Optimal Staffing in Nonstationary Service Centers with Constraints

Jerome Niyirora,[1] Jamol Pender[2]

[1] *College of Health Sciences and Management, SUNY Polytechnic Institute: Utica, NY*

[2] *School of Operations Research and Information Engineering, Cornell University: Ithaca, NY*

Received 6 July 2015; revised 23 November 2016; accepted 29 November 2016
DOI 10.1002/nav.21723
Published online 26 December 2016 in Wiley Online Library (wileyonlinelibrary.com).

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 article, 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 profit and cost service 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. © 2016 Wiley Periodicals, Inc. Naval Research Logistics 63: 615–630, 2017

Keywords: service center; dynamic queue; optimal control

1. INTRODUCTION

Managers of service centers are constantly challenged with the problem of optimal staffing. In *service* cost centers, such as most call centers [33], managers are responsible for finding the right number of servers to control expenses [59]. In *service* profit centers, such as clinical departments in the hospitals [16, 19], managers are responsible for both the revenues and the expenses, which makes profitability a key performance measure [44, 59].

This article 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 follows the square root staffing (SRS) rule [14]. 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 β is the service grade that typically depends on the probability of delay [20, 22]. This rule is commonly proposed for staffing service systems such as call centers [18, 24, 28, 58] and healthcare units [26, 60].

1.1. Literature Review

A problem similar to that of ours is considered for call centers [24] and for Emergency Departments (EDs) [53]. In both [24] and [53], variational calculus is employed to find the optimal number of servers via Lagrangian mechanics. Additionally, in [24], a fluid version of the modified offered load is proposed for SRS. In [1], a similar problem is considered, but for stationary Erlang-A queueing systems. The authors devise optimal staffing policies, under alternate SLAs, to maximize profit in outsourced call centers. The proposed policy for the horizon-based SLA systems follows a SRS rule. The optimal number of servers is found by searching for the smallest integer that satisfies the model constraint. Similarly, in [33], 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 [34], profit maximization is pursued under stationary and deterministic

Correspondence to: Jerome Niyirora (jerome.niyirora@sunyit.edu)

assumptions. Based on the optimal level for each period, a schedule of optimal shifts is obtained using mathematical programming techniques. In [25, 29], the same approach is used.

Other papers pursue the cost minimization variation of our problem. In [6], optimal staffing policies for call centers are devised where the manager's goal is to minimize both the delay and the staffing costs for Erlang-C ($M/M/s$) queueing systems. Through asymptotic approximations, an SRS rule is derived for optimal staffing with the service grade being a function of the ratio of delay over staffing costs. A similar problem for stationary Erlang-A queueing systems is discussed in [39]. In [52], 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 the optimal number of permanent and temporary servers in call centers, given the SLA. The objective is to minimize the time-average hiring and opportunity costs. In [17], dynamic programming with 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 levels at the beginning of each period with the goal of minimizing waiting and staffing costs. An admission control problem is considered in [32] for the stationary Erlang-A queueing system and is analyzed as a Markov decision process and as a diffusion control problem (DCP). The objective is to minimize infinite horizon costs associated with customer abandonments, server idleness, and the turning away of customers. In [57], a DCP is also considered but for a finite horizon problem where the queueing type is $G/M/n/B + GI$ and the objective is to minimize costs by trading off blocking versus abandonment costs. The optimal staffing decisions are made in discrete short time intervals where the arrival rate is assumed to be constant. In [2], also an admission control problem is considered but for multiclass customer and for servers with different skills. The objective is to minimize the expected costs of blocking, waiting, and defection. A linear program is used to solve a corresponding stochastic fluid approximation.

1.2. Contributions

Our article extends existing work on optimization of stationary queues (e.g., see [1, 6, 33, 39]) to address nonstationary systems. Our model differs from DCP models that also consider nonstationary queues [2, 57] 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 form.

The approach we propose resembles the Lagrangian mechanics adopted in [24, 53]. 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. The main parameter in the service grade function of our formula is a cost-to-benefit ratio. We show that staffing levels are not dictated by a preset probability of delay target but rather a fraction of the center's operating costs and expected revenues.

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 article, we pair theoretical analyses with extensive numerical experiments to illustrate our findings.

1.3. Organization of the Article

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. We provide concluding remarks in Section 5. Finally, in the Appendix, we present the proofs of the main theorems and a numerical integration algorithm of our model.

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 most frequent model notations used in the article. Particular notations are introduced in respective sections.

Notation	Description
SLA	Service-level agreement
SRS	Square root staffing
$s(t)$	Number of servers
$q(t)$	Mean queue length
$p(t)$	Shadow price
$\lambda(t)$	Customer arrival rate
a	Abandonment costs

Notation	Description
b	Benefits
c	Staffing costs
d	Delay costs
T	Planning period
ϱ	Cost-to-benefit ratio
μ	Service rate
θ	Abandonment rate
$\mathcal{O}$	Operating margin
X	Penalty costs
ϵ	Maximum allowable probability of abandonment

2.1. Stochastic Queueing Model

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

Erlang-A models have received considerable attention in the literature since they incorporate many natural features of a service system but also because they are tractable. A nonstationary Erlang-A model incorporates time-varying customer arrivals, which is a more realistic representation of a typical service system. Mathematically, a 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 [38] that the queueing system process $Q \equiv \{Q(t)|t \geq 0\}$ is represented by the following stochastic, time changed integral equation:

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

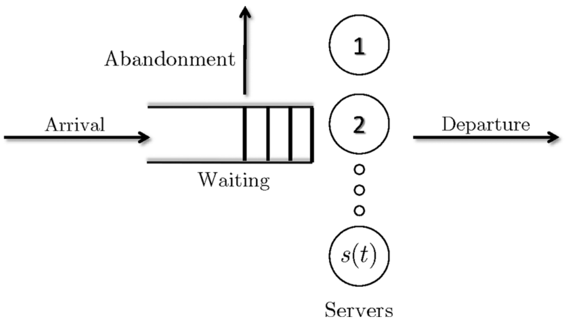


Figure 1. A nonstationary Erlang-A model for a typical service center.

where the $\wedge$ symbol indicates *minimum* and $x^+ = \max(x, 0)$. Moreover, the parameters μ and θ 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 Π_1 transforms it into a non-homogeneous Poisson arrival process with rate $\lambda(t)$ that counts the customer arrivals. Subjecting Π_2 to a random time change rate $\mu \cdot (Q(t) \wedge s(t))$, 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 Π_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*. It turns out that our sample path representation also leads us to a Markovian queueing process that obeys the following Kolmogorov forward equation for the mean queue length.

$$\begin{aligned} \overset{\bullet}{E}\ [Q(t)|Q(0) = 0] = \lambda(t) - \mu \cdot E[Q(t) \wedge s(t)] \\ - \theta \cdot E[(Q(t) - s(t))^+] \qquad (2.1) \end{aligned}$$

We use E to symbolize expectation and • to indicate the time derivative. The functional forward equations for any continuous and bounded function can be computed using the method described in [13] and are given below as

$$\begin{aligned} &\overset{\bullet}{E}\ [f(Q(t))|Q(0) = 0] \\ &\quad = \lambda(t) \cdot E[f(Q(t) + 1) - f(Q(t))] \\ &\qquad + \mu \cdot E[Q(t) \wedge s(t) \cdot (f(Q(t) - 1) - f(Q(t)))] \\ &\qquad + \theta \cdot E[(Q(t) - s(t))^+ \cdot (f(Q(t) - 1) - f(Q(t)))]. \end{aligned}$$

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 [21, 23, 24, 41, 42]. In practice, as long as the squared coefficient of variation is not far from one, the exponential assumption tends to work well [21]. Otherwise for general nonstationary distributions, simulation based algorithms may have to be used (e.g., see [10, 15]) or dynamic programs that consider optimal staffing over discrete time periods (e.g., see [2, 57]).

