computed by computing the derivatives of the Gaussian pdf and cdf functions that appear in the Hamiltonian. They are

$$\begin{aligned}
\frac{\partial\mathcal{H}}{\partial p_j}\equiv\dot{q}_j &= \lambda_j-\mu_j\cdot q_j-(\mu_j-\theta_j)\cdot\big(\chi_j\cdot\overline{\Phi}(\chi_j)-\varphi(\chi_j)\big)\cdot\sqrt{q_j} \qquad (8.75)\\
&+ \sum_{i=1}^{N}\mu_i\cdot\tau_{ij}\cdot q_i+\sum_{i=1}^{N}(\theta_i\cdot\gamma_{ij}-\mu_i\cdot\tau_{ij})\cdot\big(\chi_i\cdot\overline{\Phi}(\chi_i)-\varphi(\chi_i)\big)\cdot\sqrt{q_i}\\
-\frac{\partial\mathcal{H}}{\partial q_j}\equiv\dot{p}_j &= (\mu_j\cdot(r_j-p_j)+\theta_j\cdot(p_j+x_j))\cdot\left(\overline{\Phi}(\chi_j)+\frac{\varphi(\chi_j)}{2\sqrt{q_j}}\right) \qquad (8.76)\\
&- \mu_j\cdot(r_j-p_j)
\end{aligned}$$

8.4 Proof of Theorem 5.3

From the Pontryagin's *Maximum Principle*, the optimal control policy c^* that maximizes the Hamiltonian function in Equation 8.74, such that $\mathcal{H}(c^*,p^*,q^*,x^*,t)\geq\mathcal{H}(c,p,q,x,t)$, is obtained by $\frac{\partial\mathcal{H}}{\partial c_j}=0$. Thus, by differentiating the Hamiltonian by c_j we obtain the following expression

$$\begin{aligned}
\frac{\partial\mathcal{H}}{\partial c_j} &= r_j\cdot\mu_j\cdot\overline{\Phi}(\chi_j)-w_j-p_j\cdot(\mu_j-\theta_j)\cdot\overline{\Phi}(\chi_j)+x_j\cdot\theta_j\cdot\overline{\Phi}(\chi_j)=0\\
&= (\mu_j\cdot(r_j-p_j)+\theta_j\cdot(p_j+x_j))\cdot\overline{\Phi}(\chi_j)-w_j=0 \qquad (8.77)
\end{aligned}$$

We now solve for c_j given that $\chi_j = \frac{c_j - q_j}{\sqrt{q_j}}$:

$$\begin{aligned}
(\mu_j \cdot (r_j - p_j) + \theta_j \cdot (p_j + x_j)) \cdot \overline{\Phi}(\chi_j) &= w_j \\
\overline{\Phi}\left(\frac{c_j - q_j}{\sqrt{q_j}}\right) &= \frac{w_j}{\mu_j \cdot (r_j - p_j) + \theta_j \cdot (p_j + x_j)} \\
\Phi\left(\frac{c_j - q_j}{\sqrt{q_j}}\right) &= 1 - \frac{w_j}{\mu_j \cdot (r_j - p_j) + \theta_j \cdot (p_j + x_j)} \\
\frac{c_j - q_j}{\sqrt{q_j}} &= \Phi^{-1}\left(1 - \frac{w_j}{\mu_j \cdot (r_j - p_j) + \theta_j \cdot (p_j + x_j)}\right)
\end{aligned}$$

Finally we obtain optimal control policy c_j^* as:

$$c_j^* = q_j + \Phi^{-1}\left(1 - \frac{w_j}{\mu_j \cdot (r_j - p_j) + \theta_j \cdot (p_j + x_j)}\right) \cdot \sqrt{q_j}. \tag{8.78}$$

8.5 Proof of Theorem 5.5

Theorem 8.3 (Chen-Stein)**.** *Let Q be a random variable with values in $\mathbb{N}$. Then, Q has the Poisson distribution with mean rate q if and only if, for every bounded function $f : \mathbb{N} \to \mathbb{N}$,*

$$\mathbb{E}\left[Q \cdot f(Q)\right] = q \cdot \mathbb{E}\left[f(Q+1)\right]$$

Proof. See Peccati and Taqqu [24]. □

Lemma 8.4.

