Staffing Call Centers with Uncertain Arrival Rates and Co-sourcing

Yaşar Levent Koçağa
Sy Syms School of Business, Yeshiva University, Belfer Hall 403/A, New York, NY 10033,
kocaga@yu.edu

Mor Armony
Stern School of Business, New York University, KMC 8–62, New York, NY, 10012,
marmony@stern.nyu.edu

Amy R. Ward
Marshall School of Business, University of Southern California, Bridge Hall 401H, Los Angeles,
CA, 90089, amyward@marshall.usc.edu
Staffing Call Centers with Uncertain Arrival Rates and Co-sourcing

In a call center, staffing decisions must be made before the call arrival rate is known with certainty. Once the arrival rate becomes known, the call center may be over-staffed, in which case staff are being paid to be idle, or under-staffed, in which case many callers hang-up in the face of long wait times. Firms that have chosen to keep their call center operations in-house can mitigate this problem by co-sourcing; that is, by sometimes outsourcing calls. Then, the required staffing $N$ depends on how the firm chooses which calls to outsource in real-time, after the arrival rate realizes and the call center operates as a $M/M/N + M$ queue with an outsourcing option. Our objective is to find a joint policy for staffing and call outsourcing that minimizes the long run average cost of this two-stage stochastic program when there is a linear staffing cost per unit time and linear costs associated with abandonments and outsourcing.

We propose a policy that uses a square-root safety staffing rule, and outsources calls in accordance with a threshold rule that characterizes when the system is “too crowded”. Analytically, we establish that our proposed policy is asymptotically optimal, as the mean arrival rate becomes large, when the level of uncertainty in the arrival rate is of the same order as the inherent system fluctuations in the number of waiting customers for a known arrival rate. Through an extensive numerical study, we establish that our policy is extremely robust. In particular, our policy performs remarkably well over a wide range of parameters, and far beyond where it is proved to be asymptotically optimal.

Keywords: Call center operations; co-sourcing; staffing; overflow routing; parameter uncertainty.

History: Received: July 2013; Accepted: September 2014, by Michael Pinedo, after 2 revisions.

1. Introduction

Call centers have become ubiquitous in business. Today, every Fortune 500 company has at least one call center, and the average Fortune 500 company employs 4500 call center agents (who may be distributed across more than one site) (Gilson and Khandelwal, 2005). For many companies, the call center is a primary point-of-contact with their customers. Hence a well-run call center promotes good customer relations, and a poorly-managed one hurts them. But call center management is difficult.

A call center manager faces the classical operational challenge of determining appropriate staffing levels throughout the day and week in order to meet a random and time-varying call volume. This is extremely difficult especially when call arrival rate is itself random, as was empirically shown in Brown et al. (2005) and Maman (2009), among others. When staffing levels are too low, customers are put on hold, and many hang up in frustration while waiting for an agent to take their call. But when staffing levels are too high, the call center manager ends up paying staff to be idle.
One option in managing this uncertainty is for a company to outsource its call center operations. Then, the challenges of call center management can be handled by a vendor firm whose primary focus is call center operations. That vendor can pool demand amongst various companies, thereby lowering variability, which should allow for more accurate demand forecasts, and so better staffing decisions. However, it is also true that many companies are reluctant to relinquish control of their call center operations. This is evidenced by a recent survey from the Incoming Call Management Institute (ICMI, 2006): only 7.9% out of 279 call center professionals used an outside vendor to handle most or all of their calls. One reason for that are the “hidden costs of outsourcing” (Kharif, 2003), which include service quality costs that are hard to explicitly quantify. As a result, many of these companies prefer to co-source; that is, to outsource some, but not all, of their calls.

We study a co-sourcing structure in which the vendor charges the company a fee per call outsourced, which is consistent with the pay-per-call (PPC) co-sourcing structure analyzed in Akşin et al. (2008). We assume the company can decide on a call-by-call basis which calls to answer in-house and which calls to route to the vendor. This is helpful because call centers typically make their daily staffing decisions at least a week in advance, before the actual arrival rate to the call center for a given day is known. If the planned staffing is sufficient to handle the mean arrival rate, then the company needs to outsource only a small fraction of calls in order to handle the inherent variability that results in congestion every so often. On the other hand, if the planned staffing is insufficient to handle the mean arrival rate, then the company can outsource a large fraction of its calls, thereby preventing high congestion levels.

