Optimal Staffing in Service Centers

Jerome Niyirora
SUNY Polytechnic Institute
College of Health Sciences and Management
jerome.niyirora@sunyit.edu

Jamol Pender
Cornell University
School of Operations Research and Information Engineering
jjp274@cornell.edu

June 10, 2016

Abstract

This paper considers optimal staffing in service centers. We construct models for profit and cost centers using dynamic rate queues. To allow for practical optimal controls, we approximate the queueing process using a Gaussian random variable with equal mean and variance. We then appeal to the Pontryagin’s maximum principle to derive a closed form square root staffing (SRS) rule for optimal staffing. Unlike most traditional SRS formulas, the main parameter in our formula is not the probability of delay but rather a cost-to-benefit ratio that depends on the shadow price. We show that the delay experienced by customers can be interpreted in terms of this ratio. Throughout the paper, we provide theoretical support of our analysis and conduct extensive numerical experiments to reinforce our findings. To this end, various scenarios are considered to evaluate the change in the staffing levels as the cost-to-benefit ratio changes. We also assess the change in the service grade and the effects of a service-level agreement constraint. Our analysis indicates that the variation in the ratio of customer abandonment over service rate particularly influences staffing levels and can lead to drastically different policies between the cost or profit centers. Our main contribution is the introduction of new analysis and managerial insights into the nonstationary optimal staffing of service centers, especially when the objective is to maximize profitability of the service center.

Keywords Profit Center; Cost Center; Dynamic Queue; Optimal Control
1 Introduction

Managers of service centers are constantly challenged with the problem of optimal staffing. In cost centers, such as most call centers [34], managers are responsible for finding the right number of servers to control expenses [60]. In profit centers, such as clinical departments in the hospital [17, 20], managers are responsible for both the revenues and the expenses, which makes profitability a key performance measure in most profit centers [45, 60].

This paper is motivated by the problem of a profit center manager who is trying to find optimal staffing levels to maximize profitability given some service-level agreement (SLA). It is assumed that the center’s queueing dynamics are characterized by a nonstationary Erlang-A queue ($M(t)/M/s(t) + M$). The staffing solution we propose for this manager follows the square root staffing (SRS) rule [15]. This rule is commonly proposed for staffing service systems such as call centers [19, 25, 29, 59] and healthcare units [27, 61]. A general formula of this rule is $s = q + \beta\sqrt{q}$, where $s$ is the number of servers, $q$ is the offered load (or resource demand), and $\beta$ is the service grade that typically depends on the probability of delay [21, 23].

1.1 Literature Review

A problem similar to that of the profit center manager is considered for call centers [25] and for Emergency Departments (EDs) [54]. In both [25] and [54], variational calculus is employed to find the optimal number of servers via Lagrangian mechanics. Additionally in [25] a fluid version of the modified offered load is proposed for SRS. In [1] also a similar problem is considered, but for stationary Erlang-A queueing systems. The authors propose optimal staffing, under alternate SLAs, to maximize profit in outsourced call centers. The proposed optimal staffing for the horizon-based SLA follows a SRS rule and the optimal number of servers is found by searching for the smallest integer that satisfies the model constraint. Similarly, in [34], a stationary queueing system, $M/M/s/B+M$, is analyzed and an algorithm is introduced to find the optimal number of servers that maximizes profitability. In [35], profit maximization is sought in retail under stationary and deterministic assumptions. The optimal shift scheduling is obtained using mathematical programming techniques. In [26, 30] also shift scheduling is considered for profit maximization.

A variation of the profit center manager problem that is more common in the literature relates to cost minimization [11]. This approach is more suited for cost centers where the managerial goal is to control expenses with possible SLA constraints [34]. For example in [6], optimal staffing in call centers is pursued where the manager’s goal is to minimize both the delay and staffing costs for stationary Erlang-C ($M/M/s$) queueing systems. Through asymptotic approximations, an SRS rule is proposed for optimal staffing where the service grade is a function of the ratio of delay over staffing costs. A similar problem for the stationary Erlang-A queueing systems is discussed in [40]. In [53] appropriate staffing levels for an $M/M/s/B$ queueing system are sought on half-hour intervals to minimize staffing, time in the system, and lost service costs. In [4], a constrained dynamic optimization problem is considered to determine optimal number of permanent versus temporary servers in call centers given a SLA. The objective is to minimize the time-average hiring and opportunity costs. In [18], dynamic programming for a finite horizon is applied to study optimal staffing.
in call centers over multiple time periods where during each period the arrival rate is assumed constant. The manager optimizes staffing at the beginning of each period with the goal of minimizing waiting and staffing costs. An admission control problem is considered in [33] for the stationary Erlang-A queueing system where the Markov Decision Process and the Diffusion Control Problem (DCP) methods are used. The objective is to minimize infinite horizon costs associated with customer abandonment, server idleness, and turning away customers. In [58], DCP is also considered but for a finite horizon where the queueing type is $G/M/n/B + GI$ and the objective is to minimize costs by trading off blocking versus abandonment. The optimal staffing decisions are made in discrete short time intervals where the arrival rate is about constant. In [2], also an admission control problem is considered but for multiclass customers and different server skills. The objective is to minimize the expected sum of blocking costs, waiting costs and defection costs over a finite horizon. A linear program is used to solve a corresponding stochastic fluid approximation.

1.2 Contributions

Our paper extends existing work on optimization of stationary queues (e.g. see [1, 6, 34, 40]) into optimization of nonstationary queues. Our model differs from DCP models that also consider nonstationary queues [2, 58] in that our optimal control applies to the entire finite horizon planning period and the manager does not have to make staffing decisions over discrete time intervals. As a result, our optimal control approach is less computationally intensive since staffing policies are easily expressed in closed forms.

The approach we propose resembles the Lagrangian mechanics considered in [25, 54]. A new feature in our model is a Gaussian refinement for the queueing process to allow for the formulation of a smooth control problem and the derivation of practical staffing solutions. For our purposes, we appeal to the Pontryagin’s maximum principle and are able to derive a closed form SRS formula for optimal staffing. The main parameter in the service grade of our SRS formula is a cost-to-benefit ratio. We show that staffing levels are not dictated by a preset probability of delay target but rather the center’s operating costs and expected benefits such as revenue.

Overall our modeling approach allows for new analysis and managerial insights into nonstationary optimal staffing of service centers, especially when the objective is to maximize profitability. Throughout the paper we provide theoretical support of our analysis and conduct extensive numerical experiments to consider various scenarios to reinforce our findings.

