OPERATIONS RESEARCH
Vol. 57, No. 3, May–June 2009, pp. 714–726
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5703 | 0714

DOI 10.1287/opre.1080.0602

On a Data-Driven Method for Staffing Large Call Centers

Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
a-bassamboo@northwestern.edu

Assaf Zeevi
Graduate School of Business, Columbia University, New York, New York 10027,
assaf@gsb.columbia.edu

We consider a call center model with multiple customer classes and multiple server pools. Calls arrive randomly over time, and the instantaneous arrival rates are allowed to vary both temporally and stochastically in an arbitrary manner. The objective is to minimize the sum of personnel costs and expected abandonment penalties by selecting an appropriate staffing level for each server pool. We propose a simple and computationally tractable method for solving this problem that requires as input only a few system parameters and historical call arrival data for each customer class; in this sense the method is said to be *data-driven*. The efficacy of the proposed method is illustrated via numerical examples. An asymptotic analysis establishes that the prescribed staffing levels achieve near-optimal performance and characterizes the magnitude of the optimality gap.

Subject classifications: stochastic model applications; stochastic networks; nonstationary queues; limit theorem; approximations.
Area of review: Stochastic Models.
History: Received May 2006; revisions received February 2007, September 2007, February 2008; accepted March 2008. Published online in *Articles in Advance* March 16, 2009.

1. Introduction

1.1. Introduction and Overview of the Main Contributions

This paper is concerned with the problem of staffing a telephone call center that serves multiple customer classes using agents of multiple skill sets. In particular, we study a tractable solution method to this problem, whose salient feature is that it is *data-driven*. By that we mean that it essentially requires only historical call data to arrive at staffing decisions, and in doing so imposes minimal assumptions with regard to the nature of this data. To the best of our knowledge, this represents a departure from most studies of the staffing problem in the operations research literature, which are typically *model-based*, insofar as solutions proposed there are constructed using probabilistic structure that is assumed to characterize the data-generating process; see Gans et al. (2003) for a comprehensive survey on the topic of modeling and analysis of telephone call centers. (A review of literature relevant to this paper is deferred to the end of this section.)

Before we can explain in more detail the contributions of this paper, let us first describe in broad strokes the call center model that will be the focus of our attention. As indicated above, our model has multiple customer classes and multiple agent pools. Each of the pools consists of identical servers (agents), whose skills dictate the possible customer classes they can serve and the speed at which such service is delivered. Customers of various classes arrive randomly over time and upon arrival are either served immediately or wait in an infinite-capacity buffer.

Two important assumptions are made with regard to the call arrival process. First, *arrival rates* of incoming calls are not assumed to be constant or known. Rather, we allow these rates to be temporally varying and random; that is, there is inherent uncertainty with respect to their true value. Second, we assume that customers waiting for their service to commence might abandon before they are assigned a server. Both of these modeling assumptions capture key characteristics present in actual telephone call center operating environments; see, e.g., Gans et al. (2003) and Steckley et al. (2004) for a discussion of the latter, and Avramidis et al. (2004) and Brown et al. (2005) for discussion of the former, as well as references therein.

To describe the staffing problem, consider a fixed and given time interval, hereafter referred to as a "staffing segment," over which staffing decisions are held constant. That is, a segment represents the smallest time interval over which a staffing decision cannot be revised (typically this interval ranges from 30 minutes to two hours). We assume that there are two types of costs related to operating the call center: *personnel costs* and *abandonment costs*. The objective is then to find a staffing level for the various server pools that minimizes the sum of the two costs over a

given staffing segment. (The solution of the staffing problem usually forms the basis for more detailed workforce management decisions that assign individual agents to specific work schedules, although this level of granularity is beyond the topic of this paper.)

A major obstacle in solving for the "optimal" staffing levels is that the performance of any proposed solution requires specification of a routing policy that describes how incoming calls will be assigned to agents at any point in time, so as to minimize abandonment-related costs. Unfortunately, this dynamic control problem can rarely be solved, and thus it has become common practice to consider only relatively simple call routing rules and then rely on simulation to evaluate the performance of a given staffing level; see Gans et al. (2003) and the literature review that follows. As a consequence of this limitation, it would be difficult to discern, for example, whether a given staffing level performs poorly because the associated routing logic is deficient, or whether this is due to the staffing level itself being strictly suboptimal. (In particular, it is not clear how one identifies, even in theory, the *optimal* solution of the staffing problem; the characterization of the latter is obviously useful for purposes of benchmarking any other proposed solution.) One of the contributions of this paper is that it provides an approach for studying the staffing problem essentially in "isolation."

Main Contributions. The main *algorithmic contribution* of this paper is in proposing a computationally tractable method for obtaining prescriptive solutions to the staffing problem described above. (See §4.2.) The method builds on recent work of Harrison and Zeevi (2005), but whereas that paper develops a *model-based* approach (namely, it assumes knowledge of the probabilistic structure characterizing the mean call arrival patterns), this paper relies only on past call data as an input.

The main *theoretical and methodological contribution* of this paper is in establishing that the proposed data-driven staffing method yields prescriptions that are provably near-optimal in a suitable asymptotic sense. This analysis blends two different types of asymptotics that form the basis of our main results. The first asymptotic considers a sequence of systems in which call arrival volume, as well as abandonment and processing rates, increase without bound. This type of asymptotic is of the variety used in operations research studies of high-volume large-scale systems. In particular, we use multiscale fluid limit machinery developed in Bassamboo et al. (2005) (see also Bassamboo et al. 2006) to characterize the "approximation error" that results from applying our method to finite sized systems; see Theorems 1 and 2. The second type of asymptotic is one that is frequently used to study properties of statistical estimators, and it involves the size of the data growing large. In particular, we rely on machinery from empirical process theory (see, e.g., van der Vaart and Wellner 1996) to characterize the "estimation error" that stems from having access only to historical call arrival data; see Theorems 3 and 4.

To the best of our knowledge, the combination of the two types of asymptotics discussed above is new in the operations research literature, and this paper illustrates the benefits of this synergistic approach; see in particular Theorem 4. Using these results, the performance of the proposed solution is seen to be eventually "close" to the best achievable performance, and in that sense it is asymptotically optimal.

The above claim of asymptotic optimality appears to be somewhat peculiar in light of the earlier discussion on the difficulties of characterizing the optimal solution to the staffing problem. In particular, recall that the latter requires knowledge of the optimal routing logic that should be paired with it. To this end, imagine that one has access to an oracle that provides the optimal routing policy associated with any given staffing vector and hence can compute the optimal solution to the staffing problem. With this aid of the oracle, it is possible to assess the loss in performance that stems from using our proposed solution as opposed to the optimal one. The surprising observation is that we can characterize this optimality gap without ever having to compute the (oracle-based) optimal staffing solution. The techniques used to establish this might be of independent interest and could prove useful in other related problems of design and control of stochastic systems.

The Remainder of the Paper (and a Reading Guide). This section concludes with a review of related literature. Section 2 provides a mathematical description of our call center model, and §3 describes the staffing problem. Reading both sections is in some sense necessary to follow §4, which explains our data-driven solution method. Section 5 provides the intuition behind the proposed method. Those who are not in need of such intuition can skip this section and move on to §6, which presents numerical examples illustrating the performance of the method. Those seeking theory that supports the results observed in the numerical examples can find that in §7. Some qualitative insights and other points pertaining to these theoretical results are summarized in §8. Finally, all proofs are collected in two appendices, which are part of the online companion for the paper (available at http://or.journal.informs.org/): the main results are proved in Appendix A, and auxiliary results are proved in Appendix B.

1.2. Literature Review

Most of the literature on call center staffing focuses on a single pool of identical servers. In that realm, the case where there is only a single class of customers leads to trivial control decisions, and if the system is Markovian then the Erlang-C formula provides the main mathematical tool for solving the staffing problem. An important rule-of-thumb that arises from the Erlang-C formula is the so-called *square-root staffing rule*; see Gans et al. (2003, §4.1.1) for further discussion. Borst et al. (2004) refine the square-root rule to balance queuing and staffing costs; this type of

objective function is similar to what we use in our paper, but we take abandonments as the indicator of congestion-related costs. Extensions include Garnett et al. (2002) that incorporates abandonments, as well as Feldman et al. (2008) and recent work by Mandelbaum and Zeltyn (2008).

Staffing a single pool of servers when there are multiple customer classes involves a significant escalation in complexity because the control problem must be tackled as well. Research on this problem has started only recently, and the primary example of such work is that of Gurvich et al. (2008), which exploits many server diffusion limits in the so-called quality- and efficiency-driven regime (QED) first introduced by Halfin and Whitt (1981). Work on the staffing problem in the context of a multiclass/multipool model is still in its infancy and relies mostly on simulation-based methods; for an example of the latter see Wallace and Whitt (2005), Cezik and L'Ecuyer (2008); for further discussion see Gans et al. (2003).