$$\Gamma(s,x) = \sum_{m=s}^{\infty} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(s)} \int_0^x e^{-y} y^{s-1} dy$$

$$\overline{\Gamma}(s,x) = \sum_{m=0}^{s-1} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(s)} \int_x^{\infty} e^{-y} y^{s-1} dy.$$

where

$$\Gamma(s,x) = \frac{1}{\Gamma(s)} \int_0^x e^{-y} y^{s-1} dy \quad \text{and} \quad \overline{\Gamma}(s,x) = \frac{1}{\Gamma(s)} \int_x^{\infty} e^{-y} y^{s-1} dy$$

are the lower and upper incomplete gamma functions respectively. Moreover, we define $\Gamma^{-1}(x,\epsilon)$ and $\overline{\Gamma}^{-1}(x,\epsilon)$ to be the functional inverses of $\Gamma(s,x)$ and $\overline{\Gamma}(s,x)$ respectively.

Proof. See Janssen et al. [11]. □

Lemma 8.5.

$$\begin{aligned}
E[(Q_j - c_j)^+] &= E[(Q_j - c_j) \cdot \{Q > c_j\}] \\
&= E[Q_j \cdot \{Q > c_j\}] - s \cdot E[\{Q > c_j\}] \\
&= E[Q_j \cdot \{Q > c_j\}] - s \cdot \Gamma(c_j + 1, q) \\
&= q_j \cdot E[\{Q + 1 > c_j\}] - c_j \cdot \Gamma(c_j + 1, q_j) \\
&= q_j \cdot \Gamma(c_j, q_j) - c_j \cdot \Gamma(c_j + 1, q_j)
\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial c_j}E[(Q_j - c_j)^+] &= -\Gamma(c_j+1, q_j)\\ \frac{\partial}{\partial c_j}E[Q_j \wedge c_j] &= \Gamma(c_j+1, q_j)\end{aligned}$$

Now we prove Theorem 5.5 using the results of the profit model. In the case where $P^s = P^a$ and $\mu_i = \theta_i$, we have that

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial c_j} &= r\cdot\mu\cdot\Gamma(c_j+1,q_j) - w_j - p_j\cdot(\mu-\theta)\cdot\Gamma(c_j+1,q_j) + \mathrm{x}\cdot\theta\cdot\Gamma(c_j+1,q_j) = 0\\ &= \mu_j\cdot(r_j+\mathrm{x}_j)\cdot\Gamma(c_j+1,q) - w_j = 0.\end{aligned}$$

We now solve for c_j given that $\chi_j = \frac{c_j - q_j}{\sqrt{q_j}}$:

$$\begin{aligned}\mu_j\cdot(r_j+\mathrm{x}_j)\cdot\Gamma(c_j+1,q_j) - w_j &= 0\\ \Gamma(c_j+1,q_j) &= \frac{w_j}{\mu_j\cdot(r_j+\mathrm{x}_j)}.\end{aligned}$$

Finally by inverting the incomplete gamma function, we obtain optimal control policy c_j^* as

$$c_j^* = \Gamma^{-1}\left(q_j, \frac{w_j}{\mu_j\cdot(r_j+x_j)}\right).$$

8.6 Numerical Algorithms for Optimal Control

In this section, we explain the numerical algorithm that is needed to construct the dynamical systems that are needed to produce the optimal solution. We cannot obtain the close form solution of the optimal staffing policy indicated in equation 8.79 as $c(t)$, $q(t)$, and $p(t)$ still depend on time. To obtained needed solutions, we use the standard Euler scheme for integrating differential equations and the Forward-Backward method as described in Algorithm 8.6. When solving for $c_j(t)$,$q_j(t)$, and $p_j(t)$, Algorithm 8.6 follows from the Forward-Backward method introduced in Lenhart and Workman [16]. Our method is a slight modification in that we use standard Euler and not the Runge-Kutta method for numerically integrating equations. Here we add steps to also solving for the constant x_j when there are inequality constraints in the optimal control problem.