1.3 Organization of the Paper

In Section 2, we present our nonstationary Erlang-A queueing model and also formulate the control problem. In this section the issues with existing fluid approximations are discussed and a Gaussian refinement is introduced. In Section 3, we establish optimal control theorems for both the profit and cost centers. In section 4, we provide managerial insights into our optimal solutions, including the analysis of dynamic solutions, the mean staffing, profitability, and intuitions into the cost-to-benefit ratio. Finally we provide concluding remarks in Section 5. In the Appendix, the proofs to the main theorems and additional supporting materials are provided.
2 Queueing Model and Control Problem

In this section, we describe our queueing model and the control problem faced by the profit center manager. We start by introducing the general model notation used in the paper.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLA</td>
<td>Service-level agreement</td>
</tr>
<tr>
<td>SRS</td>
<td>Square root staffing</td>
</tr>
<tr>
<td>$s(t)$</td>
<td>Number of servers</td>
</tr>
<tr>
<td>$q(t)$</td>
<td>Mean queue length</td>
</tr>
<tr>
<td>$p(t)$</td>
<td>Shadow price</td>
</tr>
<tr>
<td>$\lambda(t)$</td>
<td>Customer arrival rate</td>
</tr>
<tr>
<td>$a$</td>
<td>Abandonment costs</td>
</tr>
<tr>
<td>$b$</td>
<td>Benefits</td>
</tr>
<tr>
<td>$c$</td>
<td>Staffing costs</td>
</tr>
<tr>
<td>$d$</td>
<td>Delay costs</td>
</tr>
<tr>
<td>$T$</td>
<td>Planning period</td>
</tr>
<tr>
<td>$\varrho$</td>
<td>Cost-to-benefit ratio</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Service rate</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Abandonment rate</td>
</tr>
<tr>
<td>$O$</td>
<td>Operating margin</td>
</tr>
<tr>
<td>$x$</td>
<td>Penalty costs</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Maximum allowable probability of abandonment</td>
</tr>
<tr>
<td>$\mathcal{E}$</td>
<td>$\epsilon \cdot \int_0^T \lambda(t)dt$</td>
</tr>
</tbody>
</table>

2.1 Stochastic Queueing Model

A graphical view of a typical service center is portrayed in Figure 1 where it is assumed that the queueing dynamics are characterized by a nonstationary Erlang-A.

Erlang-A, in general, has received considerable attention in the literature since it incorporates many natural features of a realistic service system but also it is tractable. The nonstationary Erlang-A in particular incorporates nonstationary customer arrivals, which is a realistic phenomenon in many service systems. Mathematically, the nonstationary Erlang-A model can be written in terms of time-changed Poisson processes, which makes the analysis more manageable. From a sample path perspective, it is shown in [39] that the queueing system process $Q \equiv \{Q(t)|t \geq 0\}$ is represented by the following stochastic, time changed integral equation:

$$Q(t) = Q(0) + \Pi_1 \left( \int_0^t \lambda(u)du \right) - \Pi_2 \left( \int_0^t \mu \cdot (Q(u) \wedge s(u))du \right) - \Pi_3 \left( \int_0^t \theta \cdot (Q(u) - s(u))^+ du \right),$$

where the $\wedge$ symbol indicates minimum and parameters $\mu$ and $\theta$ are service and abandonment rates, respectively. The $\Pi_i \equiv \{\Pi_i(t)|t \geq 0\}$ for $i = 1, 2, 3$ are i.i.d. standard (rate 1) Poisson
processes. The deterministic time change for $\Pi_1$ transforms it into a non-homogeneous Poisson arrival process with rate $\lambda(t)$ that counts the customer arrivals. Subjecting $\Pi_2$ to a random time change rate $\mu \cdot (Q(t) \land s(t))$, at time $t$, gives us a departure process that counts the number of customers that complete service. Here we assume that there are a deterministic number of $s(t)$ servers, at time $t$, and i.i.d. exponentially distributed service times of mean $1/\mu$. Lastly, the random time change of $\Pi_3$ gives us a counting process for the number of abandonments from $s(t)$ servers and i.i.d. exponentially distributed abandonment times of mean $1/\theta$. When the mean number in the system $E[Q(t)]$ is less than the number of servers $s(t)$ or $E[Q(t)] < s(t)$, we say that the system is underloaded. Conversely, when $E[Q(t)] > s(t)$, we say that the system is overloaded. Finally, when $E[Q(t)] = s(t)$, we say that the system is critically loaded. These assumptions also lead us to a Markovian queueing process that obeys the following Kolmogorov forward equation for the mean queue length.

$$\dot{E}[Q(t)|Q(0) = 0] = \lambda(t) - \mu \cdot E[Q(t) \land s(t)] - \theta \cdot E[(Q(t) - s(t))^+]$$  \hspace{1cm} (2.1)

We use $E$ to symbolize expectation and $\dot{}$ to indicate the time derivative. The functional forward equations for any continuous and bounded function can be computed using the method described in [14] and are given below as

$$\dot{E}[f(Q(t))|Q(0) = 0] = \lambda(t) \cdot E[f(Q(t) + 1) - f(Q(t))] + \mu \cdot E[Q(t) \land s(t) \cdot (f(Q(t) - 1) - f(Q(t)))] + \theta \cdot E[(Q(t) - s(t))^+ \cdot (f(Q(t) - 1) - f(Q(t)))] .$$

We should point out that, as observed in [5, 7], both the service time and time-to-abandon distributions tend to be non-exponential in service systems. But the exponential assumption is necessary for mathematical tractability of dynamic rate queues [22, 24, 25, 42, 43]. In practice, as long as the squared coefficient of variation is not far from 1, the exponential assumption tends to work well [22]. Otherwise for general nonstationary distributions, simulation based algorithms may have to be used (e.g. see [10, 16]) or dynamic programs that consider optimal staffing over discrete time periods (e.g. see [2, 58]).
2.2 Optimal Control Problems