Our paper is closely related to the recent work of Harrison and Zeevi (2005), which proposes a method for staffing multiclass/multipool call centers and moreover allows for temporal variation and randomness (uncertainty) in arrival rates. In Harrison and Zeevi (2005), a model-based approach is taken, in the sense that the prescribed staffing levels are computed by taking as input the probabilistic structure of the arrival rate process. This makes the approach impractical to implement because it relies on idealized information that is typically not available in any realistic setting. In contrast, this paper prescribes a solution to the staffing problem using only historical data.

The work of Bassamboo et al. (2006) establishes, using machinery of multiscale fluid limits, that the method presented in Harrison and Zeevi (2005) is asymptotically optimal. We rely on that machinery here as well, but unlike Bassamboo et al. (2006), this paper uses it to: (i) develop a key element in the estimation technique; and (ii) provide performance bounds that characterize the optimality gap (as opposed to just establishing that this gap shrinks to zero).

As indicated earlier, the output of any staffing method requires a control to be paired with it. In this paper, we do not explicitly address this issue and in fact show how one can essentially decouple the staffing problem from such considerations. Having said that, the bounds we derive on the optimality gap build on the fact that one can at least characterize an *asymptotically optimal* control. To that end, we rely on earlier work reported in Bassamboo et al. (2005), where asymptotically optimal solutions to the dynamic routing problem are derived when staffing levels are exogenously determined.

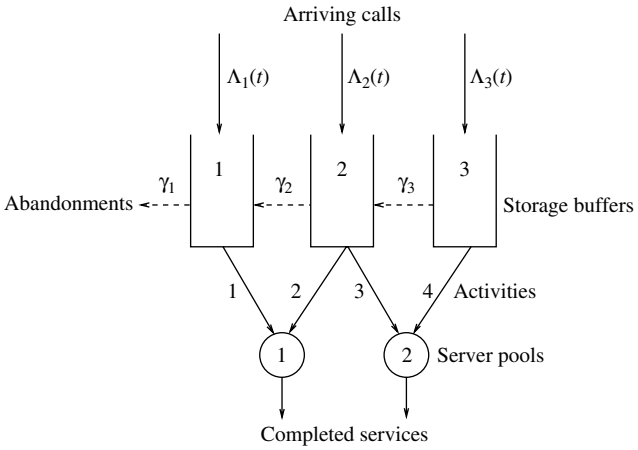
To the best of our knowledge, this paper is the first to provide a data-driven solution for staffing multiclass/multiskilled call centers and to prove that the proposed solution enjoys some optimality properties. There are few other papers in the operations research literature that rely on data-driven approaches, and one method of choice seems to be that of stochastic approximations; for an example of an application in revenue management, see van Ryzin and McGill (2000). Our work is quite different because it involves batch (off-line) optimization based on historical data, the framework being that of empirical risk minimization, which is a classical approach for obtaining "good" statistical estimators (see, e.g., van der Vaart and Wellner 1996).

2. The System Model

Our call center model has m customer classes and r server pools, each consisting of b_k identical servers ($k = 1, \ldots, r$) that can be *cross-trained* to handle customers of several different classes. Similarly, there might be several pools that are able to handle a given customer class. Customers of the various classes arrive randomly over time according to a doubly stochastic Poisson process with *instantaneous arrival rates* given by $\Lambda_1(t), \ldots, \Lambda_m(t)$; a more precise definition will be given shortly. Those customers who cannot be served immediately wait in an infinite-capacity buffer that is dedicated to their specific class. An example with $m = 3$ customer classes and $r = 2$ server pools is shown schematically in Figure 1.

Preliminaries, Notation, and Basic Modeling Assumptions. To describe server capabilities, we will use the notion of processing "activities." There are a total of ℓ processing activities available to the system manager, each of which corresponds to servers from one particular pool serving customers of one particular class (activities are denoted by solid arrows leading from buffers to server pools in Figure 1). For each activity $j = 1, \ldots, \ell$, we denote by $i(j)$ the customer class being served, by $k(j)$ the server pool involved, and by μ_j the associated mean service rate (that is, the reciprocal of the mean of the service time distribution). The actual service times are taken to be exponentially distributed random variables with the above rates, these being independent of one another and also of the

Figure 1. A call center with three customer classes, two agent pools, and four activities.



arrival processes. Note that we allow the service time distribution of a customer to depend on both the customer's class and on the pool to which the server belongs.

We define two matrices: an $m \times \ell$ matrix R with entries $R_{ij} = \mu_j$ if $i = i(j)$ and $R_{ij} = 0$ otherwise, and an $r \times \ell$ matrix A with entries $A_{kj} = 1$ if $k = k(j)$ and $A_{kj} = 0$ otherwise. Thus, one interprets R as an *input-output matrix*, its (i, j)th element specifies the average rate at which activity j removes class i customers from the system; and A is a *capacity consumption matrix*, its (k, j)th element is one if activity j draws on the capacity of server pool k and is zero otherwise. We define an $m \times \ell$ matrix B to describe which customer class is served by which activity, setting $B_{ij} = 1$ if $i(j) = i$ and $B_{ij} = 0$ otherwise.

An important assumption of our model is that customers of any given class will abandon their calls if forced to wait too long for the commencement of service; abandoned calls are represented by the horizontal dotted arrows emanating from the storage buffers in Figure 1. Specifically, each class i customer is endowed with an exponentially distributed "impatience" random variable τ that has mean $1/\gamma_i$, independent of the impatience random variables characterizing other customers and of service times and arrival processes. The customer will abandon the call when his or her waiting time in queue (exclusive of service time) reaches a total of τ time units. This assumption is quite standard in call center modeling; cf. Garnett et al. (2002), Gans et al. (2003), and Harrison and Zeevi (2004). Let $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_m)$ denote the *abandonment rate matrix*.

Consider an interval $[0, T]$ that represents the staffing segment of interest (taking zero to be the starting point is merely a convenient normalization). Let $\Lambda_i = (\Lambda_i(t)\colon 0 \leqslant t \leqslant T)$, $i = 1, \ldots, m$ denote the arrival rate process in each customer class, which is nonnegative and has continuous ample paths, such that $\mathbb{E}\int_0^T \Lambda_i(s)\,ds$ exists. Let $N_i^{(d)} = (N_i^{(d)}(t)\colon 0 \leqslant t < \infty)$ be mutually independent Poisson processes, each with unit rate, for $i = 1, \ldots, m$ and $d = 1, 2, 3$. The Poisson processes are further taken to be independent of the arrival rate processes. We use the processes $(N_1^{(1)}, \ldots, N_m^{(1)})$ to construct arrivals in our model, defining

$$F_i(t) := N_i^{(1)}\left(\int_0^t \Lambda_i(s)\,ds\right) \quad \text{for } i = 1, \ldots, m \text{ and } 0 \leqslant t \leqslant T. \tag{1}$$

This is a standard construction of a doubly stochastic Poisson process; cf. Bremaud (1981). We interpret $F_i(t)$ as the cumulative number of class i arrivals up to time t. The unit-rate Poisson processes $N_i^{(2)}$ and $N_i^{(3)}$ will be used to construct service completions and abandonments, respectively, under a given control policy, via relationships analogous to (1).

Staffing, Control Formulation, and System Dynamics. Let $b = (b_1, \ldots, b_r)$ denote a *staffing vector*, whose kth component is the number of servers to be employed during the specified planning period for server pool k. By assumption, the value of b cannot be revised as actual demand is observed during the period $[0, T]$; for ease of reference, we will refer to this time horizon as a "staffing segment." In what follows, we will relax integrality constraints and for simplicity allow b to take values in $\mathbb{R}_+^r$; because we focus on high-volume call centers, where a large number of servers are used per staffing segment, this distinction is not crucial for our purposes.

Given a staffing vector b, we define a *control* as a stochastic process $X = (X(t)\colon 0 \leqslant t \leqslant T)$ taking values in $\mathbb{R}_+^n$, whose sample paths are right continuous with left limits and Lebesgue integrable. Furthermore, we assume that X is nonanticipating, i.e., it is adapted to the filtration generated by the arrival rate processes Λ, arrivals, service completions, and abandonments. Writing $X(t) = (X_1(t), \ldots, X_\ell(t))$, we interpret $X_j(t)$ as the number of servers engaged in activity j at time t. A control X is said to be *admissible* with respect to a staffing vector b if there exist processes Z and Q, both having time domain $[0, T]$, both taking values in $\mathbb{R}_+^m$, and both necessarily unique, that jointly satisfy conditions (2)–(4) below for all $t \in [0, T]$:

$$AX(t) \leqslant b, \tag{2}$$

$$Q(t) = Z(t) - BX(t) \geqslant 0, \tag{3}$$

$$Z_i(t) = F_i(t) - N_i^{(2)}\left(\int_0^t (RX)_i(s)\,ds\right) - N_i^{(3)}\left(\int_0^t \gamma_i Q_i(s)\,ds\right) \geqslant 0, \quad i = 1, \ldots, m. \tag{4}$$

Here $Z_i(t)$ represents the number of class i customers in the system at time t (we call Z the *headcount process* and Z_i its ith component); and $Q_i(t)$ represents the number of class i customers in the buffer that are waiting for service at time t (we call Q the *queue length process* and Q_i its ith component).