Algorithm 8.6 (Steps to numerically solve for $c_j(t)$, $q_j(t)$, $p_j(t)$, and x_j)**.**

Step 0: Set initial conditions for $q(0)$ and terminal conditions for $p(T)$ and the initial guess of the control policy $\overrightarrow{c}(t)$, for all $0 < t < T$. Also initialize number of iterations $n = 0$ and the abandonment penalty $x = 0$

Step 1: Given $\{q_{n-1}(t)|0 \le t \le T\}$, solve the dynamical system $\dot{p}(t) = -\frac{\partial \mathcal{H}}{\partial q}(p_n, q_{n-1})(t)$ backward in time for all $0 \le t \le T$, starting with the terminal condition $p_n(T) = 0$

Step 2: Given $\{p_n(t)|0 \leq t \leq T\}$, solve the dynamical system $\dot{q}(t) = \frac{\partial \mathcal{H}}{\partial p}(p_n, q_n)(t)$ forward in time for all $0 \leq t \leq T$, starting with the initial condition $q_n(0) = q^0$

Step 3: For all $0 < t < T$, determine the staffing policy s_n by solving

$$c_n(t) = q_n(t) + \Phi^{-1}\left(1 - \varrho_n(t)\right) * \sqrt{(q_n)}$$

Step 4: If $\mathcal{E} - \int_0^T \theta \cdot (q_n(t) - c_n(t))^+ < 0, \forall \quad 0 < t < T$

1. n=n+1
2. $\mathrm{x}_{n+1} = x_n + \mathrm{h}$

where h is a very small number.

Step 5: Repeat Step 1-3 until the relative error is negligible, in that:

$$\int_0^T \theta \cdot (q_n(t) - c_n(t))^+ < \mathcal{E} \quad \text{and} \quad \frac{\|\overrightarrow{c}\|_n - \|\overrightarrow{c}\|_{n-1}}{\|\overrightarrow{c}\|_n} \leq \delta$$

where δ is the accepted convergence tolerance.

References

[1] N. Bäuerle and U. Rieder. Optimal control of single-server fluid networks. *Queueing Systems*, 35(1-4):185–200, 2000.

[2] N. Bäuerle et al. Optimal control of queueing networks: an approach via fluid models. *Advances in Applied Probability*, 34(2):313–328, 2002.

[3] A. Chiang. Elements of dynamic optimization. *Illinois: Waveland Press Inc*, 2000.

[4] M. Cudina and K. Ramanan. Asymptotically optimal controls for time-inhomogeneous networks. *SIAM Journal on Control and Optimization*, 49(2):611–645, 2011.

[5] F. De Véricourt and O. B. Jennings. Dimensioning large-scale membership services. *Operations Research*, 56(1):173–187, 2008.

[6] S. G. Eick, W. A. Massey, and W. Whitt. The physics of the mt/g/ queue. *Operations Research*, 41(4):731–742, 1993.

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

[8] C. Gapenski. *Fundamentals of Healthcare Finance, Second Edition*. Health Administration Press, 2012.

[9] R. Hampshire. Dynamic queueing models for the operations management of communication services, March 2007.

[10] R. C. Hampshire, O. B. Jennings, and W. A. Massey. A time-varying call center design via lagrangian mechanics. *Probability in the Engineering and Informational Sciences*, 23(02):231–259, 2009.

[11] A. Janssen, J. Van Leeuwaarden, and B. Zwart. Gaussian expansions and bounds for the poisson distribution applied to the erlang b formula. *Advances in Applied Probability*, pages 122–143, 2008.

[12] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[13] P. Khudyakov, P. D. Feigin, and A. Mandelbaum. Designing a call center with an ivr (interactive voice response). *Queueing Systems*, 66(3):215–237, 2010.

[14] Y. M. Ko and J. Pender. Diffusion limits for the (mapt/ph t/) n queueing network. 2016.

[15] Y. M. Ko and J. Pender. Strong approximations for time varying infinite-server queues with non-renewal arrival and service processes. 2016.

[16] S. Lenhart and J. T. Workman. *Optimal control applied to biological models*. CRC Press, 2007.