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 the 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 function $\zeta(s(t))$ may be formulated as follows:

$$\zeta(s(t)) = \max_{\{s(t)\geq 0:\, 0\leq t\leq T\}} \int_0^T [r \cdot \mu \cdot E[(Q(t) \wedge s(t)] - c \cdot s(t)]dt \tag{2.2}$$

Here $r > 0$ and $0 \leq c < r$ 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:

$$\int_0^T \theta \cdot E[(Q(t) - s(t))^+]dt \leq \mathcal{E} \tag{2.3}$$

where

$$\mathcal{E} \equiv \epsilon \int_0^T \lambda(t)dt \tag{2.4}$$

and ϵ is the maximum allowable probability of abandonment. This quality control is basically an isoperimetric SLA constraint [23] 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 abandonments $\mathcal{E}$. A complete optimal control problem is presented next.

PROBLEM1 2.1: (Profit Model):

$$\begin{aligned}
&\zeta_p(s(t)) \\
&= \max_{\{s(t)\geq 0:\, 0\leq t\leq T\}} \int_0^T [r \cdot \mu \cdot E[(Q(t) \wedge s(t)] - c \cdot s(t)]dt \\
&\quad \textit{subject to} \\
&\quad \dot{E}\ [Q(t)] = \lambda(t) - \mu \cdot E[Q(t) \wedge s(t)] \\
&\qquad - \theta \cdot E[(Q(t) - s(t))^+] \\
&\quad \int_0^T \theta \cdot E[(Q(t) - s(t))^+]dt \leq \mathcal{E}
\end{aligned}$$

For cost centers, where the managerial goal is to minimize costs, a corresponding optimal control problem may be formulated as

PROBLEM2 2.2: (Cost Model).

$$\begin{aligned}
\zeta_c(s(t)) = &\max_{\{s(t)\geq 0:\, 0\leq t\leq T\}} - \int_0^T [(a \cdot \theta + d) \cdot E[(Q(t) - s(t))^+] \\
&+ c \cdot s(t)]dt \\
&\textit{subject to} \\
&\dot{E}\ [Q(t)] = \lambda(t) - \mu \cdot E[Q(t) \wedge s(t)] \\
&\quad - \theta \cdot E[(Q(t) - s(t))^+]
\end{aligned}$$

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 $(x \wedge y)$ and the max $((x - y)^+)$ terms are not explicit functions of the queueing process [46]. 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 $Q(t)$ is the fluid limits based on [38]. The basic idea is to scale both the arrival rate and the number of servers by parameter $\eta > 0$ such that

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

Here $q(t)$ 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 [24] to use variational calculus to find optimal staffing levels in call centers and [53] to find optimal staffing for EDs. One issue with fluid approximations of $Q(t)$ 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 [24, 53]. Additionally, the Hamiltonian function $\mathcal{H}$ is a piecewise concave function, which leads to boundary solutions also known as *bang-bang* [8, 9, 54]. For example, in [24], the optimal staffing policy, $s^*(t)$, indicates that the manager should either staff no one or staff the system with the number of servers equivalent to the number of customers currently in the system, $q(t)$. Implementing such solutions may be challenging for the manager. In fact, it seems somewhat unreasonable for the center to open or shut their doors when it is optimal to do so since a few customers may have to be, purposely, subjected to long waits. Moreover, switching costs may be prohibitive.

We will show in the sequel that by adding a stochastic refinement to the fluid approximation of $Q(t)$, one obtains more practical staffing policies for service centers.

2.3.2. A Gaussian Refinement

In our pursuit of a refinement for the queue length process, $Q(t)$, 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/∞ queue is tractable, even when the arrival process is nonstationary. In the nonstationary M_t/G/∞ queue, we know from [11, 12] that the queue length process has a Poisson distribution with time varying rate $q^\infty(t)$ as given by Eq. (2.6). 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 [11], the mean of the queue length process $q^\infty(t)$ has the following representation

$$\begin{aligned} q^\infty(t) &\equiv E[Q^\infty(t)] && (2.6)\\ &= \int_{-\infty}^{t} \bar{G}(t-u)\lambda(u)du && (2.7)\\ &= E\left[\int_{t-S}^{t} \lambda(u)du\right] && (2.8)\\ &= E[\lambda(t-S_e)]\cdot E[S] && (2.9)\end{aligned}$$

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

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

It turns out that 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 approximation of the queue length $Q(t)$ using a Gaussian random variable with equal mean and variance such that

$$Q(t) \approx q(t) + X \cdot \sqrt{q(t)} \qquad (2.10)$$

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 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:

$$\begin{aligned}\frac{\partial}{\partial s}E[Q \wedge s] &= \frac{\partial}{\partial s}(E[Q] - E[(Q-s)^+])\\ &= -\frac{\partial}{\partial s}E[(Q-s)^+] \qquad \square\end{aligned}$$

To compute the expectations in Lemma 2.3, we exploit the Stein's Lemma for Gaussian random variables [55].

Stein's Lemma states the following:

LEMMA 2.4 (Stein's lemma [55]): 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

$$\begin{aligned} E[(Q-s)^+] &= \phi(\chi) - \chi \cdot \overline{\Phi}(\chi) && (2.11)\\ E[Q \wedge s] &= q - \phi(\chi) + \chi \cdot \overline{\Phi}(\chi) && (2.12)\end{aligned}$$

where

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

$$\overline{\Phi}(x) \equiv 1 - \Phi(x) = \int_x^\infty \phi(y)dy$$

and

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

3. OPTIMAL CONTROL

Using the results in Eqs. (2.11) and (2.12) and the Pontryagin's maximum principle [51], we next present two fundamental theorems of our optimal control model.

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

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

where

$$\varrho = \frac{c}{\mu \cdot (r - p^*) + \theta \cdot (p^* + \mathrm{x})}$$

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

$$\overset{\bullet}{q}{}^* = \lambda - \mu \cdot q^* - (\mu - \theta) \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi)) \cdot \sqrt{q^*}$$

$$\overset{\bullet}{p}{}^* = (\mu \cdot (r - p^*) + \theta \cdot (p^* + \mathrm{x})) \cdot \left(\overline{\Phi}(\chi) + \frac{\phi(\chi)}{2\sqrt{q^*}} \right) - \mu \cdot (r - p^*)$$

PROOF: see Proof of Theorem 3.1 in Appendix. □

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

$$\overset{\bullet}{q}{}^* = \lambda - \mu \cdot q^* - (\mu - \theta) \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi)) \cdot \sqrt{q^*}$$

$$\overset{\bullet}{p}{}^* = (d + \theta \cdot a + p^* \cdot (\theta - \mu)) \cdot \left(\overline{\Phi}(\chi) + \frac{\phi(\chi)}{2\sqrt{q^*}} \right) + \mu \cdot p^*$$

PROOF: see Proof of Theorem 3.2 in Appendix. □

The algorithm for obtaining solutions in both Theorem 3.1 and 3.2 is given in Numerical Integration Algorithm in Appendix. We next present managerial insights from our model.

4. MANAGERIAL INSIGHTS

4.1. The Dynamics of the Optimal Control s^*

The optimal control s^* is the recommended staffing level 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 ϱ in both Theorems 3.1 and 3.2 such that

$$\varrho = \frac{c}{b} \tag{4.13}$$

and $b > c$. Operationally speaking, b can be interpreted as *benefits* and c, as before, represents *staffing costs*. Accordingly, we refer to ϱ as the *cost-to-benefit* ratio. For a profit center, $b = \mu \cdot (r - p^*) + \theta \cdot (p^* + \mathrm{x})$, as indicated in Theorem 3.1. From the general theory of optimal control, p^* is viewed as the shadow price of one additional service unit or simply the marginal cost rate of one additional server (e.g., see [23, 54]). This means that $\mu \cdot (r - p^*)$ is the difference of the revenue r, from served customers, and the marginal cost p^*, for the rendered service. Similarly, $\theta \cdot (p^* + \mathrm{x})$ is the sum of the marginal cost p^*, for the forgone service, and the penalty cost x, for abandoned customers. As illustrated in the next theorem, a special case of $\theta = \mu$ eliminates the shadow price p^* from staffing policy decisions.