Constraint (2) requires that the number of servers in various pools that are engaged in some activity at time t cannot exceed the total number of servers in each pool. In the second constraint (3), $BX(t)$ is a vector whose components represent the numbers of servers allocated to various customer classes at time t. The constraint therefore prohibits allocating to a given class a number of servers that exceeds the headcount in that class. The third constraint (4) is the system dynamics equation, where the second term on the right-hand side is interpreted as the cumulative number of class i service completions up to time t, while the third term represents cumulative class i abandonments. The instantaneous departure rate for class i customers due to abandonments is $\gamma_i Q_i$, and the instantaneous departure rate for class i due to service completions is $\sum \mu_j X_j$, where the sum is taken over activities j that serve class i. (It is straightforward to establish that there (almost surely) exists *at most* one pair (Z, Q) satisfying (3) and (4); see Bassamboo et al. 2006.) For future purposes, we use $\mathscr{X}(b)$ to denote the set of admissible controls for a given staffing level b.

3. Problem Formulation

The Staffing Problem and Best Achievable Performance. Let $p = (p_1, \ldots, p_m)$ be the *penalty cost vector*, where p_i is the cost associated with an abandonment of a class i customer, and let $c = (c_1, \ldots, c_r)$ be the *personnel cost vector*, where c_k is the cost of employing a server in pool k for the entire planning horizon $[0, T]$, which constitutes a staffing segment. The total cost associated with a given staffing vector b and an admissible control X is given by the functional

$$\begin{aligned}\mathcal{J}(b, X) &:= c \cdot b + \sum_{i=1}^{m} p_i \mathbb{E}\left[N_i^{(3)}\left(\int_0^T \gamma_i Q_i(s)\, ds \right)\right] \\ &= c \cdot b + \mathbb{E}\left[\int_0^T p \cdot \Gamma Q_i(s)\, ds \right], \end{aligned} \tag{5}$$

using the properties of Poisson process and the vector notation introduced in §2. Here $x \cdot y$ represents the scalar product of the vectors x and y. The *minimal* expected total cost associated with a staffing vector b is given by

$$\mathcal{V}(b) = \inf_{X \in \mathcal{X}(b)} \mathcal{J}(b, X), \tag{6}$$

where the minimization is over the set of admissible controls. We refer to $\mathcal{V}(\cdot)$ as the *performance function*. The goal of the system manager is to find the *optimal staffing level* b_* that solves the following optimization problem:

$$\inf\{\mathcal{V}(b):\ b \in \mathbb{R}_+^r\}. \tag{7}$$

Given that $\mathcal{V}(b)$ is defined via (6), it is not clear a priori whether this operation results in a bona fide (measurable) function, and hence it is not clear if one can define the optimization problem in (7) using $\mathcal{V}(\cdot)$. To this end, we have the following result.

Proposition 1. *The mapping $\mathcal{V}: \mathbb{R}_+^r \to \mathbb{R}_+$ defined via the optimization problem* (6) *is Lipschitz continuous.*

Thus, $\mathcal{V}(\cdot)$ is differentiable almost everywhere with bounded derivative (where it exists). The continuity is sufficient to ensure that the minimum in (7) is achieved by a vector b_* because the domain of the optimization problem can be restricted to the compact set $\{b:\ c \cdot b \leqslant \mathcal{V}(0),\ b \in \mathbb{R}_+^r\}$. The corresponding optimal value will be denoted by

$$\mathcal{V}_* = \mathcal{V}(b_*), \tag{8}$$

a quantity that we will refer to henceforth as the *best achievable performance*.

Of course, what we have just described is a highly idealized sequence of steps. Even if the probabilistic structure of the instantaneous arrival rates is known and it is possible to compute $\mathcal{J}(b, X)$, it is virtually impossible to solve the dynamic optimization problem (6) over the space of admissible controls. Hence, one can view the derivation of the staffing problem (7) along with its solution $b_* \in \arg\min\{\mathcal{V}(b)\}$ as being made possible only with the aid of an "oracle," having at its disposal all information on the primitive processes (including arrival rates) and imaginary computational power.

The Data-Driven Staffing Objective. The system manager is assumed to have access to historical data in the form of *call arrival epochs* over n previous staffing segments with similar "characteristics" (the term will be explained shortly). All other primitive parameters, including service rates and abandonment rates, are assumed to be known. We index past segments by $l = 1, \ldots, n$ and let $\mathcal{D}_l := \{F_i^l(t):\ 0 \leqslant t \leqslant T,\ i = 1, \ldots, m\}$ be the record of arrivals for all customer classes during segment $l = 1, \ldots, n$. The complete set of data is then given by $\mathfrak{D}_n = \bigcup_{l=1}^n \mathcal{D}_l$. For simplicity, we shall assume that $\mathcal{D}_1, \ldots, \mathcal{D}_n$ are mutually independent and identically distributed. Because F is a simple counting process, one can view information contained in $\mathfrak{D}_n$ as a record of all customer arrival epochs over the past n segments.

A *data-driven solution* to the staffing problem takes as input the known problem primitives and produces a (measurable) mapping from $\mathfrak{D}_n$ to a staffing level $\hat{b}_n \in \mathbb{R}_+^r$. The performance of this staffing prescription is given by $\mathbb{E}[\mathcal{V}(\hat{b}_n)]$, the expectation being over the probability distribution corresponding to the data $\mathfrak{D}_n$. This performance will be compared against the best achievable performance $\mathcal{V}_* = \mathcal{V}(b_*)$ given in (8), which corresponds to the oracle-based solution b_*. Clearly, by optimality of b_*, we have $\mathcal{V}(b) \geqslant \mathcal{V}(b_*)$ for all vectors $b \in \mathbb{R}_+^r$, and it follows that

$$\frac{\mathbb{E}[\mathcal{V}(\hat{b}_n)]}{\mathcal{V}_*} \geqslant 1. \tag{9}$$

The ratio in (9) measures the suboptimality of the expected performance of the data-driven solution on a relative scale, i.e., percentage of excess cost relative to the best achievable performance. We are interested in developing a constructive method for computing the data-driven estimator $\hat{b}_n$ from the problem primitives and historical observations of call arrivals, such that its performance is nearly optimal in the sense that the above optimality gap is suitably small.

4. The Proposed Data-Driven Staffing Method

4.1. Preliminaries

We now proceed with a description of an optimization problem that at a first glance does not appear to be closely related to the original staffing problem given in (7). For any $\lambda \in \mathbb{R}_+^m$ and $b \in \mathbb{R}_+^r$, let $\pi(\lambda, b)$ denote the optimal value of the following linear program (LP): choose $x \in \mathbb{R}_+^\ell$ to

$$\begin{aligned} \text{minimize}\quad & p \cdot (\lambda - Rx) \\ \text{s.t.}\quad & Rx \leqslant \lambda, \quad Ax \leqslant b, \quad x \geqslant 0. \end{aligned} \tag{10}$$

Furthermore, let $\bar{b}$ be the minimizer of

$$V(b) := c \cdot b + \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b)\, dt \right],$$

where the expectation is taken with respect to the distribution of the arrival rate process, $\Lambda = (\Lambda(t)\colon 0 \leqslant t \leqslant T)$. Define the following cumulative distribution function (c.d.f.):

$$G(\lambda) = \frac{1}{T}\int_0^T \mathbb{P}(\Lambda(s) \leqslant \lambda)\, ds, \quad \lambda \in \mathbb{R}_+^m, \tag{11}$$

where $x \leqslant y$ for $x, y \in \mathbb{R}_+^n$ means the inequality holds componentwise. One interprets $G(\lambda)$ as the expected fraction of time (within the planning period $[0, T]$) during which $\Lambda(\cdot) \leqslant \lambda$. The function $V(\cdot)$ can then be expressed as follows:

$$V(b) = c \cdot b + T\int_{\lambda \in \mathbb{R}_+^m} \pi(\lambda, b)\, dG(\lambda) \tag{12}$$

and is easily seen to be convex (see Proposition 1 in Harrison and Zeevi 2005). In addition, if G is atomless, then V is differentiable. Thus, finding the minimizer $\bar{b}$ of $V(\cdot)$ is a convex programming problem. This optimization problem was first formulated in Harrison and Zeevi (2005), where extensive numerical examples indicated that both $V(\bar{b})$ as well as $\mathscr{V}(\bar{b})$ are "close" to $\mathscr{V}_*$.