[17] A. Mandelbaum and S. Zeltyn. The palm/erlang-a queue, with applications to call centers. *Faculty of Industrial Engineering & Management, Technion, Haifa, Israel*, 2005.

[18] A. Mandelbaum, W. A. Massey, and M. I. Reiman. Strong approximations for markovian service networks. *Queueing Systems*, 30(1-2):149–201, 1998.

[19] W. A. Massey and J. Pender. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75(2-4):243–277, 2013.

[20] W. A. Massey and W. Whitt. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems*, 13(1-3):183–250, 1993.

[21] Y. Nazarathy and G. Weiss. Near optimal control of queueing networks over a finite time horizon. *Annals of Operations Research*, 170(1):233–249, 2009.

[22] J. Niyirora and J. Pender. Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)*, 2016.

[23] G. Pang and M. V. Day. Fluid limits of optimally controlled queueing networks. *International Journal of Stochastic Analysis*, 2007, 2007.

[24] G. Peccati and M. Taqqu. *Wiener Chaos: Moments, Cumulants and Diagrams: A survey with Computer Implementation*, volume 1. Springer Science & Business Media, 2011.

[25] J. Pender. Laguerre polynomial approximations for nonstationary queues. 2014.

[26] J. Pender. Risk measures and their application to staffing nonstationary service systems. *European Journal of Operational Research*, 254(1):113–126, 2016.

[27] J. Pender and Y. M. Ko. Approximations for the queue length distributions of time-varying many-server queues. 2016.

[28] J. Pender and W. A. Massey. Approximating and stabilizing dynamic rate jackson networks with abandonment. *Probability in the Engineering and Informational Sciences*, 31(1):1–42, 2017.

[29] S. P. Sethi and G. L. Thompson. *Optimal Control Theory: Applications to Management Science and Economics*. Springer, 2005.

[30] C. Stein. Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7:i–164, 1986.

[31] F. d. Véricourt and O. B. Jennings. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

[32] G. B. Yom-Tov and A. Mandelbaum. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