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

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

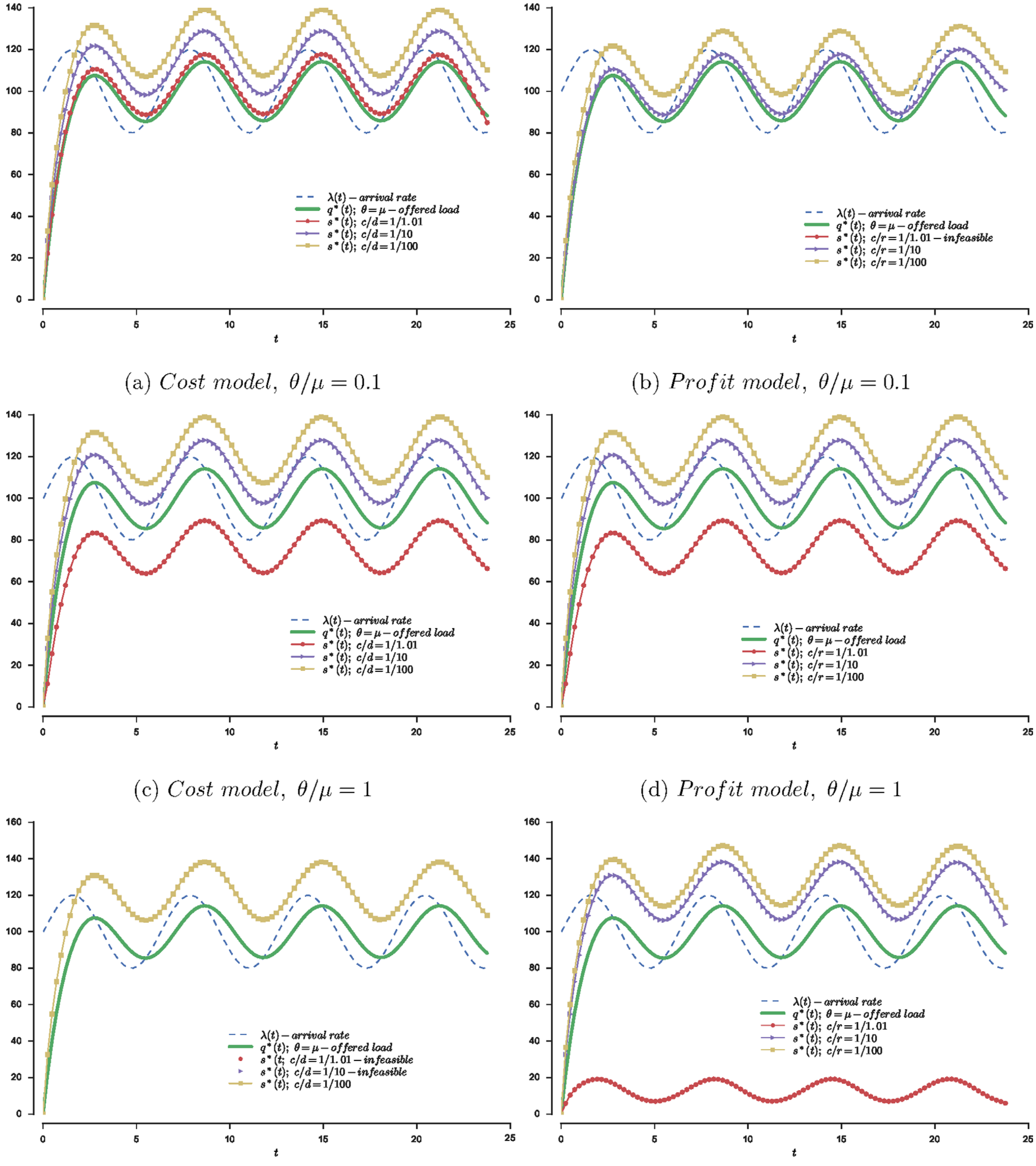
where $\Gamma^{-1}(q^*, 1 - \varrho)$ is the inverse incomplete Gamma function with parameters $(q^*, 1 - \varrho)$. Moreover, we have that

$$\varrho = \frac{c}{\mu \cdot (r + \mathrm{x})}$$

PROOF: The solution, when $\theta = \mu$, is exact since the queue length distribution is known to be Poisson (see Proof of Theorem 4.1 in Appendix). □

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 staffing solutions. In some cases, p^* induces $b \leq c$, which leads to infeasible solutions.

Figures 2a, 2c, and 2e correspond to the cost model with the abandonment costs a held to zero and delay costs d varied. Figures 2b, 2d, and 2f correspond to the profit model. In this case, the maximum allowable probability of abandonment ϵ



(a) $Cost\ model,\ \theta/\mu = 0.1$ (b) $Profit\ model,\ \theta/\mu = 0.1$

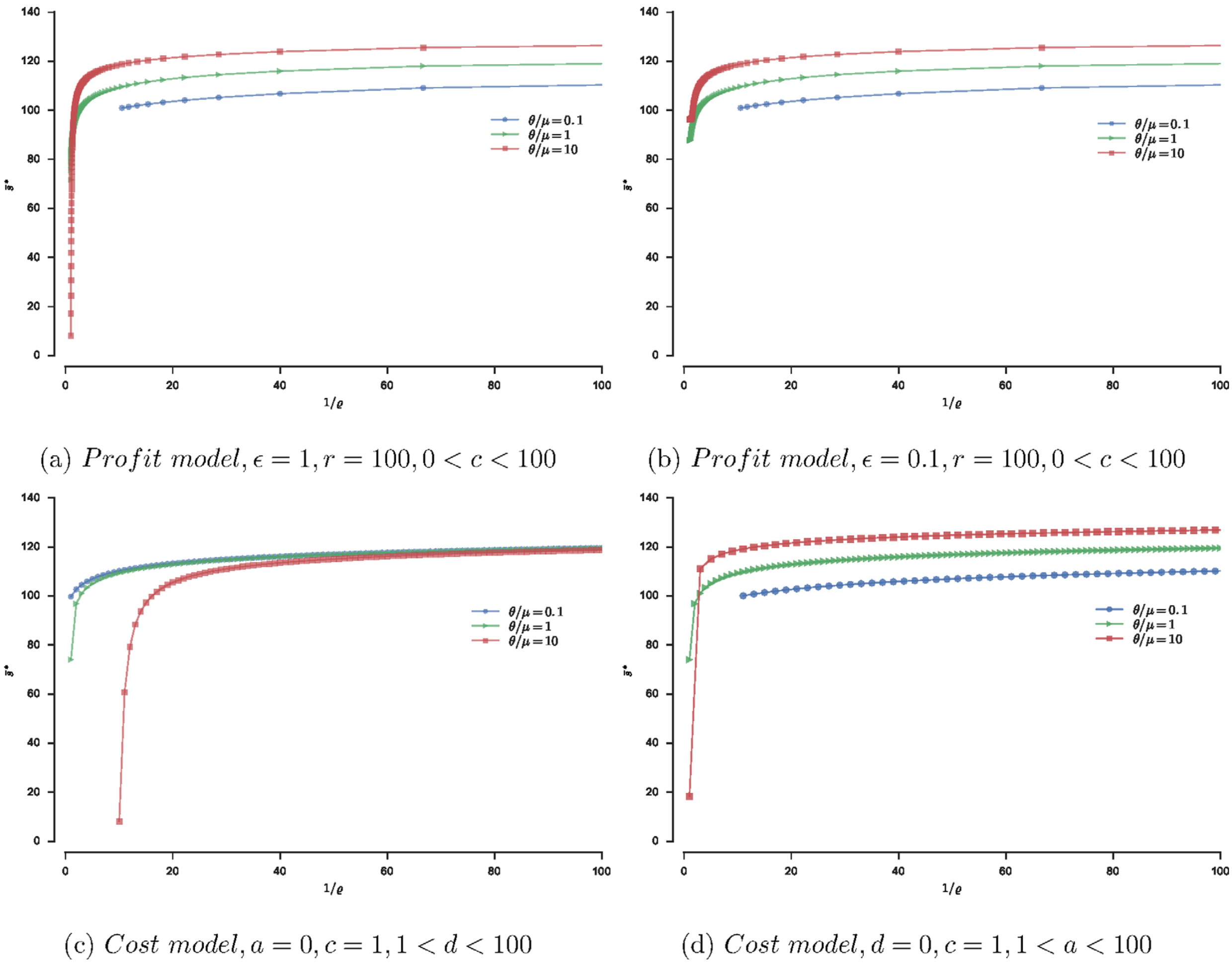
(c) $Cost\ model,\ \theta/\mu = 1$ (d) $Profit\ model,\ \theta/\mu = 1$

(e) $Cost\ model,\ \theta/\mu = 10$ (f) $Profit\ model,\ \theta/\mu = 10$

Figure 2. Optimal staffing when model parameters are varied and $\lambda(t) = 100 + 20 \cdot \sin(t)$. [Color figure can be viewed at wileyonlinelibrary.com.]

is set to 1, which means that the SLA constraint is not binding. Given the ratios c / d and c / r in Fig. 2, it can be concluded that more servers are affordable as the cost-to-benefit ratio tends to zero.