4.2. The Proposed Approach

The main idea is to use historical call arrival observations to approximate the distribution G given in (11). Consider first a scenario where one has access to n independent replications $\Lambda^1, \ldots, \Lambda^n$ of the multivariate arrival rate process $\Lambda = (\Lambda(t)\colon 0 \leqslant t \leqslant T)$. We can then construct the empirical analogue of the c.d.f. G:

$$G_n(\lambda) = \frac{1}{T}\int_0^T \frac{1}{n}\sum_{l=1}^n \mathbb{I}_{\{\Lambda^l(s) \leqslant \lambda\}}\, ds, \quad \lambda \in \mathbb{R}_+^m, \tag{13}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function and the subscript "n" is used to denote the dependence of this quantity on the "data." Of course, even if an arrival rate process exists, it is not observable, so the above estimate is not computable.

In practice, the system manager has access to data only in the form of *arrival epoches* of customers over n previous segments (each consisting of an interval of length T), which suggests the following strategy: construct estimators of the arrival rates based on the observed arrival epochs, and plug these estimates in (13). A straightforward nonparametric method for estimating the arrival rate is based on counting the number of arrivals over a small window and dividing by the window length. Specifically, let $w > 0$ be a *window size*. Then, for each previous segment $\ell = 1, \ldots, n$, we estimate the arrival rate at time $s \in [0, T]$ as follows:

$$\hat{\Lambda}^l(s) = \frac{F^l(s+w) - F^l(s)}{w}. \tag{14}$$

In §7, we discuss a rule-of-thumb for choosing a "good" window length w, based on asymptotic considerations. There are numerous other approaches for constructing such estimates; for example, regression-based models involving latent variables (see, e.g., Brown et al. 2005). What we focus on here is a general procedure that is independent of problem particulars and can be easily modified to incorporate other estimation methods by appropriately redefining (14).

We now form an empirical counterpart to (11) using the arrival rate estimators described above:

$$\widehat{G}_n(\lambda) = \frac{1}{T}\int_0^T \frac{1}{n}\sum_{l=1}^n \mathbb{I}_{\{\hat{\Lambda}^l(s) \leqslant \lambda\}}\, ds. \tag{15}$$

Based on this empirical distribution, we construct the empirical counterpart of $V(\cdot)$,

$$\widehat{V}_n(b) = c \cdot b + T\int_{\lambda \in \mathbb{R}_+^m} \pi(\lambda, b)\, d\widehat{G}_n(\lambda), \quad b \in \mathbb{R}_+^r.$$

The empirical analogue of $\bar{b}$, the minimizer of $V(\cdot)$, is then given by

$$\hat{b}_n \in \arg\min_{b \in \mathbb{R}_+^r} \widehat{V}_n(b), \tag{16}$$

which is our proposed (data-driven) solution to the staffing problem.

Computation of the Data-Driven Solution. We now describe an algorithm based on gradient descent that can be used to compute the minimizer of $\widehat{V}_n(\cdot)$. To this end, let ζ be a mapping from $\mathbb{R}_+^r \times \mathbb{R}_+^m$ to $\mathbb{R}_+^r$ such that $\zeta(\lambda, b)$ represents the optimal dual variables associated with the constraint $Ax \leqslant b$ in LP (10). The algorithm operates as follows.

Algorithm 1

Step 1. Let $i = 0$ and start with an initial "guess" $\hat{b}_n^{(0)}$.

Step 2. Calculate the expectation of $\zeta(\lambda, \hat{b}_n^{(i)})$ over λ with respect to the empirical distribution $\widehat{G}_n(\cdot)$ and compute

$$\delta b_n^{(i)} = c + \int_{\lambda \in \mathbb{R}_+^m} \zeta(\lambda, \hat{b}_n^{(i)})\, d\widehat{G}_n(\lambda).$$

Step 3. Set $\hat{b}_n^{(i)} + \alpha_i \delta b_n^{(i)} \to \hat{b}_n^{(i+1)}$.

Step 4. Increment counter $i \to i+1$ and repeat Steps 2–4 until $\|\delta b_n^{(i)}\|$ is "close" to zero.

The sequence $\{\alpha_i\}$ in Algorithm 1 represents the step-size that is chosen so that $\delta b_n^{(i)} \to 0$ as $i \to \infty$ (e.g., $\alpha_i = 1/i$), and $\|\cdot\|$ represents the Euclidean norm in $\mathbb{R}_+^r$. (See Birge and Louveaux 1997.) The above algorithm is based on the fact that $c + \int \zeta(\lambda, b)\, d\widehat{G}_n(\lambda)$ is the "gradient" of the convex function $\widehat{V}_n(\cdot)$; implicit here is an interchange argument which can be justified under mild assumptions. Because the empirical measure $\widehat{G}_n$ is not atomless, the function $\widehat{V}_n(\cdot)$ might not be differentiable at all points, in which case $\delta b_n^{(i)}$ is interpreted as a subgradient. This is a potential source of instability for the gradient descent algorithm, and to overcome this impediment one can simply use a "smoothed" version of the distribution $\widehat{G}_n$ in the implementation of the above algorithm.

5. Intuition: Why Should the Method Work?

Intuition and Supporting Logic. The function described in (12) is referred to in Harrison and Zeevi (2005) as the *pointwise stationary fluid model* (PSFM) objective, which involves two levels of approximation. First, all Poisson flows in the system are replaced with their rates, i.e., $N_i^{(1)}(\int_0^t \Lambda_i(s)\,ds) \approx \int_0^t \Lambda_i(s)\,ds$, and this reduces the system dynamics (4) to that of a *fluid model*:

$$\frac{dZ(t)}{dt} = \Lambda(t) - RX(t) - \Gamma Q(t), \quad t \in [0, T]. \tag{17}$$

The next reduction assumes that the system reaches equilibrium "instantly," implying the following instantaneous flow balance equation:

$$\Lambda(t) = RX(t) + \Gamma Q(t), \tag{18}$$

which constitutes a *pointwise stationary* approximation to (17). We can now approximate the cost functional in (5) as follows:

$$\begin{aligned}\mathcal{J}(b, X) &= c \cdot b + \mathbb{E}\left[\int_0^T p \cdot \Gamma Q(s) ds\right] \\ &\approx c \cdot b + \mathbb{E}\left[\int_0^T p \cdot (\Lambda(s) - RX(s))\,ds\right],\end{aligned} \tag{19}$$

where the second line uses (18), and the expectation there is with respect to the distribution of the process Λ. Using the above, we approximate the performance function in (6) as follows:

$$\begin{aligned}\mathcal{V}(b) &= \inf_{X \in \mathcal{X}(b)} \mathcal{J}(b, X) \\ &\overset{(a)}{\approx} c \cdot b + \mathbb{E}\left[\int_0^T \pi(\Lambda(s), b)\,ds\right] \\ &\overset{(b)}{=} c \cdot b + T \int \pi(\lambda, b)\,dG(\lambda) =: V(b),\end{aligned} \tag{20}$$

where (a) uses (19) to approximate to the cost functional $\mathcal{J}(b, X)$ and then minimizes the integrand in (19) at each instant in time, and (b) uses the definition of $V(\cdot)$ in (12). Finally, using the empirical distribution $\widehat{G}_n(\cdot)$ instead of $G(\cdot)$ in (20) and noting that for n large $\widehat{G}_n(\cdot) \approx G(\cdot)$, we expect that

$$\begin{aligned}V(b) &= c \cdot b + T \int \pi(\lambda, b)\,dG(\lambda) \\ &\approx c \cdot b + T \int \pi(\lambda, b)\,d\widehat{G}_n(\lambda) =: \widehat{V}_n(b).\end{aligned}$$

These approximations give rise to the proposed data-driven optimization problem stated in (16).

Error Sources Affecting the Performance of the Data-Driven Solution. To shed light on the approximations involved in the derivation of the data-driven solution and the effect these have on its performance, we introduce the following error decomposition:

$$\begin{aligned}\mathbb{E}[\mathcal{V}(\hat{b}_n)] - \mathcal{V}_* &= \mathbb{E}[\mathcal{V}(\hat{b}_n)] - \mathcal{V}(b_*) \\ &= \mathbb{E}[\mathcal{V}(\hat{b}_n) - V(\hat{b}_n)] + [V(b_*) - \mathcal{V}(b_*)] \\ &\quad + [\mathbb{E}[V(\hat{b}_n)] - V(b_*)] \\ &\leqslant \underbrace{\mathbb{E}[\mathcal{V}(\hat{b}_n) - V(\hat{b}_n)]}_{\text{approx. error I}} + \underbrace{[V(b_*) - \mathcal{V}(b_*)]}_{\text{approx. error II}} \\ &\quad + \underbrace{[\mathbb{E}[V(\hat{b}_n)] - V(\bar{b})]}_{\text{estimation error}},\end{aligned} \tag{21}$$

where the inequality follows because $\bar{b}$ is the minimizer of $V(\cdot)$, and all expectations are taken with respect to the distribution of the database $\mathfrak{D}_n$.

To better understand (21), and the terminology used to describe the various error terms, recall that the following two simplifications are made in deriving the data-driven solution: First, the original objective was replaced by a fluid-like (PSFM) approximation; and second, the actual distribution of the arrival rates was replaced by its empirical counterpart given by (15). These steps result in two errors: an *approximation error* and an *estimation* (*data-related*) *error*, respectively. In (21), the first and the second terms on the right-hand side measure deviations between the original objective $\mathcal{V}(\cdot)$ and the PSFM analogue $V(\cdot)$ for a fixed staffing level. The last term captures the error introduced by using the data-driven solution $\hat{b}_n$ as opposed to the V-optimal value $\bar{b}$.