From the forward equations we have that \( \{Q(t) | t \geq 0\} \) represents the total number of customers in the system (in the queue or service) at time \( t \). The term \( E[Q(t) \wedge s(t)] \equiv E[\min(Q(t), s(t))] \) is used to indicate the mean number of customers in service while term \( E[(Q(t) - s(t))^+] \equiv E[\max(0, Q(t) - s(t))] \) is used to represent the mean number of customers in the queue. For a profit center, a fundamental managerial question is how to find the optimal number of servers \( s(t) \) to maximize profitability. The corresponding objective \( \Pi(s(t)) \) may be formulated as follows:

\[
\Pi_1 = \max_{\{s(t) \geq 0 : 0 \leq t \leq T\}} \int_0^T \left[ r \cdot \mu \cdot E[Q(t) - s(t)] - c \cdot s(t) \right] dt
\]  

(2.2)

Here \( r > 0 \) and \( c \geq 0 \) are revenue and staffing costs, respectively. Equation 2.2 indicates that an optimal number of servers \( s(t) \) must be found to maximize the operating net income obtained from the difference between the operating revenue, \( r \cdot \mu \cdot E[Q(t) \wedge s(t)] \), and staffing costs, \( c \cdot s(t) \). It is natural for the manager to also aim at minimizing waiting times, which ultimately translates into fewer customers abandoning. Such quality control may be formulated as follows:

\[
\int_0^T \theta \cdot E[(Q(t) - s(t))^+] dt \leq \mathcal{E}
\]  

(2.3)

where

\[
\mathcal{E} \equiv \epsilon \int_0^T \lambda(t) dt
\]  

(2.4)

and \( \epsilon \) is the maximum allowable probability of abandonment. This quality control is basically an isoperimetric SLA constraint [24] that specifies that during the planning period \([0, T]\), the number of customers that abandon, \( \int_0^T \theta \cdot E[(Q(t) - s(t))^+] dt \), must be less or equal to the maximum allowable fraction of arrivals \( \mathcal{E} \). A complete optimal control problem is presented next.

Problem 2.1 (Profit Model).

\[
\Pi_2 = \max_{\{s(t) \geq 0 : 0 \leq t \leq T\}} \int_0^T \left[ r \cdot \mu \cdot E[Q(t) \wedge s(t)] - c \cdot s(t) \right] dt
\]

subject to

\[
\dot{E}[Q(t)] = \lambda(t) - \mu \cdot E[Q(t) \wedge s(t)] - \theta \cdot E[(Q(t) - s(t))^+]
\]

\[
\int_0^T \theta \cdot E[(Q(t) - s(t))^+] dt \leq \mathcal{E}
\]

For cost centers, where the managerial goal is to minimize costs, a corresponding optimal control problem may be formulated as
Problem 2.2 (Cost Model).

\[ \Pi(s(t)) = \max_{\{s(t) \geq 0 : 0 \leq t \leq T\}} \left\{ \int_0^T (a \cdot \theta + d) \cdot E[(Q(t) - s(t))^+] + c \cdot s(t) dt \right\} \]

subject to

\[ \dot{E}[Q(t)] = \lambda(t) - \mu \cdot E[Q(t) \wedge s(t)] - \theta \cdot E[(Q(t) - s(t))^+] \]

where \( a \geq 0 \) and \( d \geq 0 \) are abandonment and delay costs, respectively. SLA constraints are not necessary in Problem 2.2 since penalty costs for delay and abandonment are assumed incorporated into \( a \) and \( d \).

Both Problems 2.1 and 2.2 are not glaringly difficult, but they pose non-trivial mathematical challenges. The main issue is that the forward equations are neither a closed system nor autonomous since the min and max functions are not explicit functions of the queueing process [47]. The same issue also applies to the objective and constraint functions. The other issue, discussed in the sequel, relates to the challenges of approximating the queueing process \( Q(t) \).

2.3 Approximating the Queueing Process

2.3.1 The Fluid Approach

The distribution of the queueing process \( Q(t) \) is essentially unknown and intractable. One method of simplification commonly used to characterize the behaviors of \( Q(t) \) is the fluid limits based on [39] where the arrival rate and the number of servers are scaled by parameter \( \eta > 0 \) such that

\[ \lim_{\eta \to \infty} \sup_{0 \leq t \leq T} \frac{1}{\eta} Q^\eta(t) = q(t) \quad a.s \ u.o.c. \] (2.5)

where \( q(t) \) solves the following ordinary differential equation

\[ \dot{q}(t) = \lambda(t) - \mu \cdot (q(t) \wedge s(t)) - \theta \cdot (q(t) - s(t))^+. \]

This limit theorem, which approximates a stochastic queue length process with a deterministic dynamical system, allowed [25] to use variational calculus to find optimal staffing levels in call centers and [54] to find optimal staffing for EDs. One issue with fluid approximations is that the Langrangian function is not differentiable everywhere since it still contains the min and max functions. Thus, to find optimal staffing levels, special methods are imperative such as the Competing Lagrangians approach used in [25, 54]. Additionally the Hamiltonian function \( \mathcal{H} \) is a piecewise concave function, which leads to boundary solutions also known as bang-bang [8, 9, 55]. For example in [25], the optimal staffing policy, \( s^*(t) \), indicates that the manager should either staff no one or staff the system with as many customers as currently in the system, \( q(t) \). Implementing such solutions may be challenging for the manager. In fact it seems somewhat impractical for the center to open and shut their doors when it is optimal to do so since customer satisfaction is likely to suffer but also switching costs could
be prohibitive. In the case of a healthcare center such as ED [54], closing the doors during the hours of operations could in addition put patients’ lives at risk.

We will show in the sequel that by adding a stochastic refinement to the fluid approximation of the queue one obtains more practical optimal staffing solutions.

2.3.2 A Gaussian Refinement

In our pursuit of a refinement for the queue length process we choose to use the infinite server queue as our motivation. The infinite server queue is 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 \( M/G/\infty \) queue is tractable, even when the arrival process is nonstationary. In the nonstationary \( M_t/G/\infty \) queue, we know from [12, 13] that the queue length process has a Poisson 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 [12], the mean of the queue length process \( q^\infty(t) \) has the following representation

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

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

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

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 \( M_t/G/\infty \) queue are equal to one another when initialized with a Poisson distribution or at zero. This cumulant moment property of the \( M_t/G/\infty \) queue motivates our approximation of the queue length \( Q(t) \) using a Gaussian random variable with equal mean and variance such that

\[
Q(t) \approx q(t) + X \cdot \sqrt{q(t)}
\]

Here \( X \) is a standard Gaussian random variable with mean 0 and variance 1.

One important property that will be useful for calculating the optimal control solution is the following derivative property of min and max functions. From now on we disregard the time dependence \( t \) to simplify notation.