The relevant question for this paper is: how do we decide on staffing levels when the arrival rate is uncertain and the aforementioned co-sourcing option is present? To answer this question, we begin with one simple and widely used queueing model of a call center, the Erlang A or M/M/N + M (see, for example, Section 4.2.2 in Gans et al. (2003)), and add uncertainty in the arrival rate and an outsourcing option. Then, our model for the call center is a multi-server queue with a doubly stochastic time-homogeneous Poisson arrival process, exponential service times, and exponential times to abandonment. Although this model ignores the time-varying nature of the arrival rate over the course of each day, there is call center literature that discusses how to use the Erlang A model to make staffing decisions for time-varying arrival rates, using the “stationary independent period by period approach” (SIPP); see Green et al. (2001), Gans et al. (2003), Akşin et al. (2007), and Liu and Whitt (2012) for more discussion. We suppose that a similar approach can be adopted here to accommodate the added feature of arrival rate uncertainty.

The control decisions in our model are (a) an upfront staffing decision and (b) real-time call
outsourcing (routing) decisions. Recall that staffing decisions are made on a much longer time horizon and well before the timing of the control decisions. In particular, these decisions are made on two different time scales. This means that we have a two stage stochastic program: The staffing decisions are made in the first stage, before the arrival rate is known, and the outsourcing decisions are made in the second stage, after the arrival rate is known. Then, the outsourcing decisions can depend on the actual arrival rate even though the staffing decisions cannot.

Our objective is to propose a policy for staffing and outsourcing under the assumption of linear staffing cost and linear abandonment and outsourcing costs. For each arriving customer that cannot be immediately served, there is a tension between choosing to outsource that customer (and paying the outsourcing fee) or having that customer wait for an in-house agent (and risking incurring an abandonment cost). In summary, we are solving a joint staffing and routing control problem for a (modified) Erlang A model with an uncertain arrival rate and an outsourcing option.

The three main contributions in this paper are:

- The *modeling* contribution is the formulation of a joint staffing and outsourcing problem for a call center that has access to co-sourcing, and must make staffing decisions when there is arrival rate uncertainty. This modeling framework can be used to study more general joint staffing and control problems in call centers that have been previously studied in the literature under the assumption of a known arrival rate.

- The *application* contribution is the development of a square root safety staffing and threshold outsourcing policy that we numerically show to be extremely robust over the entire parameter space. This robustness may come as no surprise for readers who are familiar with related literature such as Borst et al. (2004) and Gurvich et al. (2014). However, the existing literature has not addressed the issue of robustness in the context of random arrival rates and dynamic control, nor can this robustness be readily explained using existing results.

- The *technical* contribution is the proof that our proposed square root safety staffing and threshold outsourcing policy is asymptotically optimal, as the mean arrival rate becomes large, when the level of uncertainty in the arrival rate is of the same order as the inherent system stochasticity (which is of the order of the square root of the mean of the arrival rate).

The remainder of this paper is organized as follows. First, we review the most relevant literature. Next, in Section 2, we describe our model in detail. In Section 3, we present the exact (non-asymptotic) analysis which leads to an algorithm to compute the optimal policy numerically.
However, that algorithm does not provide insight into the structure of an optimal policy, and so, in Section 4, we perform an asymptotic analysis under the assumption that the level of uncertainty in the arrival rate is of the same order as the inherent system stochasticity. That asymptotic analysis motivates us to propose, in Section 5, a square root safety staffing and threshold outsourcing policy that is universal in the sense that there is no assumption on the level of uncertainty in the arrival rate. We evaluate the performance of our universal policy numerically in Section 6. We make concluding remarks in Section 7. All proofs and additional numerical results can be found in the electronic companion (EC).

**Literature Review**

Previous work on joint staffing and routing problems in call centers includes [Gurvich et al. (2008)] who study staffing and dynamic routing in call centers with multiple customer classes and a single server pool, [Armony and Mandelbaum (2011)] who consider the symmetric case of a single customer type and a heterogeneous server pool, and [Gurvich and Whitt (2010)] who consider multiple customer classes and a heterogeneous server pool. These papers study the staffing and dynamic routing problems within the Halfin-Whitt heavy traffic regime, pioneered by [Halfin and Whitt (1981)], and extended to include abandonments by [Garnett et al. (2002)]. This is also known as the Quality and Efficiency Driven (QED) heavy traffic regime. The key idea is to approximate the behavior of call centers that are modeled as multi-server queues with that of their limiting diffusions. The limiting diffusion arises from a specific relationship between the arrival rate and the staffing level as both grow large without bound. Our work is different than the aforementioned papers in that our model is pertinent to situations where the arrival rate is not known when staffing decisions are made, and thus has to be inferred or forecasted using available historical data.