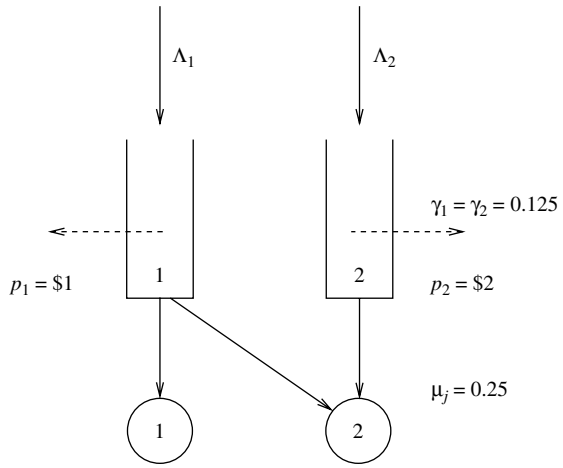
If the system is such that the volume of arrivals is "large" and customers are processed "quickly," we expect that $\mathcal{V}(\cdot) \approx V(\cdot)$. If in addition the number of historical observations n is large, then one can expect $\widehat{V}_n(\cdot) \approx V(\cdot)$ and thus $\hat{b}_n \approx \bar{b}$. Under these conditions, all error terms on the right-hand side of (21) should be small and hence the performance of the data-driven solution should be close to the best achievable performance, namely, $\mathbb{E}[\mathcal{V}(\hat{b}_n)] \approx \mathcal{V}_*$. The next two sections investigate the validity of this logic both via numerical experiments and theoretical analysis.

6. Accuracy of the Proposed Method: An Illustrative Numerical Example

In this section, we study numerically the efficacy of the proposed data-driven staffing method as a function of arrival rate variability and the number of data points (corresponding to past call arrival observations).

Description of the Example and Numerical Experiments. The example we consider has two customer classes ($m = 2$), which are served by two server pools ($r = 2$). There are three processing activities ($\ell = 3$): servers in pool 1 can serve only class 1 customers (activity 1),

Figure 2. Schematic of the two-class/two-pool system used for the numerical study.



and servers in pool 2 are cross-trained and can serve both class 1 and class 2 customers (activities 2 and 3, respectively). All the services are exponentially distributed with mean equal to four minutes, that is, $\mu_j = 1/4$ customers per minute for $j = 1, 2, 3$. (For simplicity, it is assumed that all services can be preempted.) Customers of both classes abandon at rate $\gamma_1 = \gamma_2 = 1/8$ customers per minute. The nominal abandonment penalties for class 1 and class 2 are $p_1 = \$1$ per customer and $p_2 = \$2$ per customer, respectively. (In what follows, we also discuss sensitivities with respect to these parameters.) A schematic of the system together with various parameter values is depicted in Figure 2.

We focus on a two-hour staffing segment over which staffing is kept constant, so $T = 120$ minutes. The cost of a server in pool 1 for the two-hour period is \$15, and \$30 in pool 2 where the servers are cross-trained (these servers can process requests from either customer class). Customers from classes 1 and 2 arrive according to a Poisson process with random rates Λ_1 and Λ_2, which will be specified shortly. The cost structure described above has been judiciously chosen so that the optimal control that minimizes the cost functional $\mathcal{J}(b, X)$ can be easily identified. In particular, this control has all servers in pool 2 giving strict priority to class 2 customers. With the control in place, one can compute the performance function and proceed to optimize it as in (7) by means of simulation. It is probably worth noting that in general it is not possible to identify the optimal control, and hence it is also impossible to compute the optimal staffing level and best achievable system performance (even via simulation).

Consistent with the system model that was described earlier in the paper, arrivals of calls/customers occur according to a Poisson process whose rate, for each class, is given by $\Lambda_1 = 2\Lambda_2 = \max\{0, Y\}$, where Y is a random variable whose distribution is specified below. That is, we assume that arrivals are generated by a homogenous Poisson process with an intensity that is constant but random. We perform several experiments that illustrate the (in)sensitivity of our method to the characteristics of this intensity. In particular, we consider three distributions for Y, with two scenarios considered in each case: one where Y has relatively low variance (LV) and one with high variance (HV).

- **Normal distribution:** Y has mean 5 and standard deviation 0.5 (LV) and 1.5 (HV).
- **Uniform distribution:** $Y \sim U[4.1, 5.9]$ (LV) and $Y \sim U[2.4, 7.6]$ (HV).
- **Discrete distribution:** Y puts equal mass on two points: $\{4.5, 5.5\}$ (LV) and $\{3.5, 5.5\}$ (HV).

The parameters are chosen so as to maintain the same mean and variance for all distributions (thus keeping the coefficient of variation constant). For each of the three distributions outlined above, we simulate the system under various staffing levels and estimate the value of the performance function $\mathcal{V}(\cdot)$ for the low and high variability scenarios, respectively. (The number of simulation runs for each staffing level is 1,000.) Based on this, we obtain a simulation-based estimate of the optimal staffing levels b^* and the value function $\mathcal{V}^* := \mathcal{V}(b^*)$.

We study properties of the data-driven solution for two sizes of historical data: a "small" sample consisting of arrival epochs that have been recorded over $n = 5$ past staffing segments, and a "large" sample consisting of arrival epochs for $n = 100$ past segments. For each arrival rate scenario, we construct $N = 5{,}000$ replication of the data set, and for each replication we use the data-driven staffing algorithm described in §4 to obtain the recommended staffing levels $\hat{b}_n$ as in (16). In doing so, we use a window length $w = 20$ minutes to estimate rates from arrival epochs. (We have found this to be a reasonable choice in terms of performance.) Computing this estimate involves solving the LP with recourse that was described in §4. This operation took a few seconds or less in all scenarios considered on a standard desktop computer running MATLAB. The estimates of the performance function $\mathcal{V}(\cdot)$, obtained in the previous simulation experiments for both the LV and HV scenarios, are then used to compute an estimate of the performance of the data-driven solution. Specifically, we estimate $\mathbb{E}\mathcal{V}(\hat{b}_n)$ by taking an average of $\mathcal{V}(\hat{b}_n)$ over the $N = 5{,}000$ replications.

Discussion of the Main Results. The main results are summarized in Tables 1 and 2. Table 1 details the *performance* of the data-driven staffing method and the loss relative to the optimal performance $\mathcal{V}^*$. Table 2 depicts the properties (mean and variance) of the data-driven estimator itself, and contrasts that with the optimal staffing level b^*. The results reported encompass all cases outlined above, and the 95% confidence intervals are given in parentheses as relative percentage values. The results given in Table 1 illustrate that the data-driven solution achieves near-optimal performance in all cases, in particular, the ratio $\mathbb{E}[\mathcal{V}(\hat{b}_n)]/\mathcal{V}_*$ is almost always very close to one. Moving now to Table 2, we observe that the mean of the

Table 1. Performance of the data-driven staffing method.

		Optimal	Large sample		Small sample	
		$\mathcal{V}_* = \mathcal{V}(b_*)$	$\mathbb{E}[\mathcal{V}(\hat{b}_{100})]$	$\mathbb{E}[\mathcal{V}(\hat{b}_{100})]/\mathcal{V}_*$	$\mathbb{E}[\mathcal{V}(\hat{b}_5)]$	$\mathbb{E}[\mathcal{V}(\hat{b}_5)]/\mathcal{V}_*$
LV	Normal	690.4	696.9 [±0.03%]	1.01	698.7 [±0.2%]	1.01
	Uniform	690.5	701.3 [±0.01%]	1.02	701.9 [±0.03%]	1.02
	Two point	690.3	700.8 [±0.01%]	1.02	701.6 [±0.03%]	1.02
HV	Normal	747.7	754.7 [±0.1%]	1.01	771.1 [±0.2%]	1.03
	Uniform	747.8	763.1 [±0.03%]	1.02	781.7 [±0.08%]	1.05
	Two point	749.5	774.0 [±0.05%]	1.03	792.3 [±0.08%]	1.06

Notes. This table reports the performance of the data-driven staffing solution $\mathcal{V}(\hat{b}_n)$ for small and large sample sizes, for low variability (LV) and high variability (HV) scenarios and the various arrival rate distributions (normal, uniform, and discrete).

data-driven solution is close to the value of the optimal staffing level in all cases. The variance of the data-driven estimator is much higher, as expected, when the variability of arrival rates is large and increases when the sample size is small. (We also performed sensitivity analysis relative to other input parameters, for example, penalties p and costs c, and we observed results similar to those reported above over a wide range of values; for brevity these results are not reported herein.)