To finalize this sub-section, we point to the improvement of our solutions over the fluid approximations used in [24, 53]. For example, using the competing Lagrangian method in the case of $\mu = \theta$, the maximum staffing level would have been



(a) $Profit\ model, \epsilon = 1, r = 100, 0 < c < 100$

(b) $Profit\ model, \epsilon = 0.1, r = 100, 0 < c < 100$

(c) $Cost\ model, a = 0, c = 1, 1 < d < 100$

(d) $Cost\ model, d = 0, c = 1, 1 < a < 100$

Figure 3. Planes of $1/\varrho$ versus the meaning staffing $\bar{s}^*$ when $\lambda(t) = 100+20\cdot\sin(t)$. [Color figure can be viewed at wileyonlinelibrary.com.]

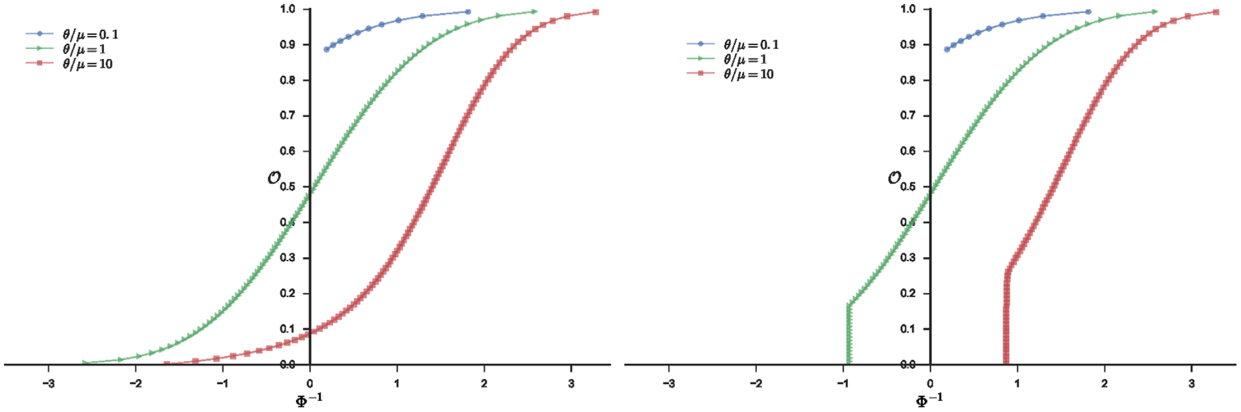
the offered-load, q; $\dot{q} = \lambda - \mu q$, meaning that the changes in c / r would have no effect on the staffing levels. In contrast, our solutions indicate that staffing levels can be increased, to maximize profitability, as c / r decreases (see Fig. 2d).

4.2. Mean Staffing Analysis

The dynamic solutions in Fig. 2 do not allow for in-depth analysis of the changes in staffing levels as the cost-to-benefit ratio changes. In an attempt to examine such changes, we analyze the evolution of the mean staffing levels by plotting $\bar{s}^* \equiv \frac{1}{T}\int_0^T s^* dt$ against $1/\varrho$ (see Fig. 3). In the cost model, $1/\varrho$ represents the ratio of delay over staffing costs whereas for the profit model $1/\varrho$ represents the ratio of revenue over staffing costs. As also observed in [6, 39], Fig. 3 confirms that the mean staffing levels converge to particular values as $1/\varrho$ increases. But there is a clear difference in staffing policies dictated by the ratio θ/μ and the type of model being considered (profit versus cost). For example, in the profit model, when ϵ is binding, there is a minimum number of servers required to ensure that the SLA constraint is satisfied (see Fig. 3b for the truncated tail at the lower end of $1/\varrho$). For the cost model, staffing policies are also influenced by the setting of abandonment costs a versus that of the delay costs d. When $a > 0$, it is expected that as θ/μ increases, also staffing levels should increase to minimize abandonment costs (see Fig. 3d). Likewise, it is expected that more servers would be needed as $\theta/\mu \to 0$, for the case of $d > 0$, to minimize delay costs. What is remarkable for the latter case is that the increase in θ/μ, as $1/\varrho$ increases, has little impact on staffing levels (see Fig. 3c).

4.3. Profitability Analysis

Hamiltonian functions in optimal control theory are generally interpreted as profit rates [54]. For our purposes, such interpretation is natural for profit centers. Accordingly, from Theorem 3.1, we obtain the following Hamiltonian function



(a) $\epsilon = 1$ (b) $\epsilon = 0.1$

Figure 4. Planes of Φ^{-1} versus $\mathcal{O}$ when $\lambda(t) = 100 + 20 \cdot \sin(t)$, $r = 100$, and $0 < c < 100$. [Color figure can be viewed at wileyonlinelibrary.com.]

(see derivation in Derivation of the Hamiltonian Function for the Profit Model in Appendix):

$$\begin{aligned}
&\mathcal{H}(s,p,q,x)\\
&\quad = (r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))) - c \cdot s)\\
&\qquad + p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))\\
&\qquad - \mathrm{x} \cdot \theta \cdot (\phi(\chi) - \chi \cdot \overline{\Phi}(\chi)) \cdot \sqrt{q} \qquad (4.14)
\end{aligned}$$

The first term of Eq. (4.14), $r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))) - c \cdot s$, represents the operating income. This implies that $r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi)))$ is the operating revenue and that $c \cdot s$ is the operating cost. The second term, $p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))$, represents the shadow income from the marginal customer $\dot{q}$. In optimality, the Pontryagin's maximum principle [51] guarantees that the marginal revenue $p^* \cdot \dot{q}^*$ equals the marginal cost $\dot{p}^*$ [54]. The third term, $\mathrm{x} \cdot \theta \cdot (\phi(\chi) - \chi \cdot \overline{\Phi}(\chi)) \cdot \sqrt{q}$, represents penalty costs for violating the SLA constraint.

Under optimal conditions, the shadow income and the penalty costs do not materialize. These quantities rather serve as a guide for pricing the marginal increase in the number of servers. On the contrary, the operating income does materialize and from it we are able to measure profitability using the operating margin, $\mathcal{O}$, defined as follows:

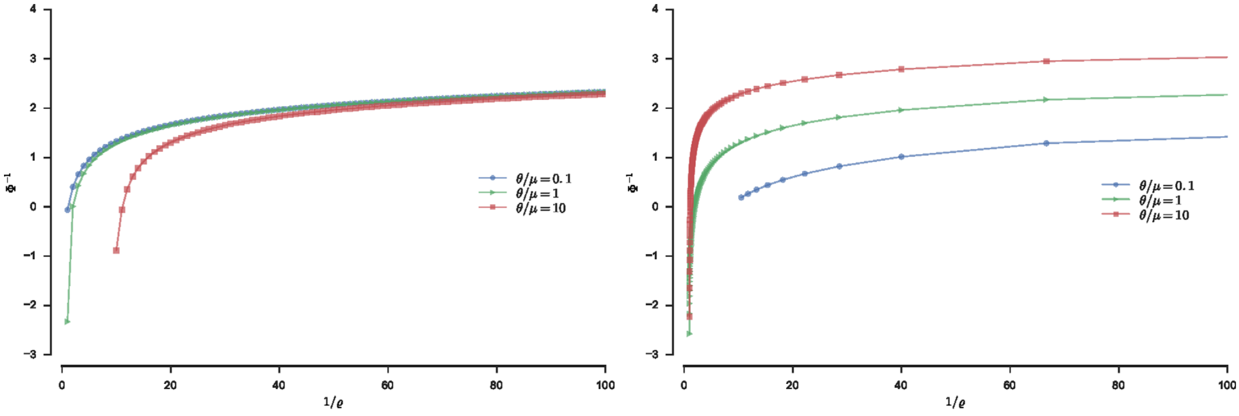
$$\mathcal{O} = \frac{1}{T}\int_0^T \frac{r \cdot \mu \cdot (q^* + \sqrt{q^*} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))) - c \cdot s^*}{r \cdot \mu \cdot (q^* + \sqrt{q^*} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi)))} dt,$$
$$0 < \mathcal{O} < 1 \qquad (4.15)$$