**Lemma 2.3.** Let \( Q \) be any random variable and \( s(t) \) be a deterministic function of time, then we have that

\[
\frac{\partial}{\partial s} E[Q \wedge s] = -\frac{\partial}{\partial s} E[(Q - s)^+]
\]
Proof.

\[
\frac{\partial}{\partial s} E[Q \wedge s] = \frac{\partial}{\partial s} \left( E[Q] - E[Q - s]^+ \right) = -\frac{\partial}{\partial s} E[(Q - s)^+] \]

To compute the expectations in Lemma 2.3, we exploit Stein’s Lemma [56] for Gaussian random variables.

Stein’s Lemma states the following:

**Lemma 2.4** *(Stein’s lemma [56])*.

\(X\) is a standard Gaussian random variable mean 0 and variance 1 if and only if

\[ E[X \cdot f(X)] = E[f'(X)] \]

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

Using Lemma 2.4 we obtain the expectation of the min and max functions as

\[
E[(Q - s)^+] = \phi(\chi) - \chi \cdot \Phi(\chi) \\
E[Q \wedge s] = q - \phi(\chi) + \chi \cdot \Phi(\chi)
\]

where

\[
\phi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) \equiv \int_{-\infty}^{x} \phi(y)dy, \quad \Phi(x) \equiv 1 - \Phi(x) = \int_{x}^{\infty} \phi(y)dy
\]

and

\[
\chi \equiv \frac{s - q}{\sqrt{q}}.
\]

### 3 Optimal Control

Using the results in Equations 2.7 and 2.8 and the Pontryagin’s maximum principle [52], we next present two fundamental theorems of optimal control in service centers.

**Theorem 3.1** *(Optimal control in a profit center).* The optimal control \(s^*\) under the managerial goals of maximizing profitability is given by

\[
s^* = q^* + \Phi^{-1}(1 - \rho) \cdot \sqrt{q^*}
\]

where

\[
\rho = \frac{c}{\mu \cdot (r - p^*) + \theta \cdot (p^* + x)}
\]

and the optimal queue dynamics \(q^*\) and the shadow price \(p^*\) conform to

\[
\dot{q}^* = \lambda - \mu \cdot q^* - (\mu - \theta) \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)) \cdot \sqrt{q^*}
\]

\[
\dot{p}^* = (\mu \cdot (r - p^*) + \theta \cdot (p^* + x)) \cdot \left( \Phi(\chi) + \frac{\phi(\chi)}{2\sqrt{q^*}} \right) - \mu \cdot (r - p^*)
\]
Proof. see Appendix A.3

**Theorem 3.2 (Optimal control in a cost center).** The optimal control $s^*$ under the managerial goals of minimizing costs is given by

$$ s^* = q^* + \Phi^{-1}(1 - \varrho) \cdot \sqrt{q^*} $$

where

$$ \varrho = \frac{c}{d + \theta \cdot a + p^* \cdot (\theta - \mu)} $$

and the queue dynamics $q^*$ and the shadow price $p^*$ conform to

\[
\begin{align*}
q^* &= \lambda - \mu \cdot q^* - (\mu - \theta) \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)) \cdot \sqrt{q^*} \\
p^* &= (d + \theta \cdot a + p^* \cdot (\theta - \mu)) \cdot (\Phi(\chi) + \frac{\phi(\chi)}{2\sqrt{q^*}}) + \mu \cdot p^*
\end{align*}
\]

Proof. see Appendix A.4

We next present managerial insights from our optimal control results.

## 4 Managerial Insights

### 4.1 The Dynamics of the Optimal Control $s^*$

The optimal control $s^*$ is the recommended staffing levels to maximize profitability in a profit center (see Theorem 3.1) or to minimize costs in a cost center (see Theorem 3.2). We now let $b$ represent the denominator of $\varrho$ in both Theorems 3.1 and 3.2 such that

$$ \varrho = \frac{c}{b} \quad (4.9) $$

Operationally speaking, $b$ can be interpreted as benefits and $c$, as before, represents staffing costs, which then makes $\varrho$ a cost-to-benefit ratio. Naturally, $b > c$ for a rational business. For a profit center, $b = \mu \cdot (r - p^*) + \theta \cdot (p^* + x)$, as indicated in Theorem 3.1. In optimal control theory, $p^*$ is generally viewed as the shadow price of one additional service unit or simply the opportunity cost of one additional server (e.g. see [24], [55]). This means that $\mu \cdot (r - p^*)$ is the revenue rate from serviced customers minus the marginal cost of the rendered service and $\theta \cdot (p^* + x)$ is the penalty cost rate of abandoned customers plus the marginal cost of the forgone service. A special case of equal abandonment and service rates, $\theta = \mu$, eliminates the shadow price $p^*$ from staffing policy decisions, which leads us to the following theorem:

**Theorem 4.1 (Exact Optimal Staffing Policy when $\theta = \mu$).** The optimal control policy when $\theta = \mu$ is given by

$$ s^* = \Gamma^{-1}(q^*, 1 - \varrho) $$

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

$$ \varrho = \frac{c}{\mu \cdot (r + x)} $$
Proof. The solution when $\theta = \mu$ is exact since the queue length distribution is known to be Poisson in this case. We show how to derive the optimal staffing function in the Appendix A.3 using the Chen-Stein Lemma of [14].

Figure 2 portrays the dynamics of optimal solutions for both the cost and profit centers. In the special case of $\theta = \mu$, both the profit and the cost models yield similar staffing policies. For the cases of $\theta \neq \mu$, $p^*$ influences the staffing solutions. Consequently, depending on the values of $p^*$, some solutions are infeasible if $p^*$ makes $b \leq c$.

Sub-figures 2a, 2c, and 2e correspond to the cost model where we hold abandonment costs $a$ to zero and vary delay costs $d$. Sub-figures 2b, 2d, and 2f correspond to the profit model when the maximum allowable probability of abandonment $\epsilon = 1$ meaning that the SLA constraint is not binding. The ratios of staffing over delay costs $(c/d)$ and staffing costs over revenue $(c/r)$ are presented to give an idea of how staffing solutions are changing with respect to the change in these ratios. Clearly, the smaller these ratios get, the most staffing is required.