The behavior of the estimator observed in Table 2 makes the results reported in Table 1 perhaps even more surprising. In particular, one would expect that as the estimator's accuracy degrades, so will its performance. The reason this is not the case hinges on the effects of arrival rate variability on the properties of the performance function $\mathcal{V}(\cdot)$. Specifically, when the variability of the arrival rates increases, the performance function becomes relatively flat in the vicinity of its minimum; see also Harrison and Zeevi (2005). Consequently, the performance of the system does not degrade in a significant manner even when $\hat{b}_n$ is not that "close" to b_*. Of course, when the variability of the arrival rates is low, then the performance function is not as flat in the vicinity of the optimal staffing vector b_*. Roughly speaking, the reason one still observes good performance in this case, even for small sample sizes, is that the low variability implies that a small number of data points suffices for accurately estimating the arrival rate distribution. (That is, in this case there is more statistical "information" in each sample.) Due to these two contradicting effects, our data-driven estimator of the optimal staffing level performs well in all scenarios.

Sample Size Effects: A Hint Toward Asymptotic Theory. The purpose of the following experiment is to indicate how the performance of the data-driven staffing level is affected by the number data points n. We consider the same system model introduced earlier in this section, but with the following parameters. Service rates are $\mu_j = 1$ call per minute for each activity, and abandonment rates are $\gamma_i = 0.5$ call per minute in each class. The time horizon is again taken to be 120 minutes, and the cost of servers in each pool are $(c_1, c_2) = (30, 60)$ over that time interval. The abandonment penalties are taken to be $(p_1, p_2) = (1, 2)$.

The arrival rate in this example is such that Λ_1, measuring number of arrivals per minute, is drawn from a 15-point distribution placing equal mass on each of the points $50, 55, 60, \ldots, 120$ and $\Lambda_2 = 0.5\Lambda_1$. The number of runs to determine the function $\mathcal{V}(\cdot)$ was taken to be 405. We compute the value function for staffing levels in the range of 80–120 in pool 1 and 40–60 in pool 2. (Confidence intervals are ±2%–3% in all experiments.) The simulation-based estimate of the optimal staffing level is found to be $b^* = (100, 53)$, and the optimal expected total cost is $\mathcal{V}^* = 6{,}887$.

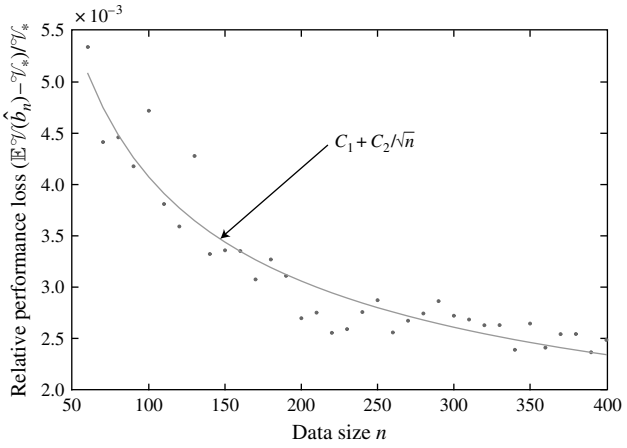
We consider samples sizes that contain data for $n = 50, 60, \ldots, 400$ past segments, each such segment comprising arrival records over the interval $[0, T]$. We generate

Table 2. Properties of the data-driven estimator.

		Optimal	Large sample		Small sample	
		b_*	$\mathbb{E}[\hat{b}_{100}]$	$\mathrm{Var}[\hat{b}_{100}]$	$\mathbb{E}[\hat{b}_5]$	$\mathrm{Var}[\hat{b}_5]$
LV	Normal	(17, 10)	(19.5, 10.0)	(0.25, 0.23)	(19.4, 10.3)	(1.3, 0.4)
	Uniform	(17, 11)	(19.2, 10.7)	(0.16, 0.20)	(19.3, 10.6)	(1.6, 0.5)
	Two point	(18, 10)	(19.2, 10.6)	(0.2, 0.2)	(19.3, 10.6)	(1.6, 0.4)
HV	Normal	(16, 10)	(19.4, 10.3)	(0.58, 0.25)	(19.4, 10.4)	(9.7, 2.4)
	Uniform	(18, 12)	(18.3, 10.3)	(13.1, 1.2)	(19.5, 10.4)	(20.5, 5.3)
	Two point	(15, 11)	(18.9, 10.3)	(13.12, 1.24)	(19.5, 10.4)	(20.6, 5.3)

Notes. Mean and variance of the data-driven staffing estimator $\hat{b}_n$ for small and large sample sizes, for low variability (LV) and high variability (HV) scenarios, and the three distributions for the arrival rates (normal, uniform, and discrete). Here b^* denotes the optimal staffing level, and both $\hat{b}_n$ and b^* give the staffing level in each of the two server pools.

Figure 3. The dots represent performance of the data-driven solution ($\mathbb{E}\mathscr{V}(\hat{b}_n)$) relative to the optimal performance ($\mathscr{V}_*$), as a function of the sample size (n).



Note. The superimposed curve is obtained by regressing against the function $C_1 + C_2/\sqrt{n}$.

$N = 100$ such replications for each value of n, and compute the data-driven estimator $\hat{b}_n$. We then estimate the performance of the data-driven solution $\mathbb{E}\mathscr{V}(\hat{b}_n)$ by averaging over the N replications. Figure 3 displays the optimality gap (on a relative scale) as a function of n. That is, it plots $\mathbb{E}[\mathscr{V}_*(\hat{b}_n)]/\mathscr{V}_* - 1$. From the results in Tables 1 and 2, we expect this difference to be small, but our interest here is in assessing the rate at which it decreases as the sample size grows large. The curve in Figure 3 is obtained by regressing the function $g(n) = C_1 + C_2/\sqrt{n}$ against the observed performance of $\hat{b}_n$ for each sample size. More precisely, it is derived by regressing the relative loss in performance, $\mathbb{E}\mathscr{V}(\hat{b}_n)/\mathscr{V}_* - 1$, against n. The fit is evident, and the power of $n^{-1/2}$ is found to be statistically significant (the values of the constants are $C_1 = 6.07 \times 10^{-4}$ and $C_2 = 3.4 \times 10^{-2}$). One again sees good performance of the data-driven estimator even for small sample sizes (this is also reflected in the moderate values of the constant C_i in the fitted curve). In the next section, we will show that the empirical observations drawn from Figure 3 can be formalized in a precise mathematical sense, articulating the rate at which the data-dependent error associated with the performance data-driven staffing level converges to zero.

7. Accuracy of the Proposed Method: Asymptotic Analysis

7.1. Preliminaries

To evaluate the performance of our data-driven solution method, we introduce a scaling of model parameters that characterizes systems with high volume, rapid turnover and many servers. Let $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ be a superlinear function, i.e., $x^{-1}f(x) \to \infty$ as $x \to \infty$, and for each positive integer κ put $R^\kappa = \kappa R$, $\Gamma^\kappa = \kappa\Gamma$, and $\Lambda^\kappa(\cdot) = f(\kappa)\Lambda(\cdot)$. We also scale the cost of servers by a factor κ, $c^\kappa = \kappa c$, because they now process work at κ-times the original rate. (Note that because the arrivals are scaled up by a superlinear function $f(\cdot)$ while the service rates are scaled up only linearly, the number of servers required for nominal operation should increase without bound.)

We will indicate the dependence of various state processes, cost functionals, etc., on the above scaled parameters by using the superscript "κ." In particular, let Z^κ and Q^κ denote the headcount and queue-length processes that jointly solve (2)–(4) with the input parameters R^κ, Γ^κ, and Λ^κ and with control that belongs to $\mathscr{U}^\kappa(b)$, the corresponding admissible control set. Let $\mathscr{J}^\kappa(b, X)$ be the cost functional for the κ-scaled system for a staffing vector $b \in \mathbb{R}^r_+$ and control $X \in \mathscr{U}^\kappa(b)$, and let the κ-scaled performance function be $\mathscr{V}^\kappa(b) = \inf_{X\in\mathscr{U}^\kappa(b)} \mathscr{J}^\kappa(b, X)$. Similarly, let $\mathscr{V}^\kappa_*$ denote the *best achievable performance* for the κ-scaled system

$$\mathscr{V}^\kappa_* = \inf_{b\in\mathbb{R}^r_+} \mathscr{V}^\kappa(b). \tag{22}$$

In light of Proposition 1, there exists an optimal staffing level b^κ_* that achieves the infimum in (22) and thus $\mathscr{V}^\kappa_* = \mathscr{V}^\kappa(b^\kappa_*)$.

We also need to define the counterparts of G, V, and π for the κ-scaled system. Let $G^\kappa(\lambda)$ be as in (11) with arrival rate Λ replaced by Λ^κ, namely, for $\lambda \in \mathbb{R}^m_+$ set $G^\kappa(\lambda) = T^{-1}\int_0^T \mathbb{P}(\Lambda^\kappa(s) \leqslant \lambda)\, ds$. Put

$$V^\kappa(b) = c^\kappa \cdot b + T\int_{\lambda\in\mathbb{R}^m_+} \pi^\kappa(\lambda, b)\, dG^\kappa(\lambda), \tag{23}$$

where $\pi^\kappa(\lambda, b)$ is the value of the LP (10), where the matrix R is replaced by the scaled matrix R^κ.