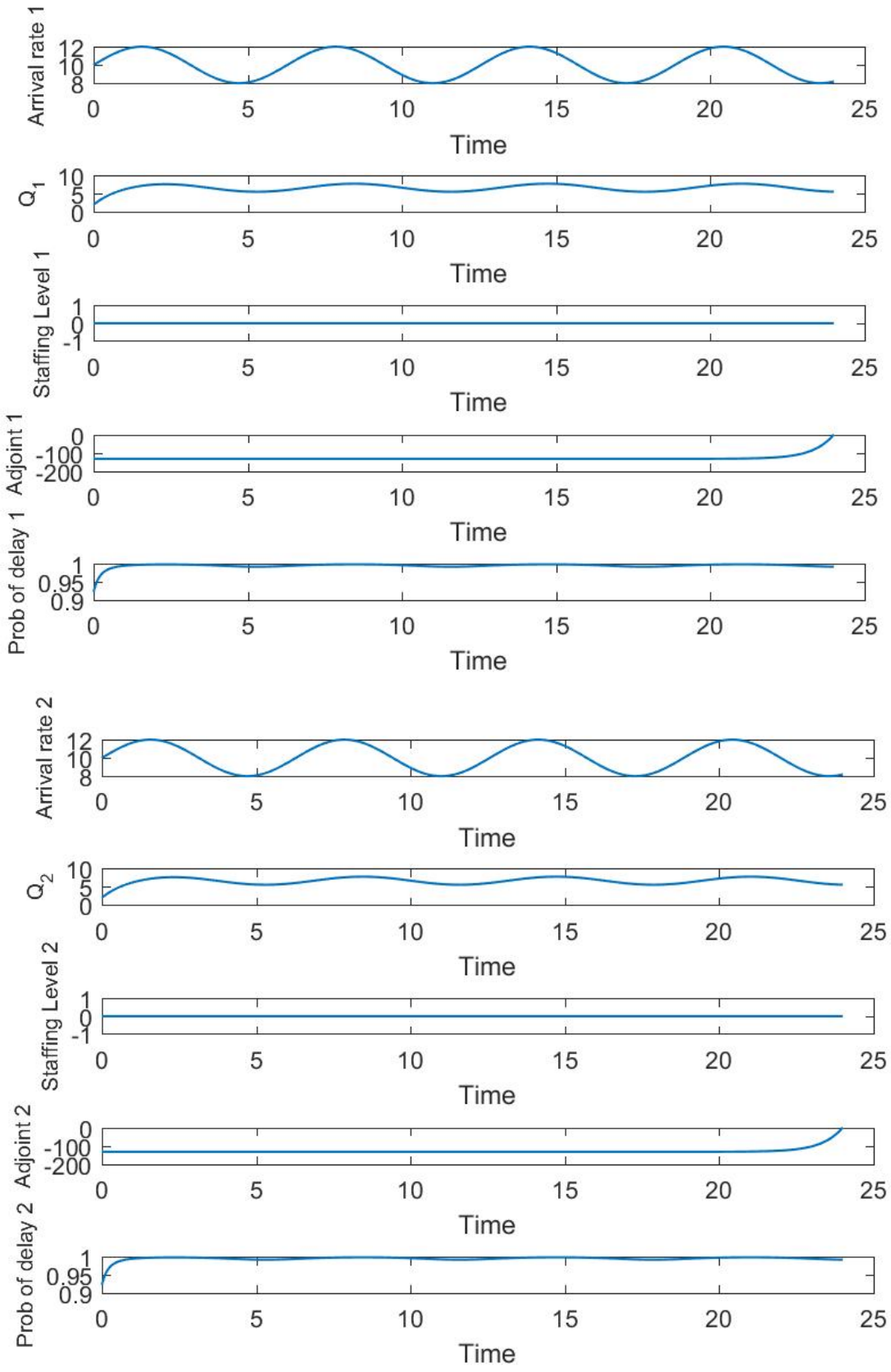

Figure 4: $\lambda_1 = \lambda_2 = 10 + 2\sin(t)$, $\mu_1 = \mu_2 = 1$, $\theta_1 = \theta_2 = 2$, $r_1 = r_2 = 20 + 10\sin(t)$ $w_1 = w_2 = 10$, $x_1 = x_2 = 100$ $\tau_{11} = \tau_{22} = 0$, $\tau_{12} = \tau_{21} = .25$, $\tau_{10} = \tau_{20} = .75$ $\gamma_{11} = \gamma_{22} = 0$, $\gamma_{12} = \gamma_{21} = .25$, $\gamma_{10} = \gamma_{20} = .75$.