To finalize this sub-section, we point out the improvement of our solutions over the fluid approximations used in [25, 54]. For example using the competing Lagrangian method in the case of $\mu = \theta$, the maximum staffing level would have been the offered-load, $q^* = \lambda - \mu q$, meaning that the changes in $c/r$ would’ve had no effect on the maximum staffing levels. Nonetheless it makes operational sense to increase staffing to maximize profitability as $c/r$ decreases (see our sub-figure 2d).
The dynamic solutions in Figure 2 don’t allow for a clear appreciation of the marginal change in the staffing levels as ρ becomes very small. In an attempt to assess such change, we analyze

4.2 Mean Staffing Analysis

The dynamic solutions in Figure 2 don’t allow for a clear appreciation of the marginal change in the staffing levels as ρ becomes very small. In an attempt to assess such change, we analyze
the evolution of the mean staffing. In order to relate our mean solutions to the stationary results in [6, 40], we plot \( \bar{s}^* \equiv \frac{1}{T} \int_0^T s^* \, dt \) against \( \frac{1}{\varrho} \) (see Figure 3). In the cost model, \( \frac{1}{\varrho} \) represents the ratio of delay over staffing costs while for the profit model \( \frac{1}{\varrho} \) represents the ratio of revenue over staffing costs. As observed in [6, 40], Figure 3 confirms that staffing levels converge to particular values as \( \frac{1}{\varrho} \) becomes very small. But there is a clear difference in staffing levels depending on the ratio \( \theta/\mu \) and the type of model being considered (profit versus cost). For example in the profit model, when the abandonment constraint is binding, there is a minimum number of servers required to ensure that the constraint is met (see sub-figure 3b for lower values of \( \frac{1}{\varrho} \)). For the cost model, there is a clear difference in staffing policies depending on whether the abandonment costs \( a \) are set to zero or that the delay costs are set to zero. It makes sense that as \( \theta/\mu \to \infty \) more staffing would be needed for the case of \( d = 0 \) and \( a > 0 \) (see sub-figure 3d). Likewise, it is expected that more staffing would be needed as \( \theta/\mu \to 0 \) and \( d > 0 \) while \( a = 0 \) (see sub-figure 3c).

![Figure 3: Planes of \( \frac{1}{\varrho} \) versus the meaning staffing \( \bar{s}^* \) when \( \lambda(t) = 100 + 20 \cdot \sin(t) \).](image-url)
4.3 Profitability Analysis

Hamiltonian functions in optimal control theory are generally viewed as profit rates (e.g. see [55]). For our purposes, such interpretation only makes sense for profit centers. The Hamiltonian function of interest in this case is shown in Equation 4.10 (see derivation in Appendix A.1).

\[
H(s,p,q,x) = (r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)))) - c \cdot s + p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi))) - x \cdot \theta \cdot (\phi(\chi) - \chi \cdot \Phi(\chi)) \cdot \sqrt{q}.
\]

The first term of Equation 4.10, \(r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)))) - c \cdot s\), represents the operating income given that \(r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi))))\) is the operating revenue and \(c \cdot s\) is the operating cost. The second term, \(p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)))\) represents the shadow income from the marginal customer \(q\). In optimality, the Pontryagin’s maximum principle [52] guarantees that the marginal revenue \(p^* \cdot q^*\) equals the marginal cost \(p^*\) [55]. The third term, \(x \cdot \theta \cdot (\phi(\chi) - \chi \cdot \Phi(\chi)) \cdot \sqrt{q}\) represents the penalty cost for violating the abandonment SLA constraint.

Under optimal conditions, the shadow income and the penalty costs do not materialize. These quantities rather serve as a guide for the manager on how to price the marginal service. Nonetheless the operating income does materialize and from it we are able to measure profitability using the operating margin [20], \(\mathcal{O}\), defined as follows:

\[
\mathcal{O} = \frac{1}{T} \int_{0}^{T} \frac{r \cdot \mu \cdot (q^* + \sqrt{q^*} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)))) - c \cdot s^*}{r \cdot \mu \cdot (q^* + \sqrt{q^*} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi))))} dt, \quad 0 < \mathcal{O} < 1
\]

The close to 1 \(\mathcal{O}\) is, the more profitable the profit center, otherwise the center is less profitable. As it can be appreciated in Equation 4.11, the conditions leading to a lower profit margin include higher staffing costs \(c\) or lower operating revenue \(r\). The relationship between the quality of the service grade \(\Phi^{-1}(1 - \varrho) \equiv \Phi^{-1}\) [21] and the operating margin \(\mathcal{O}\) is captured in Figure 4. The abandonment SLA constraint in sub-figure 4a is non-binding with \(\epsilon = 1\) while the constraint is binding in sub-figure 4b with \(\epsilon = 0.1\). We recall that \(\epsilon\) is the maximum allowable probability of abandonment. It follows that with a constant revenue rate \(r\), as the change in \(\theta/\mu\) makes the \(\mathcal{O}\) curve shift to the right, more staffing is required, thus less profit is expected. Also with \(\epsilon = 0.1\) there is a minimum service grade required implying that the center is expected to be less profitable under SLA constraints (see sub-figure 4b). In the case of a small \(\theta/\mu\) ratio, that is very patient customers, the operating margin is less sensitive to the SLA constraint and profitability is expected to be high since most customers are eventually served and almost all revenue is collected.
4.4 Probability of Delay

We have earlier interpreted \( \varrho \) as the cost-to-benefit ratio. In addition, \( \varrho \) can also be interpreted as the probability of delay. Since delay for service happens when the queue length is larger than the number of servers available, then our new interpretation of \( \varrho \) is justified by the following:

\[
\mathbb{P}(Q \geq s) \approx \mathbb{P}(Q^\infty \geq s) \\
\approx \mathbb{P}(q + \sqrt{q} \cdot X \geq q + \sqrt{q} \cdot \Phi^{-1}(1 - \varrho)) \\
= \mathbb{P}(X \geq \Phi^{-1}(1 - \varrho)) \\
= 1 - \Phi(\Phi^{-1}(1 - \varrho))
\]

\( (4.12) \)

We can now give a performance measure interpretation of the delay experienced by customers in terms of \( \varrho \). It follows that as \( \varrho \) increases, we should expect more delay since servers are expensive in light of either increasing staffing costs or decreasing benefits. Similarly, as \( \varrho \) goes down, we should expect the delay to decrease since servers are relatively cheaper.

In the context of the newsvendor problem, \( \varrho \) can conceivably be interpreted as a "stock-out probability" where \( \mathbb{P}(Q \geq s) \) symbolizes the probability of the demand \( Q \) being greater that the current stock \( s \) (e.g. see [3]).