7.2. Performance Bounds

To derive our main results, we require the following technical assumption with regard to the arrival rate processes.

ASSUMPTION 1. *The path of $\Lambda_i(\cdot)$ is nonnegative, uniformly bounded, and Lipschitz continuous (a.s.) over $[0, T]$ for $i = 1, \ldots, m$. Furthermore, the Lipschitz constant is uniformly bounded and the number of sign changes in the derivative of $\Lambda_i(\cdot)$, $i = 1, \ldots, m$ (where it exists) is uniformly bounded.*

This assumption ensures that paths of the arrival rate process are suitably "smooth" and do not oscillate too much. (One expects these conditions to hold in any reasonable practical setting.)

We first analyze an idealized scenario where data consists of "observed" arrival rates. This assumption eliminates any error associated with inferring the arrival rates from arrival epochs and hence will lead to an optimistic performance bound for our data-driven solution, irrespective of the method used to construct arrival rate estimates. Subsequently, we will analyze the performance of our proposed data-driven method, which uses a window-based scheme to estimate arrival rates.

Performance Bounds: The Case of "Observed" Arrival Rates. Using the prescription in §4, construct the empirical analogue of the c.d.f. $G^\kappa(\cdot)$,

$$\check{G}_n^\kappa(\lambda) = \frac{1}{T}\int_0^T \frac{1}{n}\sum_{l=1}^{n} \mathbb{I}_{\{\Lambda^{\kappa,l}(s)\leqslant\lambda\}}\, ds, \quad \lambda \in \mathbb{R}_+^m,$$

and put

$$\check{b}_n^\kappa \in \arg\min_{b\in\mathbb{R}_+^r}\left\{c^\kappa \cdot b + T\int_{\lambda\in\mathbb{R}_+^m} \pi^\kappa(\lambda, b)\, d\check{G}_n^\kappa(\lambda)\right\}. \tag{24}$$

The estimator $\check{b}_n^\kappa$ is clearly an idealized quantity because it is constructed based on a sample that consists of "*observed*" *rates*; the notation used here is meant to distinguish this estimator from the one that is constructed based on past *arrival epochs* data $\hat{b}_n$, which is discussed in what follows.

The total cost of running the system with this staffing level is given by $\mathscr{V}^\kappa(\check{b}_n^\kappa)$. Let

$$\bar{b}^\kappa \in \arg\min_{b\in\mathbb{R}_+^n} V^\kappa(b). \tag{25}$$

Note that $V_n^\kappa(0) \leqslant Mf(\kappa)$ and $V^\kappa(0) < Mf(\kappa)$, where M is a finite constant such that $\mathbb{E}[\int_0^T p\cdot\Lambda(s)\,ds] \leqslant M$. Thus, the minimization in (24) and (25) can be taken over the compact convex set $\mathscr{B}^\kappa = \{b\colon c^\kappa \cdot b \leqslant Mf(\kappa)\}$.

Using Proposition 1, there exists a staffing level b_*^κ such that $\mathscr{V}^\kappa(b_*^\kappa) = \mathscr{V}_*^\kappa$ for each $\kappa \in \mathbb{N}$. Using the error decomposition in §5, we get the following counterpart of (21) that bounds the relative error between the performance of the staffing level $\check{b}_n^\kappa$ and the optimal performance:

$$\frac{\mathbb{E}[\mathscr{V}^\kappa(\check{b}_n^\kappa)]}{\mathscr{V}_*^\kappa} \leqslant 1 + \frac{\mathbb{E}[\mathscr{V}^\kappa(\check{b}_n^\kappa) - V^\kappa(\check{b}_n^\kappa)]}{\mathscr{V}_*^\kappa} + \frac{[V^\kappa(b_*^\kappa) - \mathscr{V}^\kappa(b_*^\kappa)]}{\mathscr{V}_*^\kappa} + \frac{[\mathbb{E}[V^\kappa(\check{b}_n^\kappa)] - V^\kappa(\bar{b}^\kappa)]}{\mathscr{V}_*^\kappa}. \tag{26}$$

We next state three theorems that characterizes the asymptotic behavior of each error term on the right-hand side of (26). For this purpose, it will be useful to introduce the following notation: for a real-valued sequence $\{a^\kappa\}$ and a positive real-valued sequence $\{d^\kappa\}$, we say that $a^\kappa = \mathscr{O}(d^\kappa)$ as $\kappa\to\infty$ if $\limsup_{\kappa\to\infty} a^\kappa/d^\kappa < \infty$.

Theorem 1 (Approximation Error I). *Let Assumption* 1 *hold. Then,*

$$\sup_{b\in\mathscr{B}^\kappa} \frac{(\mathscr{V}^\kappa(b) - V^\kappa(b))}{\mathscr{V}_*^\kappa} \to 0 \quad \textit{as } \kappa\to\infty.$$

If in addition $(\Lambda(t)\colon 0\leqslant t\leqslant T)$ *is time homogenous, then*

$$\sup_{b\in\mathscr{B}^\kappa} \frac{\mathscr{V}^\kappa(b) - V^\kappa(b)}{\mathscr{V}_*^\kappa} = \mathscr{O}\left(\sqrt{\frac{\kappa}{f(\kappa)}}\right) \quad \textit{as } \kappa\to\infty.$$

This result ensures that $\mathscr{V}^\kappa(\cdot)$ is uniformly "close" to its fluid analogue $V^\kappa(\cdot)$ in the sense of relative error. With regard to the second approximation error in (26), we have the following theorem.

Theorem 2 (Approximation Error II). *Let Assumption* 1 *hold. Then,*

$$\frac{V^\kappa(b_*^\kappa) - \mathscr{V}^\kappa(b_*^\kappa)}{\mathscr{V}_*^\kappa} = \mathscr{O}\left(\frac{1}{\sqrt{f(\kappa)}}\right) + \mathscr{O}\left(\frac{1}{\kappa}\right) \quad \textit{as } \kappa\to\infty.$$

We observe that Theorem 2 has two terms characterizing the error. The first term, which is $\mathscr{O}(1/\sqrt{f(\kappa)})$, corresponds to the "fluid approximation," based on which we replace all Poisson flows in the system by their rates. This error term stems from the functional central limit theorem, which in the case of Poisson processes states that the deviations from the mean path are of order square root of the mean. The term $\mathscr{O}(1/\kappa)$ corresponds to the *pointwise stationary approximation*, which was used in the reduction of the fluid dynamics equation (17) to the instantaneous flow balance equation (18). Roughly speaking, the fluid model converges to its steady-state exponentially fast with parameter κ (because both service and abandonment rates are scaled by κ), hence the relaxation time should be $\mathscr{O}(1/\kappa)$.

Finally, with regard to the third term on the right-hand side of (26), we have the following theorem.

Theorem 3 (Estimation Error). *Let Assumption* 1 *hold. Then, for all* $\kappa = 1, 2, \ldots,$

$$\frac{\mathbb{E}[V^\kappa(\check{b}_n^\kappa)] - V^\kappa(\bar{b}^\kappa)}{\mathscr{V}_*^\kappa} = \mathscr{O}\left(\frac{1}{\sqrt{n}}\right) \quad \textit{as } n\to\infty.$$

The term appearing in the right-hand side above is in fact uniform in κ. Roughly speaking, the above theorem states that having a finite number of past observations results in deviations of $\mathscr{O}(1/\sqrt{n})$ between the fluid (PSFM) objective values $V(\bar{b}^\kappa)$ and $V(\check{b}_n^\kappa)$.

Combining Theorems 1–3, we see that the performance of $\check{b}_n^\kappa$ relative to the best achievable performance is given by

$$\frac{\mathbb{E}[\mathscr{V}^\kappa(\check{b}_n^\kappa)]}{\mathscr{V}_*^\kappa} = 1 + \mathscr{O}\left(\frac{1}{\kappa}\right) + \mathscr{O}\left(\sqrt{\frac{\kappa}{f(\kappa)}}\right) + \mathscr{O}\left(\frac{1}{\sqrt{n}}\right) \quad \text{as } \kappa\to\infty \text{ and } n\to\infty, \tag{27}$$

where for a real-valued double-indexed sequence $\{a_n^\kappa\colon n = 1, 2, \ldots \text{ and } \kappa = 1, 2, \ldots\}$ and positive real-valued functions $h_1(\cdot)$ and $h_2(\cdot)$, we say that $a_n^\kappa = \mathscr{O}(h_1(n)) + \mathscr{O}(h_2(\kappa))$ if there exists a finite constant C such that $a_n^\kappa \leqslant C(h_1(n) + h_2(\kappa))$ for all $\kappa = 1, 2, \ldots$ and $n = 1, 2, \ldots$.