A common interpretation of the operating margin is that the closer to 1 $\mathcal{O}$ is, the more profitable a business is [19]. As it can be noted in Eq. (4.15), factors leading to a lower profit margin include higher staffing costs c or lower operating revenues r. The relationship between the quality of the service grade Φ^{-1} and the operating margin $\mathcal{O}$ is captured in Fig. 4. The SLA constraint in Fig. 4a is non-binding, with $\epsilon = 1$. In this figure, it is apparent that as θ/μ increases the $\mathcal{O}$ curve shifts to the right, which increases the service grade Φ^{-1}. In turn, staffing levels are increased, which eventually leads to lower profitability. In Fig. 4b, the constraint is binding, with $\epsilon = 0.1$. As a result, there is a minimum service grade required, which limits the profitability region of the service center. It should be remarked that as θ/μ gets smaller (see both Figs. 4a and 4b), more profitability is likely since customers are more patient, hence more of them will eventually be served.

The graphical results in Fig. 5 add more insights into the profitability of the service center, given the service grade Φ^{-1}. In this figure, we observe that as ϱ decreases, or simply as the benefits far outweigh the costs, the service grade Φ^{-1} converges to a particular value. This conclusion was also reached in [39], where it was observed that in cost centers, $\Phi^{-1} < 2$, when $d/c \leq 20$, and that $\Phi^{-1} < 3$, when $d/c \leq 500$. We also observe similar results for the cost model in Fig. 5a. The results in Fig. 5b suggest that the service grades remain remarkably apart under various ratios of θ/μ, even as $1/\varrho$ becomes large. The operational intuition is that the less patient customers are, the more servers are needed to maximize profitability.

4.4. Probability of Delay

We have earlier interpreted ϱ as the cost-to-benefit ratio. In addition, ϱ can also be interpreted as the probability of delay. Since delay for service happens when the queue length is



(a) $Cost\ model, a = 0, c = 1, 0 < d < 100$ (b) $Profit\ model, \epsilon = 1$

Figure 5. Planes of $1/\varrho$ versus Φ^{-1} when $\lambda(t) = 100 + 20 \cdot \sin(t)$. [Color figure can be viewed at wileyonlinelibrary.com.]

larger than the number of servers available, then our new interpretation of ϱ is justified, given that

$$\begin{aligned}\mathbb{P}(Q \geq s) &\approx \mathbb{P}(Q^{\infty} \geq s)\\ &\approx \mathbb{P}(q + \sqrt{q} \cdot X \geq q + \sqrt{q} \cdot \Phi^{-1}(1-\varrho))\\ &= \mathbb{P}(X \geq \Phi^{-1}(1-\varrho))\\ &= 1 - \Phi(\Phi^{-1}(1-\varrho))\\ &= \varrho \qquad (4.16)\end{aligned}$$

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

In the context of the newsvendor problem [3], ϱ 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.

A high-level view of the relationship between Φ^{-1} and ϱ is as follows:

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

This relationship is also graphically displayed in Fig. 6. Additionally, Fig. 6 portrays the relationship of ϱ and Φ^{-1} versus the operating margin $\mathcal{O}$. It can be concluded that the higher ϱ, the lower $\mathcal{O}$, implying that when the probability of delay is high, low profitability is expected. An operational explanation of such phenomenon is that more servers cannot be afforded, which leads to a high probability of delay and low profitability since fewer customers are served.

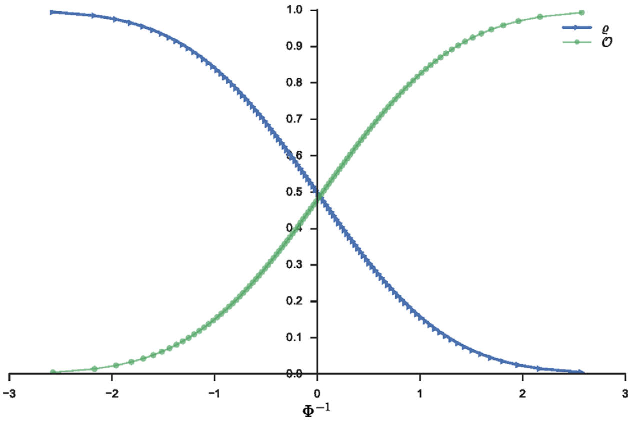
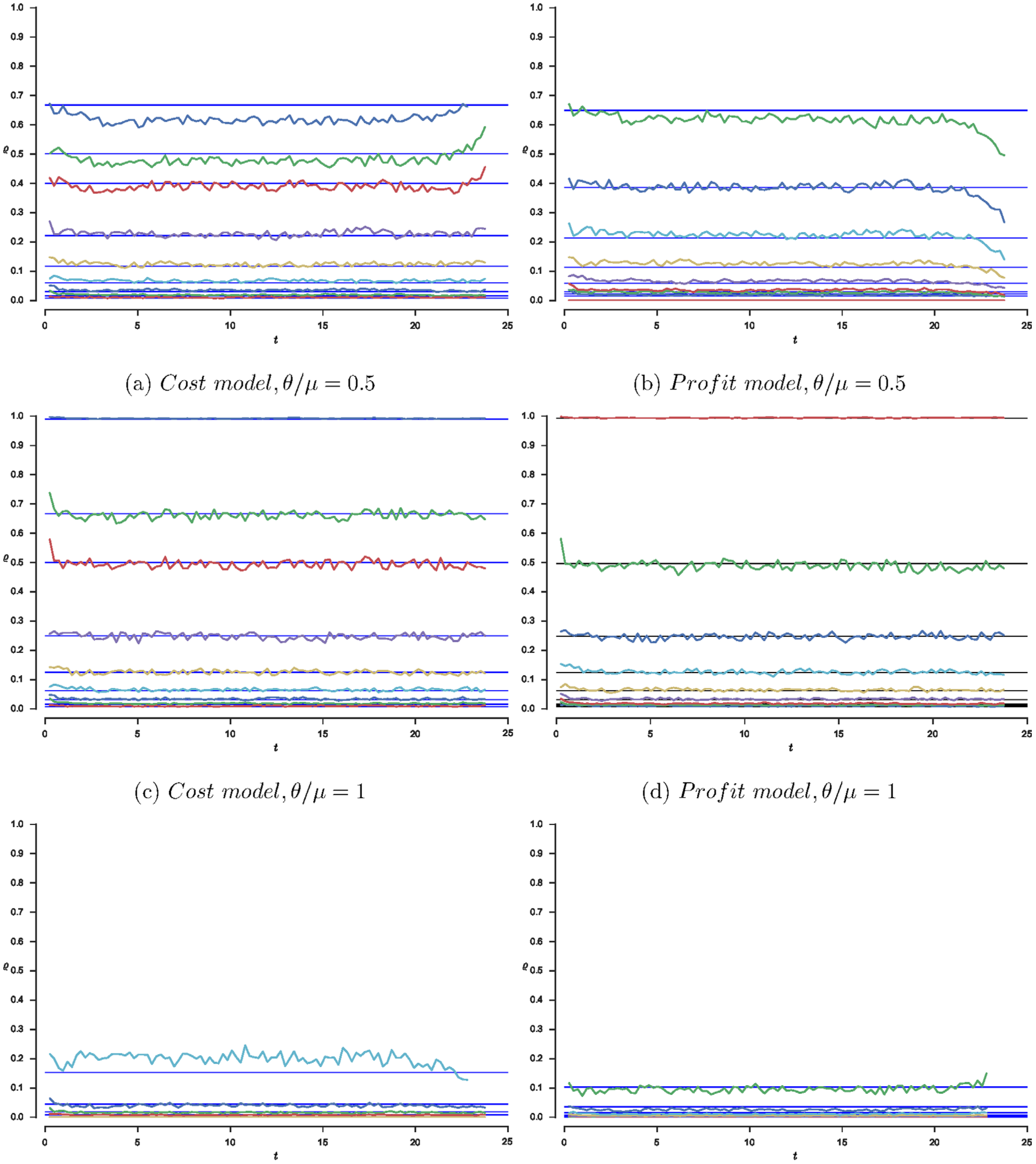


Figure 6. Plane Φ^{-1} versus ϱ and $\mathcal{O}$ when $\lambda(t) = 100+20\cdot\sin(t)$ and $\theta = \mu$. [Color figure can be viewed at wileyonlinelibrary.com.]