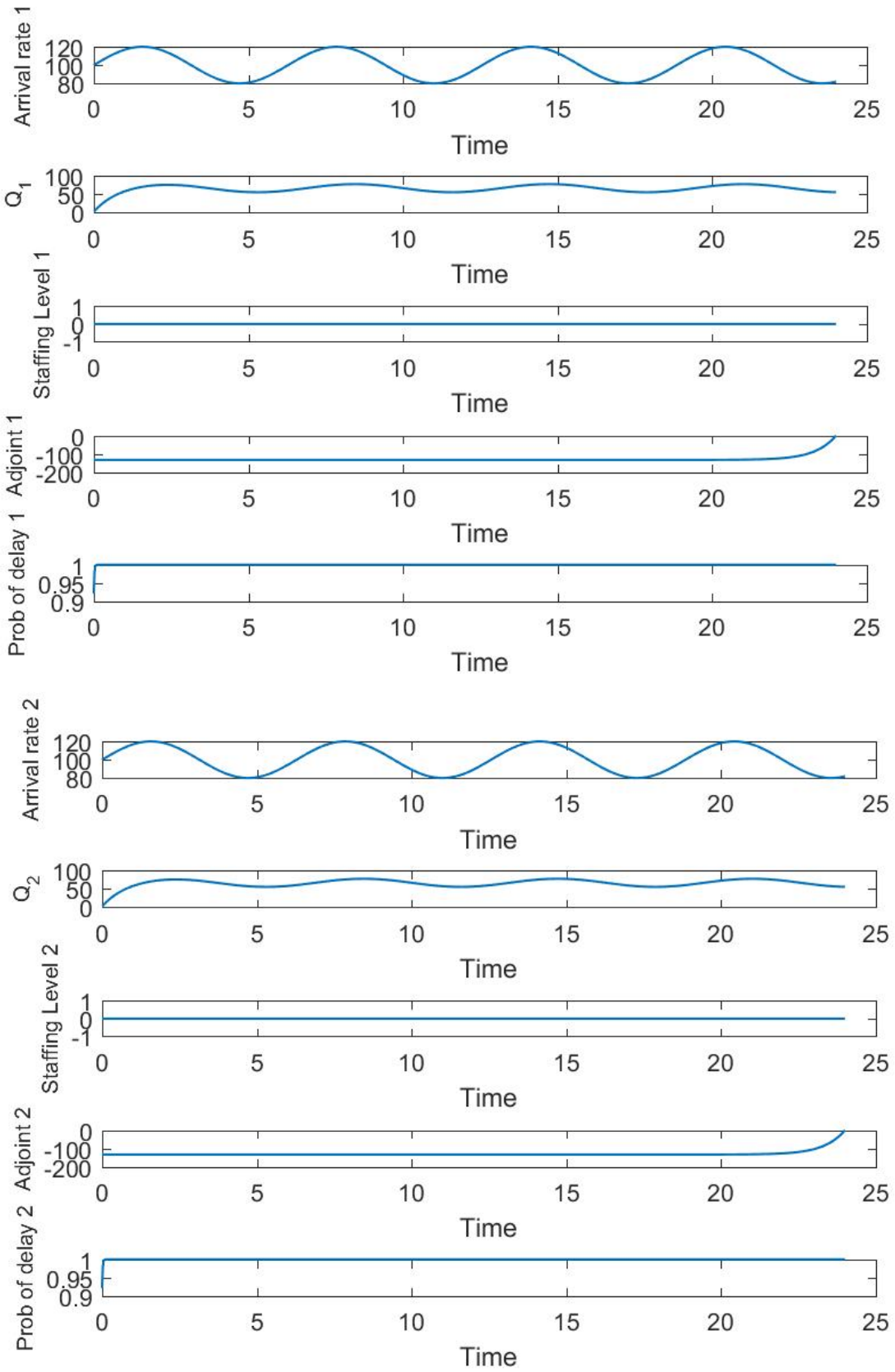

Figure 5: $\lambda_1 = \lambda_2 = 100 + 20\sin(t)$, $\mu_1 = \mu_2 = 1$, $\theta_1 = \theta_2 = 2$, $r_1 = r_2 = 20 + 10\sin(t)$ $w_1 = w_2 = 10$, $x_1 = x_2 = 100$ $\tau_{11} = \tau_{22} = 0$, $\tau_{12} = \tau_{21} = .25$, $\tau_{10} = \tau_{20} = .75$ $\gamma_{11} = \gamma_{22} = 0$, $\gamma_{12} = \gamma_{21} = .25$, $\gamma_{10} = \gamma_{20} = .75$.