The relationship between the mean \( \Phi^{-1} \) and the mean \( \varrho \), as graphically portrayed in Figure 5, is such that

\[
\Phi^{-1} \begin{cases} 
< 0 & \text{when } \varrho > 0.5 \\
= 0 & \text{when } \varrho = 0.5 \\
> 0 & \text{when } \varrho < 0.5 
\end{cases}
\]

The operating margin \( \mathcal{O} \) is also shown in Figure 5. As expected the higher \( \varrho \) the lower \( \mathcal{O} \), implying that when the probability of delay is high (e.g. servers are more expensive) the profitability is low.
The graphical results in Figure 6 confirm that as $\rho$ becomes small or simply as the benefits far outweigh the costs, the service grade $\Phi^{-1}$ converges to a particular value depending on the model being considered and the ratio $\theta/\mu$. In [40], it is observed that for the cost model, when $d/c \leq 20$, $\Phi^{-1} < 2$ and when $d/c \leq 500$, $\Phi^{-1} < 3$. We also observe similar results in for the cost model in sub-figure 6a. The results in sub-figure 6b are new in the literature and they suggest that the service grades $\Phi^{-1}$ remain remarkably different under various ratios of $\theta/\mu$, even as $1/\rho$ becomes large. The operational intuition is that the less patience customers are, the more servers are needed to maximize profitability.

Finally, we expect our optimal staffing policies to stabilize the probability of delay as portrayed in Figure 7 where the simulated probability of delay is compared to the mean $\rho$;
only feasible cases are shown. For more precise stabilization, it appears that a refinement factor [29] may be necessary for our SRS rule in some cases so $\varrho$ is exactly met. For more information on the stabilization of the queueing parameters see [16, 41, 43, 37, 48, 49].
Figure 7: Simulated probability of delay versus mean $\rho$ (straight horizontal lines) when $\lambda(t) = 100 + 20 \cdot \sin(t)$. For the profit model, $c = 1$, $r = \{1.01, 1.5, 2, 4, 8, 16, 32, 128\}$, and $\epsilon = 1$. For the cost model, $c = 1$, $d = \{1.01, 1.5, 2, 4, 8, 16, 32, 128\}$, and $a = 0$. 
5 Concluding remarks and further research

We constructed models for profit and cost centers using the concepts of queueing and optimal control theories. We justified the Gaussian approximation of the queueing process and invoked the Pontryagin’s maximum principle to derive a SRS optimal staffing policy.

Unlike in most traditional SRS formulas, the main parameter in our formula was not the probability of delay but rather a cost-to-benefit ratio that depends on the shadow price. We showed that the delay experienced by customers can be interpreted in terms of this ratio. One of the conclusions was that as the cost-to-benefit ratio increased, customers experienced more delay since it was more expensive for the center to increase the number of servers. To this end, we showed a special relationship between profitability and the probability of delay.

Throughout the paper we provided theoretical support of our analysis and conducted extensive numerical experiments to verify the optimality and the stability of our solutions. Also various scenarios were considered to assess the change in the staffing levels as the cost-to-benefit ratio changed. We also assessed the change in the service grade and the effects of SLA constraints. We found that as the cost-to-benefit ratio became smaller both the mean staffing and the service grade converged to particular values. The SLA constraints only seemed to matter when the cost-to-benefit ratio was close to one. In all cases, we pointed out that the variation in the ratio of abandonment over service rate particularly affected staffing levels and led to drastically different policies between the cost and profit centers.

For further research, other queueing approximation techniques will be contemplated. Also an extension to more complicated networks of queues will be studied for optimal staffing when customers seek service in different types of centers of the same organization. It seems reasonable that our approach might extend to Jackson networks with abandonment. Another area for further research is to consider optimal staffing of queues with non-Markovian dynamics like in the work of Ko and Pender [32, 31], Pender and Ko [51]. Lastly, it would be interesting to use risk measures like in the work of Pender [50] and generate optimal control policies for nonstationary queues using risk measures in the objective function or constraints.

A Appendix

A.1 Derivation of the Hamiltonian Function for the Profit Model

We follow the methods of optimal control theory (e.g. [8, 9, 36, 55]) and proceed to construct a Hamiltonian function of our profit model using the approximation in Equation 2.6.

\[
\mathcal{H}(s, p, q, x) = r \cdot \mu \cdot E[(Q \land s)] - c \cdot s + p \cdot (\lambda - \mu \cdot E[(Q \land s)] - \theta \cdot E[(Q - s)^+])
\]

\[
- x \cdot \theta \cdot E[(Q - s)^+]
\]

\[
= r \cdot \mu \cdot E[((q + \sqrt{q} \cdot X) \land s)] - c \cdot s
\]

\[
+ p \cdot (\lambda - \mu \cdot E[((q + \sqrt{q} \cdot X) \land s)] - \theta \cdot E[((q + \sqrt{q} \cdot X) - s)^+])
\]

\[
- x \cdot \theta \cdot E[((q + \sqrt{q} \cdot X) - s)^+].
\]

(1.13)
The $p$ in the Hamiltonian function $\mathcal{H}$ is the shadow price and $x$ is the multiplier (interpreted as penalty cost) of some auxiliary variable $Z$ given by

$$Z = -\int_0^t \theta \cdot E \left[ \left( (q(u) + X \cdot \sqrt{q(u)}) - s \right)^+ \right] \, du$$

$$\dot{Z} = -\theta \cdot E[[q + X \cdot \sqrt{q}) - s]^+]$$

where $Z(T) \geq -\mathcal{E}$. Again $\mathcal{E} = \epsilon \cdot \int_0^T \lambda \, dt$ with $\epsilon$ being the maximum allowable probability of abandonment. Since $Z$ does not appear in Equation 1.13, then $x = -\partial \mathcal{H} / \partial Z = 0$, meaning that $x$ is a constant that satisfies the following complementary of slackness equation:

$$x \cdot \left[ \mathcal{E} - \int_0^T \theta \cdot E[[q + X \cdot \sqrt{q}) - s]^+] \, dt \right] = 0 \tag{1.14}$$