The final point about the probability of delay in our model, is that our proposed staffing solutions mostly stabilize this performance as portrayed in Fig. 7. In some cases, a refinement factor may be necessary for better stability. This will be one of the subjects of our future research. For more discussion on stabilizing queueing parameters see [28, 15, 40, 42, 36, 47, 48].

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



(a) $Cost\ model, \theta/\mu = 0.5$

(b) $Profit\ model, \theta/\mu = 0.5$

(c) $Cost\ model, \theta/\mu = 1$

(d) $Profit\ model, \theta/\mu = 1$

(e) $Cost\ model, \theta/\mu = 10$

(f) $Profit\ model, \theta/\mu = 10$

Figure 7. Simulated probability of delay (squiggly lines) versus mean ϱ (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 $a=0$. Only feasible solutions are presented. [Color figure can be viewed at wileyonlinelibrary.com.]

and invoked the Pontryagin's maximum principle to derive a closed form SRS rule for optimal staffing.

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. Additionally, it was established that as the probability of delay increased, profitability decreased.

Throughout the article, we provided theoretical support of our analysis and conducted extensive numerical experiments of our findings. To this end, 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. In all cases, we have observed that the ratio of abandonment over service rate particularly affected staffing levels and, in some instances, led to drastically different policies between the cost and profit centers.

For future research, we will investigate further the stabilization of performance from our model. Also, other queueing approximation techniques will be contemplated. 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 [31, 30, 50]. Lastly, it would be interesting to use risk measures like in the work of [49] and generate optimal control policies for nonstationary queues using risk measures in the objective function or constraints.

APPENDIX

In "Derivation of the Hamiltonian Function for the Profit Model," we construct the Hamiltonian function using optimal control theory. In "Necessary Conditions," we appeal to the Pontryagin's maximum principle to obtain the necessary conditions of our optimal solutions. Proofs to Theorems 3.1, 3.2, and 4.1 are, respectively, provided in "Proof of Theorem 3.1," "Proof of Theorem 3.2," and "Proof of Theorem 4.1." Lastly, the algorithm we use to numerically integrate our dynamical systems is presented in "Numerical Integration Algorithm.".

Derivation of the Hamiltonian Function for the Profit Model

We follow the methods of optimal control theory (e.g. [8, 9, 35, 54]) and proceed to construct a Hamiltonian function of our profit model using the approximation in Eq. (2.10). From Problem 2.1, we obtain the Hamiltonian function as follows:

$$\begin{aligned}\mathcal{H}(s,p,q,x) &= r\cdot\mu\cdot E[(Q\wedge s)] - c\cdot s\\ &\quad + p\cdot(\lambda - \mu\cdot E[(Q\wedge s)] - \theta\cdot E[(Q-s)^+])\\ &\quad - \mathrm{x}\cdot\theta\cdot E[(Q-s)^+]\\ &= r\cdot\mu\cdot E[((q+\sqrt{q}\cdot X)\wedge s)] - c\cdot s\\ &\quad + p\cdot(\lambda - \mu\cdot E[((q+\sqrt{q}\cdot X)\wedge s)]\\ &\qquad - \theta\cdot E[((q+\sqrt{q}\cdot X)-s)^+])\\ &\quad - \mathrm{x}\cdot\theta\cdot E[((q+\sqrt{q}\cdot X)-s)^+]. \end{aligned} \tag{1.17}$$

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

$$\begin{aligned} Z &= -\int_0^t \theta\cdot E[((q(u)+X\cdot\sqrt{q(u)})-s)^+]du\\ \dot{Z} &= -\theta\cdot E[((q+X\cdot\sqrt{q})-s)^+] \end{aligned}$$

where $Z(T)\geq -\mathcal{E}$. Again $\mathcal{E} = \epsilon\cdot\int_0^T \lambda dt$ with ϵ being the maximum allowable probability of abandonment. Since Z does not appear in Eq. (1.17), then $\dot{\mathrm{x}} = -\partial\mathcal{H}/\partial Z = 0$, meaning that x is a constant that satisfies the following complementary of slackness equation:

$$\mathrm{x}\cdot[\mathcal{E} - \int_0^T \theta\cdot E[((q+X\cdot\sqrt{q})-s)^+]dt] = 0 \tag{1.18}$$

Accordingly, $\mathrm{x}=0$ when $\mathcal{E} - \int_0^T \theta\cdot E[((q+X\cdot\sqrt{q})-s)^+]dt > 0$, else $\mathrm{x}>0$. It follows that

$$\begin{aligned} &\mathcal{H}(s,p,q,\mathrm{x})\\ &\quad\approx r\cdot\mu\cdot E[(q+X\cdot\sqrt{q})\wedge s] - c\cdot s\\ &\qquad + p\cdot(\lambda-\mu\cdot E[(q+X\cdot\sqrt{q})\wedge s]\\ &\qquad - \theta\cdot E[((q+X\cdot\sqrt{q})-s)^+] - \mathrm{x}\cdot\theta\cdot E[((q+X\cdot\sqrt{q})-s)^+]\\ &\quad = r\cdot\mu\cdot(q+E[(X\wedge\chi)]\cdot\sqrt{q}) - c\cdot s\\ &\qquad + p\cdot(\lambda-\mu\cdot(q+E[(X\wedge\chi)]\cdot\sqrt{q}) - \theta\cdot E[(X-\chi)^+]\cdot\sqrt{q})\\ &\qquad - \mathrm{x}\cdot\theta\cdot E[(X-\chi)^+]\cdot\sqrt{q} \end{aligned}$$

where

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

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

$$\begin{aligned} E[(X-\chi)^+] &= E[(X-\chi)\cdot\{X\geq\chi\}]\\ &= E[X\cdot\{X\geq\chi\}] - \chi\cdot P\{X\geq\chi\}\\ &= \int_{-\infty}^{\infty}\delta_\chi(y)\cdot\phi(y)dy - \chi\cdot\overline{\Phi}(\chi)\\ &= \phi(\chi) - \chi\cdot\overline{\Phi}(\chi). \end{aligned}$$

Given

$$E[X\wedge\chi] = E[X] - E[(X-\chi)^+],$$

we easily compute

$$\begin{aligned}E[X \wedge \chi] &= E[X - (X - \chi)^{+}]\\ &= 0 - E[(X - \chi)^{+}]\\ &= \chi \cdot \overline{\Phi}(\chi) - \phi(\chi)\end{aligned}$$

Finally, we obtain

$$\begin{aligned}&\mathcal{H}(s, p, q, \mathrm{x})\\ &\quad = (r \cdot \mu \cdot (q + \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))) - c \cdot s)\\ &\qquad + p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))\\ &\qquad - \mathrm{x} \cdot \theta \cdot (\phi(\chi) - \chi \cdot \overline{\Phi}(\chi)) \cdot \sqrt{q} \qquad (1.19)\end{aligned}$$