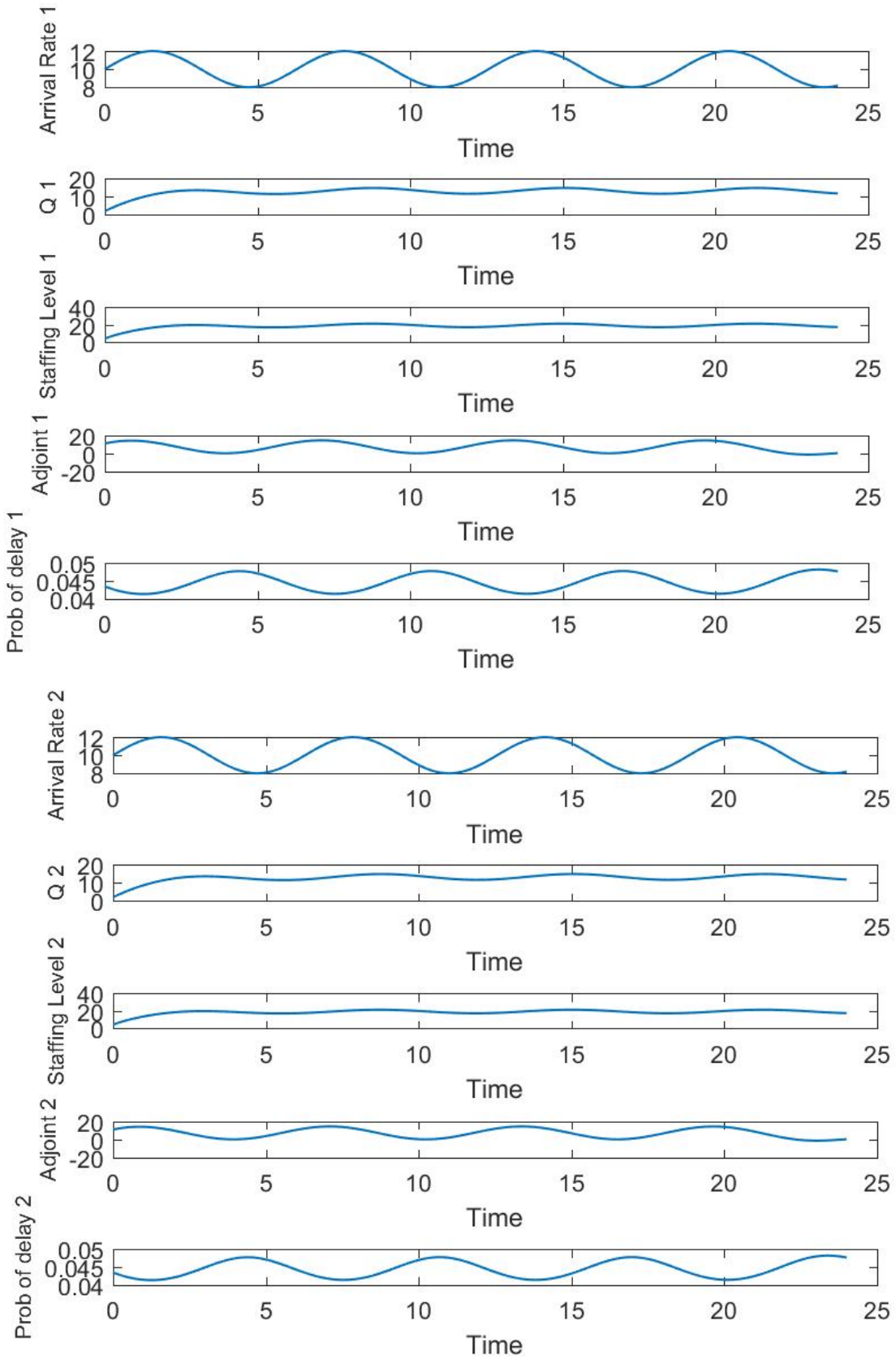

Figure 6: $\lambda_1 = \lambda_2 = 10 + 2\sin(t)$, $\mu_1 = \mu_2 = 1$, $\theta_1 = \theta_2 = 2$, $r_1 = r_2 = 20 + 10\sin(t)$ $w_1 = w_2 = 10$, $x_1 = x_2 = 100$ $\tau_{11} = \tau_{22} = 0$, $\tau_{12} = \tau_{21} = .25$, $\tau_{10} = \tau_{20} = .75$ $\gamma_{11} = \gamma_{22} = 0$, $\gamma_{12} = \gamma_{21} = .25$, $\gamma_{10} = \gamma_{20} = .75$.