Accordingly, $x = 0$ when $\mathcal{E} - \int_0^T \theta \cdot E[[q + X \cdot \sqrt{q}) - s]^+] \, dt > 0$, else $x > 0$. I follows that $\mathcal{H}(s, p, q, x) \approx r \cdot \mu \cdot E[(q + X \cdot \sqrt{q}) \wedge s] - c \cdot s$

$$+ p \cdot (\lambda - \mu) \cdot E[(q + X \cdot \sqrt{q}) \wedge s]$$

$$- \theta \cdot E[((q + X \cdot \sqrt{q}) - s)^+] - x \cdot \theta \cdot E[((q + X \cdot \sqrt{q}) - s)^+]$$

$$= r \cdot \mu \cdot (q + (X \wedge \chi) \cdot \sqrt{q}) - c \cdot s$$

$$+ p \cdot (\lambda - \mu) \cdot (q + (X \wedge \chi) \cdot \sqrt{q}) - \theta \cdot (X - \chi)^+ \cdot \sqrt{q} - x \cdot \theta \cdot (X - \chi)^+ \sqrt{q}$$

$$= r \cdot \mu \cdot (q + E[(X \wedge \chi)] \cdot \sqrt{q}) - c \cdot s$$

$$+ p \cdot (\lambda - \mu) \cdot (q + E[(X \wedge \chi)] \cdot \sqrt{q}) - \theta \cdot E[(X - \chi)^+] \cdot \sqrt{q}$$

$$- x \cdot \theta \cdot E[(X - \chi)^+] \cdot \sqrt{q}$$

where

$$\chi \equiv \frac{s - q}{\sqrt{q}}.$$  

By appealing to Stein’s Lemma 2.4 we are able to compute

$$E[(X - \chi)^+] = E[(X - \chi) \cdot \{X \geq \chi\}]$$

$$= E[X \cdot \{X \geq \chi\}] - \chi \cdot P\{X \geq \chi\}$$

$$= \int_{-\infty}^{\infty} \delta_{\chi}(y) \cdot \phi(y) \, dy - \chi \cdot \Phi(\chi)$$

$$= \phi(\chi) - \chi \cdot \Phi(\chi).$$

Given
\[ E[X \wedge \chi] = E[X] - E[(X - \chi)^+] , \]
we easily compute
\[ E[X \wedge \chi] = E[X - (X - \chi)^+] \]
\[ = 0 - E[(X - \chi)^+] \]
\[ = \chi \cdot \Phi(\chi) - \phi(\chi). \]

Finally we obtain
\[ \mathcal{H}(s, p, q, x) = (r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi))) - c \cdot s) \]
\[ + p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi))) \]
\[ - x \cdot \theta \cdot (\phi(\chi) - \chi \cdot \Phi(\chi)) \cdot \sqrt{q} \] (1.15)

### A.2 Necessary Conditions

In order to prove that our staffing solutions are 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 of \( \mathcal{H} \) in Equation 1.15 with respect to the queue length \( q \) and to the shadow price \( p \). We proceed as follows:
\[
\frac{\partial}{\partial q} (\chi \sqrt{q} \cdot \Phi) = \frac{\partial}{\partial q} (\sqrt{q} \cdot \phi) \]
\[ = \chi \cdot \phi \cdot \left( \frac{s + q}{2q} \right) - \Phi - \chi \cdot \phi \cdot \left( \frac{s + q}{2q} \right) - \phi \cdot \frac{1}{2\sqrt{q}} \]
\[ = -\left( \Phi + \frac{\phi}{2\sqrt{q}} \right), \]

\[
\frac{\partial}{\partial q} (\chi \sqrt{q} \cdot \Phi) = \frac{\partial}{\partial q} ((s - q) \cdot \Phi) \]
\[ = (s - q) \cdot \phi \cdot \left( \frac{s + q}{2q \cdot \sqrt{q}} \right) - \Phi \]
\[ = \chi \cdot \phi \cdot \left( \frac{s + q}{2q} \right) - \Phi, \]

and
\[
\frac{\partial}{\partial q} (\sqrt{q} \cdot \phi) = \sqrt{q} \cdot \chi \cdot \phi \cdot \left( \frac{s + q}{2q \sqrt{q}} \right) + \frac{1}{2\sqrt{q}} \cdot \phi \]
\[ = \chi \cdot \phi \cdot \left( \frac{s + q}{2q} \right) + \frac{\phi}{2\sqrt{q}}. \]
The necessary conditions are then given by:

\[
\frac{\partial \mathcal{H}}{\partial p} \equiv \dot{q} = \lambda - \mu \cdot q - (\mu - \theta) \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)) \cdot \sqrt{q}
\]

\[
- \frac{\partial \mathcal{H}}{\partial q} \equiv \dot{p} = -r \cdot \mu + r \cdot \mu \cdot \left( \frac{\partial}{\partial q} (\chi \cdot \sqrt{q} \cdot \Phi) - \frac{\partial}{\partial q} (\sqrt{q} \cdot \phi) \right)
\]

\[
- p \left( -\mu + (\mu - \theta) \cdot \left( \frac{\partial}{\partial q} (\chi \cdot \sqrt{q} \cdot \Phi) - \frac{\partial}{\partial q} (\sqrt{q} \cdot \phi) \right) \right)
\]

\[
+ x \cdot \theta \cdot \left( \frac{\partial}{\partial q} (\chi \cdot \sqrt{q} \cdot \Phi) - \frac{\partial}{\partial q} (\sqrt{q} \cdot \phi) \right)
\]

\[
= (\mu \cdot (r - p) + \theta \cdot (p + x)) \cdot \left( \Phi + \frac{\phi}{2\sqrt{q}} \right) - \mu \cdot (r - p)
\]

**A.3 Proof of Theorem 3.1**

From the Pontryagin’s maximum principle, the optimal control policy \( s^* \) that maximizes the Hamiltonian function in Equation 1.15, such that \( \mathcal{H}(s^*, p^*, q^*, x^*, t) \geq \mathcal{H}(s, p, q, x, t) \), is obtained by \( \frac{\partial \mathcal{H}}{\partial s} = 0 \).