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 $\mathcal{H}$ with respect to the queue length q and the shadow price p. We proceed as follows:

$$\begin{aligned}\frac{\partial}{\partial q}(\chi\sqrt{q} \cdot \overline{\Phi}) - \frac{\partial}{\partial q}(\sqrt{q} \cdot \phi) &= \chi \cdot \phi \cdot \left(\frac{s+q}{2q}\right) - \overline{\Phi} - \chi \cdot \phi \cdot \left(\frac{s+q}{2q}\right)\\ &\quad - \frac{\phi}{2\sqrt{q}}\\ &= -\left(\overline{\Phi} + \frac{\phi}{2\sqrt{q}}\right),\\ \frac{\partial}{\partial q}(\chi\sqrt{q} \cdot \overline{\Phi}) &= \frac{\partial}{\partial q}((s-q) \cdot \overline{\Phi})\\ &= (s-q) \cdot (\phi) \cdot \left(\frac{s+q}{2q \cdot \sqrt{q}}\right) - \overline{\Phi}\\ &= \chi \cdot \phi \cdot \left(\frac{s+q}{2q}\right) - \overline{\Phi},\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial q}(\sqrt{q} \cdot \phi) &= \sqrt{q} \cdot \chi \cdot \phi \cdot \left(\frac{s+q}{2q\sqrt{q}}\right) + \frac{1}{2\sqrt{q}} \cdot \phi\\ &= \chi \cdot \phi \cdot \left(\frac{s+q}{2q}\right) + \frac{\phi}{2\sqrt{q}}.\end{aligned}$$

The necessary conditions are then given by:

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial p} &\equiv \overset{\bullet}{q} = \lambda - \mu \cdot q - (\mu - \theta) \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi)) \cdot \sqrt{q}\\ -\frac{\partial \mathcal{H}}{\partial q} &\equiv \overset{\bullet}{p} = -r \cdot \mu + r \cdot \mu \cdot \left(\frac{\partial}{\partial q}(\chi \cdot \sqrt{q} \cdot \overline{\Phi}) - \frac{\partial}{\partial q}(\sqrt{q} \cdot \phi)\right)\\ &\quad - p\left(-\mu + (\mu - \theta) \cdot \left(\frac{\partial}{\partial q}(\chi \cdot \sqrt{q} \cdot \overline{\Phi}) - \frac{\partial}{\partial q}(\sqrt{q} \cdot \phi)\right)\right)\\ &\quad - \mathrm{x} \cdot \theta \cdot \left(\frac{\partial}{\partial q}(\chi \cdot \sqrt{q} \cdot \overline{\Phi}) - \frac{\partial}{\partial q}(\sqrt{q} \cdot \phi)\right)\\ &= (\mu \cdot (r-p) + \theta \cdot (p + \mathrm{x})) \cdot \left(\overline{\Phi} + \frac{\phi}{2\sqrt{q}}\right) - \mu \cdot (r-p)\end{aligned}$$

Proof of Theorem 3.1

From the Pontryagin's maximum principle, the optimal control policy s^* that maximizes the Hamiltonian function in Eq. (1.19), 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

$$\begin{aligned}\frac{\partial}{\partial s}(\chi\sqrt{q} \cdot \overline{\Phi}) &= \frac{\partial}{\partial s}(s-q) \cdot \overline{\Phi}\\ &= (s-q) \cdot \frac{-\phi}{\sqrt{q}} + \overline{\Phi}\\ &= \overline{\Phi} - \chi \cdot \phi\end{aligned}$$

and

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

We obtain

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial s} &= r \cdot \mu \cdot \overline{\Phi}(\chi) - c - p \cdot (\mu - \theta) \cdot \overline{\Phi}(\chi) + x \cdot \theta \cdot \overline{\Phi}(\chi) = 0\\ &= (\mu \cdot (r-p) + \theta \cdot (p + \mathrm{x})) \cdot \overline{\Phi}(\chi) - c = 0\end{aligned}$$

We now solve for s by recalling that $\chi = \frac{s-q}{\sqrt{q}}$. We proceed as follows:

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

Finally we obtain the optimal staffing s^* given by

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

Proof of Theorem 3.2

The Hamiltonian function associated with the cost model (see Problem 2.2) is given by

$$\begin{aligned}\mathcal{H}(s, p, q, x) = &-c \cdot s - (d + \theta \cdot a) \cdot \sqrt{q} \cdot (\phi(\chi) - \chi \cdot \overline{\Phi}(\chi))\\ &+ p \cdot (\lambda - \mu \cdot q - (\mu - \theta) \cdot \sqrt{q} \cdot (\chi \cdot \overline{\Phi}(\chi) - \phi(\chi))\\ &\qquad (1.20)\end{aligned}$$

The resulting necessary conditions follow.

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

The optimal staffing policy s^* is obtained by

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

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 Ref. [45]. □

LEMMA A.2:

$$\Gamma(s,x) = \sum_{m=s}^{\infty} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(s)} \int_0^x e^{-y} y^{s-1} dy$$
$$\overline{\Gamma}(s,x) = \sum_{m=0}^{s-1} e^{-x} \cdot \frac{x^m}{m!} = \frac{1}{\Gamma(s)} \int_x^{\infty} e^{-y} y^{s-1} dy.$$

where

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

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

PROOF: See Ref. [27]. □

LEMMA A.3:

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

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

$$\begin{aligned} \frac{\partial\mathcal{H}}{\partial s} &= r\cdot\mu\cdot\Gamma(s+1,q) - c - p\cdot(\mu-\theta)\cdot\Gamma(s+1,q) \\ &\quad + \mathrm{x}\cdot\theta\cdot\Gamma(s+1,q) = 0 \\ &= \mu\cdot(r+\mathrm{x})\cdot\Gamma(s+1,q) - c = 0 \end{aligned}$$

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

$$\begin{aligned} \mu\cdot(r+\mathrm{x})\cdot\Gamma(s+1,q) - c = 0 \\ \Gamma(s+1,q) = \frac{c}{\mu\cdot(r+\mathrm{x})}. \end{aligned}$$

Finally we obtain optimal control policy s^* as:

$$s^* = \Gamma^{-1}\left(q, \frac{c}{\mu\cdot(r+\mathrm{x})}\right).$$

Numerical Integration Algorithm

Our algorithm, detailed next, is based on the Forward-Backward method [35]. The new element to the algorithm is step 4 to allow for the computation of the complementary of slackness in Eq. (1.18).

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

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

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

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

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

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

where h is a very small increment.

Step 5: Repeat Step 1–3 until the relative error is negligible, in accordance to

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

where δ is the accepted convergence tolerance.

For further discussion on convergence of forward-backward algorithms see [37, 43, 56].

REFERENCES

[1] O. Baron and J. Milner, Staffing to maximize profit for call centers with alternate service-level agreements, Oper Res 57 (2009), 685–700.

[2] A. Bassamboo, J.M. Harrison, and A. Zeevi, Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits, Queueing Syst 51 (2005), 249–285.

[3] A. Bensoussan, M. Çakanyildirim, and S.P. Sethi, A multiperiod newsvendor problem with partially observed demand, Math Oper Res 32 (2007), 322–344.

[4] A. Bhandari, A. Scheller-Wolf, and M. Harchol-Balter, An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers, Manage Sci 54 (2008), 339–353.

[5] V. Bolotin, Telephone circuit holding time distributions, Proc ITC 14 (2013), 125–134.

[6] S. Borst, A. Mandelbaum, and M.I. Reiman, Dimensioning large call centers, Oper Res 52 (2004), 17–34.

[7] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, Statistical analysis of a telephone call

center: A queueing-science perspective, J Am Stat Assoc 100 (2005), 36–50.

[8] M.R. Caputo, Foundations of dynamic economic analysis: Optimal control theory and applications, Cambridge University Press, New York, NY, 2005.

[9] A. Chiang, Elements of dynamic optimization, Waveland Press Inc, Illinois, 2000.

[10] M. Defraeye and I. Van Nieuwenhuyse, Controlling excessive waiting times in small service systems with time-varying demand: An extension of the isa algorithm, Decision Supp Syst 54 (2013), 1558–1567.

[11] S.G. Eick, W.A. Massey, and W. Whitt, The physics of the mt/g/ queue, Oper Res 41 (1993), 731–742.

[12] S.G. Eick, W.A. Massey, and W. Whitt, Mt/g/ queues with sinusoidal arrival rates, Manage Sci 39 (1993), 241–252.