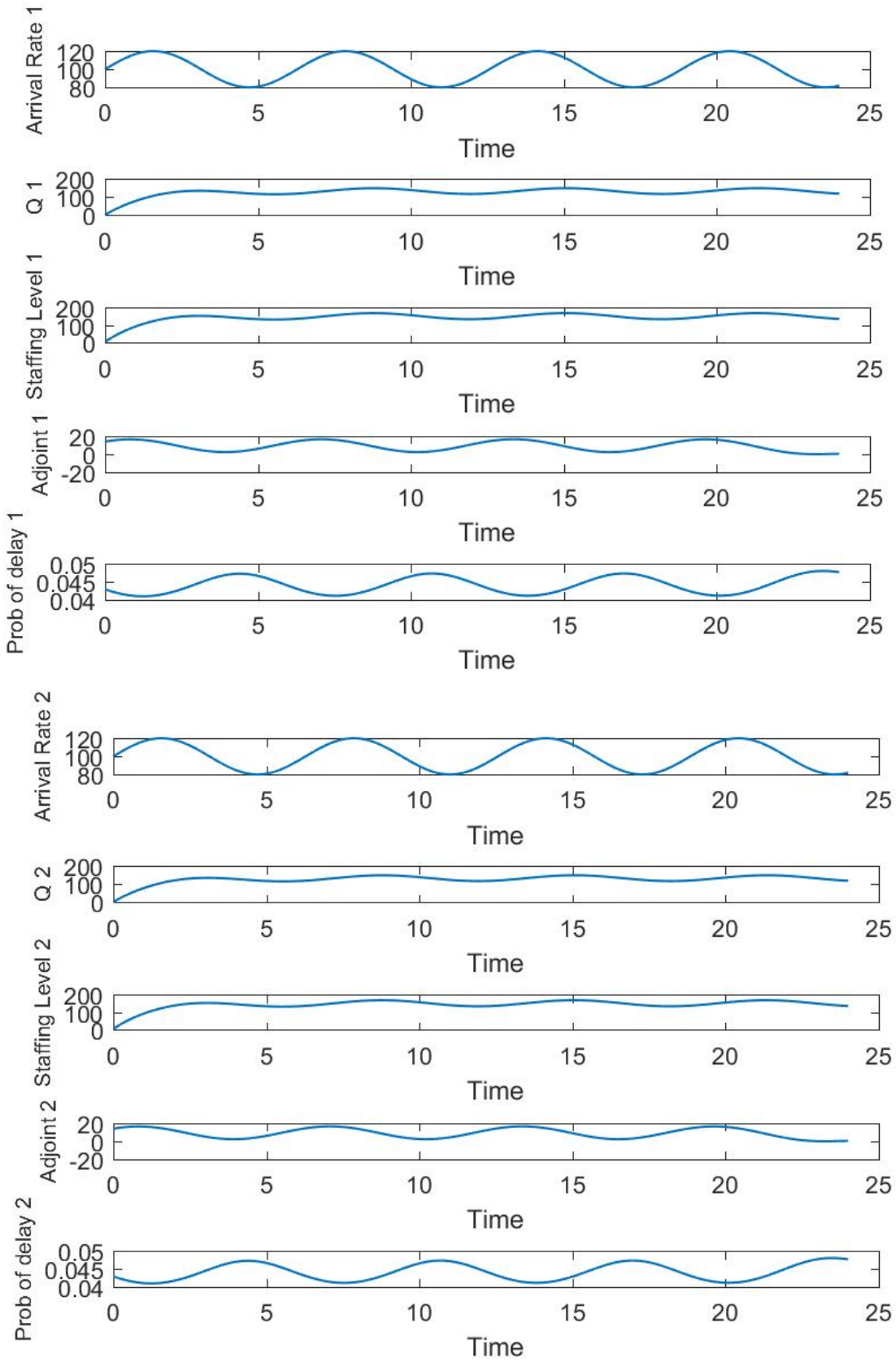

Figure 7: $\lambda_1 = \lambda_2 = 100 + 20\sin(t)$, $\mu_1 = \mu_2 = 1$, $\theta_1 = \theta_2 = 2$, $r_1 = r_2 = 20 + 10\sin(t)$ $w_1 = w_2 = 10$, $x_1 = x_2 = 100$ $\tau_{11} = \tau_{22} = 0$, $\tau_{12} = \tau_{21} = .25$, $\tau_{10} = \tau_{20} = .75$ $\gamma_{11} = \gamma_{22} = 0$, $\gamma_{12} = \gamma_{21} = .25$, $\gamma_{10} = \gamma_{20} = .75$.