Performance Bounds for the Data-Driven Solution. We now extend the results derived above to cover the case where the observations consist of past arrival epochs in each class. Let

$$\mathfrak{D}_n^\kappa = \{F_i^{\kappa,l} \text{ for } l = 1, \ldots, n \text{ and } i = 1, \ldots, m\},$$

where $F_i^{\kappa,l} = (F_i^{\kappa,l}(t)\colon 0\leqslant t\leqslant T)$ is the arrival process for customer class i in the κ-scaled system over the lth past segment. We use the window-based estimator defined in (14) to "convert" arrival epochs to arrival rate estimates.

Let $(\hat{\Lambda}^{\kappa,l}(t)\colon 0 \leqslant t \leqslant T)$ be the estimated arrival rate process that is generated as follows:

$$\hat{\Lambda}^{\kappa}(t) = \frac{1}{w^{\kappa}}(F^{\kappa,l}(t) - F^{\kappa,l}(t - w^{\kappa})) \tag{28}$$

for $t \in [w^{\kappa}, T]$. Here $w^{\kappa} := g(\kappa)^{-1}$ represents the length of a κ-scaled *sliding window* over which arrivals are counted, where $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function. Recall that $F^{\kappa}(t) = (F_1^{\kappa}(t), \ldots, F_m^{\kappa}(t))$ is the vector of cumulative arrivals up until time t in each customer class.

From these estimators of the arrival rates, we construct the empirical c.d.f.

$$\widehat{G}_n^{\kappa}(\lambda) = \frac{1}{T}\int_0^T \frac{1}{n}\sum_{l=1}^{n} \mathbb{I}_{\{\hat{\Lambda}^{\kappa,l}(s)\leqslant\lambda\}}\, ds \qquad \text{for } \kappa = 1, 2, \ldots \text{ and } n = 1, 2, \ldots.$$

For each n and κ, let the data-driven solution be given by

$$\hat{b}_n^{\kappa} \in \arg\min\left\{ c^{\kappa}\cdot b + \int_{\lambda\in\mathbb{R}_+^m} \pi^{\kappa}(\lambda, b)\, d\widehat{G}_n^{\kappa}(\lambda)\right\}, \tag{29}$$

where $\pi^{\kappa}(\cdot,\cdot)$ is the value of the LP (10), where the matrix R is replaced by the scaled matrix R^{κ}.

The performance of $\hat{b}_n^{\kappa}$ is again affected by approximation errors identical to those characterized in the previous section. However, here the data-related error consists of two components: an *observation error* and an *estimation error*. The former stems from estimating arrival rates from arrival epoches, and the latter is identical to that stated in Theorem 3. For the specific arrival rate estimator given in (28), the observation error shrinks to zero as the arrival rates grow large and the window size shrinks accordingly. The counterpart of Theorem 3 that characterizes both the observation error and estimation error is stated next.

Theorem 4 (Estimation and Observation Error). *Let Assumption* 1 *hold. Then,*

$$\frac{\mathbb{E}[V^{\kappa}(\hat{b}_n^{\kappa})] - V^{\kappa}(\bar{b}^{\kappa})}{\mathscr{V}_*^{\kappa}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{1}{g(\kappa)}\right) + \mathcal{O}\left(\frac{g(\kappa)}{\sqrt{f(\kappa)}}\right) \quad \text{as } \kappa\to\infty \text{ and } n\to\infty. \tag{30}$$

Discussion. Note that the first term on the right-hand side of (30) is identical to that in Theorem 3 and corresponds to the estimation error; the last two terms correspond to the observation error, and both vanish if $g(\kappa)\to\infty$ and $g(\kappa)/\sqrt{f(\kappa)}\to 0$ as $\kappa\to\infty$. Roughly speaking, the window size should be large enough to have sufficiently many observations fall within its support and simultaneously not too large so that the arrival rate can be estimated consistently over the entire interval of interest. In particular, the order of the sum of the last two terms on the right-hand side of (30) is minimized by setting $g(\kappa) = \mathcal{O}(f(\kappa)^{1/4})$, which suggests, as a rule-of-thumb, to set the window size proportional to the inverse of the fourth root of the maximal arrival rate. For example, this suggests that for staffing purposes if the arrival rate is around 200 customers per hour, then an estimation window whose length is around 15 minutes is needed.

With this choice of windowing scheme, it is possible, by examining the order of magnitude of quantities in (27) and (30), to determine which error term will dominate in various regimes. Specifically, if the rate of call arrivals is close in magnitude to the rate of turnover in the system ($f(\kappa) \ll \kappa^2$), then the fluid-model error articulated in Theorem 1 dominates. On the other hand, if the arrival rate to the system is moderately larger than the rate of turnover in the system ($\kappa^2 \ll f(\kappa) \ll \kappa^4$), then the windowing scheme introduces the dominant error term. Finally, if the arrival rate to the system is significantly larger than the rate of turnover in the system ($f(\kappa) \gg \kappa^4$), then the error stemming from the pointwise stationary approximation dominates. A simple rule of thumb that emerges is that the method we propose should produce adequate performance when arrival rates are large and significantly larger than the turnover times.

8. Discussion and Qualitative Insights

On the Choice of Windowing Scheme. The analysis in §7, particularly the discussion following Theorem 4, suggests that a "good" choice of window size for estimating arrival rates from past arrival epochs should be such that it is significantly larger then the characteristic interarrival time. The logic for choosing this window size is that the arrival rate should not vary by a large amount within the window, and still a large number of arrivals should occur within a window. Furthermore, when volumes of incoming calls are high, recorded arrival epochs should be sufficient to produce "good" estimates of the arrival rate. In particular, in such environments the observation error is expected to be small. However, if the arrival rates are of moderate size, then the observation error can be significant. To efficiently implement the proposed data-driven staffing method, one can resort to more explicit modeling of the arrival rates—for example, using latent variables that could include time of day, day of the week, seasonality effects, etc; see Brown et al. (2005) for an example of this type of approach.

Asymptotic Optimality of the Data-Driven Staffing Method. From Theorems 1, 2, and 4, we see that if both n and κ increase without bound and $g(\kappa)$ is taken such that $g(\kappa)\to\infty$ and $g(\kappa)/\sqrt{f(\kappa)}\to 0$ as $\kappa\to\infty$, the performance of the data-driven solution approaches the best achievable performance $\mathscr{V}_*$. That is, we have proved that the data-driven solution $\hat{b}_n^{\kappa}$ is *asymptotically optimal* in the sense that

$$\frac{\mathbb{E}[\mathscr{V}(\hat{b}_n^{\kappa})]}{\mathscr{V}_*^{\kappa}} \to 1 \quad \text{as } \kappa, n\to\infty. \tag{31}$$

(More precisely, there exists a sequence n_κ such that $n_\kappa \to \infty$ as $\kappa \to \infty$ and the above ratio approaches one.) The example studied in §6 illustrates that the asymptotics in (31) might be observed in practice even for moderate values of system parameters and a relatively small number of past segment observations.

On the Control Associated with the Data-Driven Solution. Our analysis of the staffing problem essentially "assumes away" the issue of call routing logic as the performance of any staffing rule (whether optimal or not) was evaluated by pairing it with its associated optimal control. This approach allows us to discuss the staffing problem in isolation but leaves open an obvious question: What control should be implemented in conjunction with the proposed data-driven staffing solution to guarantee "good performance"? A careful look at the proofs of the main results suggests one possible answer to that question. In particular, the proofs use a construction of a control that is similar to the one introduced in Bassamboo et al. (2006) and involves a repeated solution of a linear program. This identifies a family of easily implementable controls which are "good" candidates for being paired with the output of our proposed data-driven staffing method.

9. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgments

The authors thank the associate editor and two referees for their helpful comments that improved the presentation and structuring of the paper. This research was supported in part by NSF grant DMI-0447652.

References

Avramidis, A. N., A. Deslauriers, P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Sci.* **50** 896–908.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2005. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems, Theory and Appl.* (*QUESTA*) **51** 249–285.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54** 419–435.

Birge, J. R., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer-Verlag, New York.

Borst, S., A. Mandelbaum, M. I. Reimann. 2004. Dimensioning large call centers. *Oper. Res.* **52** 17–34.

Bremaud, P. 1981. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, New York.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Cezik, M. T., P. L'Ecuyer. 2008. Staffing multiskill call centers via linear programming and simulation. *Management Sci.* **54**(2) 310–323.

Feldman, Z., A. Mandelbaum, B. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.

Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* **52** 243–257.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7** 20–36.

Mandelbaum, A., S. Zeltyn. 2008. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. Working paper, Technion—Israel Institute of Technology, Haifa, Israel.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2004. Service system planning in the presence of a random arrival rate. Technical report, Cornell University, Ithaca, NY.

van der Vaart, A. W., J. A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer, New York.

van Ryzin, G. J., J. McGill. 2000. Revenue management without forecasting of optimization: An adaptive algorithm for determining airline seat protection levels. *Management Sci.* **46** 760–775.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* **7** 276–294.