[13] S. Engblom and J. Pender, Approximations for the moments of nonstationary and state dependent birth-death queues, 2014, arXiv preprint arXiv:1406.6164. Available at: https://arxiv.org/abs/1406.6164v2.

[14] A. Erlang, "On the rational determination of the number of circuits," in: E. Brockmeyer, H.L. Halstrom, and A. Jensen A (Editors), The life and works of A.K. Erlang, The Copenhagen Telephone Company, Copenhagen, 1948, 216–221.

[15] Z. Feldman, A. Mandelbaum, W.A. Massey, and W. Whitt, Staffing of time-varying queues to achieve time-stable performance, Manage Sci 54 (2008), 324–338.

[16] S.A. Finkler, D. Ward, and J.J. Baker, Essentials of cost accounting for health care organizations, 3rd Ed., Jones & Bartlett Learning, Mississauga, Ontario Canada, 2007.

[17] M.C. Fu, S.I. Marcus, and I.-J. Wang, Monotone optimal policies for a transient queueing staffing problem, Oper Res 48 (2000), 327–331.

[18] N. Gans, G. Koole, and A. Mandelbaum, Telephone call centers: Tutorial, review, and research prospects, Manuf Service Oper Manag 5 (2003), 79–141.

[19] C. Gapenski, Fundamentals of Healthcare Finance, 2nd Ed, Health Administration Press, Chicago, IL, 2012.

[20] O. Garnett, A. Mandelbaum, and M. Reiman, Designing a call center with impatient customers, Manuf Service Oper Manage 4 (2002), 208–227.

[21] L.V. Green, P.J. Kolesar, and W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system, Prod Oper Manage 16 (2007), 13–39.

[22] S. Halfin and W. Whitt, Heavy-traffic limits for queues with many exponential servers, Oper Res 29 (1981), 567–588.

[23] R.C. Hampshire and W.A. Massey, Dynamic optimization with applications to dynamic rate queues, TUTORIALS in Operations Research, INFORMS Society, 2010, pp. 210–247.

[24] R.C. Hampshire, O.B. Jennings, and W.A. Massey, A time-varying call center design via Lagrangian mechanics, Prob Eng Inf Sci 23 (2009), 231–259.

[25] S. Helber and K. Henken, Profit-oriented shift scheduling of inbound contact centers with skills-based routing, impatient customers, and retrials, Or Spectr, 32 (2010), 109–134.

[26] N. Izady and D. Worthington, Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments, Eur J Oper Res 219 (2012), 531–540.

[27] A. Janssen, J. Van Leeuwaarden, and B. Zwart, Gaussian expansions and bounds for the Poisson distribution applied to the Erlang b formula, Adv Appl Probab (2008), 122–143.

[28] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt, Server staffing to meet time-varying demand, Manage Sci 42 (1996), 1383–1394.

[29] Ö. Kabak, F. Ülengin, E. Aktaş, Ş. Önsel, and Y.I. Topcu, Efficient shift scheduling in the retail sector through two-stage optimization, Eur J Oper Res 184 (2008), 76–90.

[30] Y.M. Ko and J. Pender, Diffusion limits for the (mapt/ph t/) n queueing network, 2016. Available at: https://people.orie.cornell.edu/jpender/MAP_MAP_INF_Network.pdf

[31] Y.M. Ko and J. Pender, Strong approximations for time varying infinite-server queues with non-renewal arrival and service processes, 2016. Available at: https://people.orie.cornell.edu/jpender/MAP_MAP_INF_Network.pdf

[32] Y.L. Koçağa and A.R. Ward, Admission control for a multi-server queue with abandonment, Queueing Syst 65 (2010), 275–323.

[33] G. Koole and A. Pot, A note on profit maximization and monotonicity for inbound call centers, Technical Report, Working Paper. Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands, 2006.

[34] S. Lam, M. Vandenbosch, and M. Pearce, Retail sales force scheduling based on store traffic forecasting, J Retail 74 (1998), 61–88.

[35] S. Lenhart and J.T. Workman, Optimal control applied to biological models, CRC Press, Boca Raton, FL, 2007.

[36] Y. Liu and W. Whitt, Stabilizing performance in networks of queues with time-varying arrival rates, Probab Eng Inf Sci 28 (2014), 419–449.

[37] S. Lv, R. Tao, and Z. Wu, Maximum principle for optimal control of anticipated forward–backward stochastic differential delayed systems with regime switching, Optim Control Appl Methods 37 (2016), 154–175.

[38] A. Mandelbaum, W.A. Massey, and M.I. Reiman, Strong approximations for Markovian service networks, Queueing Syst 30 (1998), 149–201.

[39] Z. Mandelbaum, Service engineering class 13-qed (qd, ed) queues-Erlang-a (m/m/n+g) in the qed & ed regime, University Lecture, Technion - Israel Institute of Technology, 2016. Available at: http://ie.technion.ac.il/serveng/Lectures/QED_lecture_Erlang_A.pdf

[40] W. Massey and J. Pender, Approximation and stabilizing Jackson networks with abandonment, Technical Report, Cornell University, Working Paper, 2013. Available at: https://people.orie.cornell.edu/jpender/Loss_Network.pdf

[41] W.A. Massey and J. Pender, Poster: Skewness variance approximation for dynamic rate multiserver queues with abandonment, ACM Sigmetrics Perform Eval Rev 39 (2011), 74–74.

[42] W.A. Massey and J. Pender, Gaussian skewness approximation for dynamic rate multi-server queues with abandonment, Queueing Syst 75 (2013), 243–277.

[43] M. McAsey, L. Mou, and W. Han, Convergence of the forward-backward sweep method in optimal control, Comput Optim Appl 53 (2012), 207–226.

[44] K.A. Merchant, Rewarding results: Motivating profit center managers, Harvard Business School Press, Brighton, Massachusetts, 1989.

[45] G. Peccati and M. Taqqu, Wiener chaos: Moments, cumulants and diagrams: A survey with computer implementation, Vol. 1, Springer Science & Business Media, Milan, Italy, 2011.

[46] J. Pender, Laguerre polynomial approximations for nonstationary queues, 2014. Available at: https://people.orie.cornell.edu/jpender/LSA.pdf

[47] J. Pender, Gram Charlier expansion for time varying multiserver queues with abandonment, SIAM J Appl Math 74 (2014), 1238–1265.

[48] J. Pender, Nonstationary loss queues via cumulant moment approximations, Probab Eng Inf Sci 29 (2015), 27–49.

[49] J. Pender, Risk measures and their application to staffing nonstationary service systems, Eur J Oper Res 254 (2016), 113–126.

[50] J. Pender and Y.M. Ko, Approximations for the queue length distributions of time-varying many-server queues, 2016. Available at: https://people.orie.cornell.edu/jpender/Phase_paper.pdf

[51] L.S. Pontryagin, Mathematical theory of optimal processes, CRC Press, New York, NY, 1987.

[52] P. Quinn, B. Andrews, and H. Parsons, Allocating telecommunications resources at ll bean, inc, Interfaces 21 (1991), 75–91.

[53] H. Rudolph, Optimal staffing in a hospital's emergency department through dynamic optimization: A queueing perspective, 2011.

[54] S.P. Sethi and G.L. Thompson, Optimal control theory: Applications to management science and economics, Springer, New York, NY, 2005.

[55] C. Stein, Approximate computation of expectations, Lect Notes-Monogr Ser, 7 (1986), i–164.

[56] P. Tseng, A modified forward-backward splitting method for maximal monotone mappings, SIAM J Control Optim 38 (2000), 431–446.

[57] A. Weerasinghe and A. Mandelbaum, Abandonment versus blocking in many-server queues: Asymptotic optimality in the qed regime, Queueing Syst 75 (2013), 279–337.

[58] W. Whitt, Engineering solution of a basic call-center model, Manage Sci 51 (2005), 221–235.

[59] J.R. Williams, S.F. Haka, M.S. Bettner, and J.V. Carcello, Financial & managerial accounting the basis for business decisions, McGraw-Hill Education, New York, NY, 2014.

[60] G.B. Yom-Tov and A. Mandelbaum, Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing, Manuf Service Oper Manage 16 (2014), 283–299.