Given

\[
\frac{\partial}{\partial s} (\chi \sqrt{q} \cdot \Phi) = \frac{\partial}{\partial s} (s - q) \cdot \Phi
\]

\[
= (s - q) \cdot \frac{-\phi}{\sqrt{q}} + \Phi
\]

\[
= \Phi - \chi \cdot \phi
\]

and

\[
\frac{\partial}{\partial s} (\sqrt{q} \cdot \phi) = -\frac{\chi \cdot \phi(\chi)}{\sqrt{q}} \cdot \sqrt{q} = -\chi \cdot \phi
\]

We obtain

\[
\frac{\partial \mathcal{H}}{\partial s} = r \cdot \mu \cdot \Phi(\chi) - c - p \cdot (\mu - \theta) \cdot \Phi(\chi) + x \cdot \theta \cdot \Phi(\chi) = 0
\]

\[
= (\mu \cdot (r - p) + \theta \cdot (p + x)) \cdot \Phi(\chi) - c = 0
\]
We now solve for $s$ by recalling that $\chi = s - q/\sqrt{q}$. We proceed as follows:

$$
\begin{align*}
\mu \cdot (r - p) + \theta \cdot (p + x) \cdot \Phi \left( \frac{s - q}{\sqrt{q}} \right) &= c \\
\Phi \left( \frac{s - q}{\sqrt{q}} \right) &= \frac{c}{\mu (r - p) + \theta (p + x)} \\
\Phi \left( \frac{s - q}{\sqrt{q}} \right) &= 1 - \frac{c}{\mu (r - p) + \theta (p + x)} \\
\frac{s - q}{\sqrt{q}} &= \Phi^{-1} \left( 1 - \frac{c}{\mu (r - p) + \theta (p + x)} \right)
\end{align*}
$$

Finally we obtain the optimal staffing $s^*$ given by

$$
s^* = q + \Phi^{-1} \left( 1 - \frac{c}{\mu (r - p) + \theta (p + x)} \right) \cdot \sqrt{q}.
$$

### A.4 Proof of Theorem 3.2

The Hamiltonian function associated with the cost model is given by

$$
\begin{align*}
\mathcal{H}(s, p, q, x) &= -c \cdot s - (d + \theta \cdot a) \cdot \sqrt{q} \cdot \left( \phi(\chi) - \chi \cdot \Phi(\chi) \right) \\
&\quad + p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)) \\
&= -s \cdot \mu - (\theta + a) \cdot \sqrt{q} \cdot \left( \phi(\chi) - \chi \cdot \Phi(\chi) \right)
\end{align*}
$$

The resulting necessary conditions follow:

$$
\begin{align*}
\frac{\partial \mathcal{H}}{\partial p} &\equiv q = \lambda - \mu \cdot q - (\mu - \theta) \cdot (\chi \cdot \Phi(\chi) - \phi(\chi)) \cdot \sqrt{q} \\
-\frac{\partial \mathcal{H}}{\partial q} &\equiv p = (d + \theta \cdot a + p \cdot (\theta - \mu)) \cdot \left( \Phi + \frac{\phi}{2\sqrt{q}} \right) + \mu \cdot p
\end{align*}
$$

The optimal staffing policy $s^*$ is obtained by

$$
\frac{\partial \mathcal{H}}{\partial s} = 0 \implies s^* = q + \Phi^{-1} \left( 1 - \frac{c}{d + \theta \cdot a + p \cdot (\theta - \mu)} \right) \cdot \sqrt{q}
$$

### A.5 Proof of Theorem 4.1

**Theorem A.1** (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 $f : \mathbb{N} \to \mathbb{N},$

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

**Proof.** See [46].
Lemma A.2.

\[
\Gamma(s,x) = \sum_{m=s}^{\infty} e^{-x} \frac{x^m}{m!} = \frac{1}{\Gamma(s)} \int_{0}^{x} e^{-y^{s-1}} dy
\]

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

where

\[
\Gamma(s,x) = \frac{1}{\Gamma(s)} \int_{0}^{x} e^{-y^{s-1}} dy \quad \text{and} \quad \overline{\Gamma}(s,x) = \frac{1}{\Gamma(s)} \int_{x}^{\infty} e^{-y^{s-1}} dy
\]

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 [28].

Lemma A.3.

\[
E[(Q-s)^+] = E[(Q-s) \cdot \{ Q > s \}]
\]
\[
= E[Q \cdot \{ Q > s \}] - s \cdot E[\{ Q > s \}]
\]
\[
= E[Q \cdot \{ Q > s \}] - s \cdot \Gamma(s+1,q)
\]
\[
= q \cdot E[\{ Q + 1 > s \}] - s \cdot \Gamma(s+1,q)
\]
\[
= q \cdot \Gamma(s,q) - s \cdot \Gamma(s+1,q)
\]

\[
\frac{\partial}{\partial s} E[(Q-s)^+] = -\Gamma(s+1,q)
\]
\[
\frac{\partial}{\partial s} E[Q \wedge s] = \Gamma(s+1,q)
\]

Now we prove Theorem 4.1 using the results of the profit model. In the case where \( \theta = \mu \), we have that

\[
\frac{\partial \mathcal{H}}{\partial s} = r \cdot \mu \cdot \Gamma(s+1,q) - c - p \cdot (\mu - \theta) \cdot \Gamma(s+1,q) + x \cdot \theta \cdot \Gamma(s+1,q) = 0
\]
\[
= \mu \cdot (r + x) \cdot \Gamma(s+1,q) - c = 0.
\]

We now solve for \( s \) given that \( \chi = \frac{s-q}{\sqrt{q}} \):

\[
\mu \cdot (r + x) \cdot \Gamma(s+1,q) - c = 0
\]
\[
\Gamma(s+1,q) = \frac{c}{\mu \cdot (r + x)}.
\]

Finally we obtain optimal control policy \( s^* \) as:

\[
s^* = \Gamma^{-1}\left(q, \frac{c}{\mu \cdot (r + x)}\right).
\]
B Numerical Integration Algorithm

Algorithm B.1 is based on the Forward-Backward method detailed in [36]. We added Step 4 to accommodate the computation of the complementary of slackness in Equation 1.14.

Algorithm B.1 (numerical solutions for $s(t)$, $q(t)$, and $p(t)$).

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

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

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

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

$$s_n(t) = q_n(t) + \Phi^{-1}(1 - \omega_n(t)) \cdot \sqrt{q_n}$$

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

1. $n = n + 1$
2. $x_{n+1} = x_n + h$

where $h$ is a very small number.

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

$$\int_0^T \theta \cdot (q_n(t) - s_n(t))^+ < \mathcal{E} \quad \text{and} \quad \|\bar{s}_n\|_{n-1}^n - \|\bar{s}_n\|_n < \delta$$

where $\delta$ is the accepted convergence tolerance.

For further discussion on the convergence of forward-backward algorithms see [38, 44, 57].

References