Given a staffing level and a realized arrival rate, our dynamic outsourcing decision is equivalent to the admission control problem studied in [Koçağa and Ward (2010)]. The equivalence follows because we do not explicitly model the vendor firm and assume it has ample capacity to handle the outsourced calls. [Koçağa and Ward (2010)] show that a threshold admission control policy is optimal, and characterize a simple form for the threshold level that is asymptotically optimal when the staffing level is assumed to be such that the system operates in the QED regime. In contrast, this paper explicitly models the staffing decisions and has a random arrival rate.

Robbins and Harrison (2010), Bassamboo et al. (2010), Mehrotra et al. (2010), Gans et al. (2012), and Zan et al. (2014). The two works most closely related to ours are Maman (2009) and Bassamboo et al. (2010), and we discuss each in turn.

The focus of Maman (2009) is to extend the QED staffing formula under a general form of arrival uncertainty. Our asymptotic optimality result assumes a special case of the form of the arrival rate uncertainty presented in that paper. However, that paper does not explicitly study the cost minimizing staffing and does not model routing decisions, as we do.

Bassamboo et al. (2010) propose a staffing policy for an $M/M/N + M$ queue in which the arrival rate is random, and there is no outsourcing option. They establish that a simple newsvendor based staffing policy performs extremely well when the order of uncertainty in the arrival rate exceeds the order of the inherent system stochasticity. In contrast, we establish the asymptotic optimality of our proposed policy when the aforementioned two magnitudes are the same. In our numeric study, we adapt their policy to our setting with outsourcing, in order to evaluate policy robustness.

In relation to the literature on call center outsourcing (see, for example, Zhou and Ren (2011)), our paper is most similar to Akşin et al. (2008). In contrast to most papers in this literature, which assumes all calls will be outsourced, Akşin et al. (2008) considers the contract design problem of a company which faces an uncertain call volume, and can outsource part of its calls by choosing between a capacity-based and volume-based contract that is pay-per-call. Although both Akşin et al. (2008) and our model study co-sourcing decisions which are driven by call volume uncertainty, Akşin et al. (2008) focuses on the optimal contract choice, whereas we focus on the in-house staffing and dynamic routing decisions and assume the contract.

2. Model Description

We model the in-house call center (which we henceforth refer to as “the call center” or “the system”) as an $M/M/N + M$ queueing system in which the arrival rate is uncertain. We let $\Lambda$ denote the random arrival rate with known cdf $F_{\Lambda}$, and we let $l$ denote a particular realization of $\Lambda$. We assume that $\Lambda$ is a non-negative random variable with mean $E[\Lambda] = \lambda$. For ease of exposition, we assume the mean of the exponential service time is 1, so that we can think of measuring time in terms of the mean time to serve an arrival. Each customer call put on hold has an exponential patience time with mean $1/\gamma$ and abandons the system if not answered within this time. We assume that $\Lambda$ is independent of the service and patience times, and that, given a realization $l$ of $\Lambda$, the inter-arrival, service and patience times are independent and identically distributed.
The call center manager must make two decisions: the upfront staffing level \( N \), and the dynamic outsourcing decision. The staffing level \( N := N(F_\Lambda) \) must be set before the arrival rate \( \Lambda \) is realized, based on the knowledge of its distribution. After the arrival rate \( \Lambda \) is realized as \( l \), every arriving call can be either accepted into the system, or routed to the outsourcing vendor. The routing control policy \( \pi := \pi(N, l) \) is in general a function of the staffing level \( N \) and \( l \). The notation \( \pi(N, \Lambda) \) refers to a routing policy that may depend on the actual realization \( l \) of \( \Lambda \). Any stationary routing control policy \( \pi = (\pi_n : n \in \{0, 1, \ldots\}) \) is a vector, where \( \pi_n \in [0, 1] \) denotes the probability that a customer is accepted into the system when there are \( n \) customers currently present there. We let \( \Pi \) be the set of all such vectors. An admissible policy

\[
u := (N, \pi(N, \Lambda)) = (N, (\pi_n(N, \Lambda) : n \in \{0, 1, \ldots\}))\]

sets the staffing level as a non-negative integer \( N \), and, after the arrival rate \( \Lambda \) realizes as \( l \), controls outsourcing decisions dynamically by routing calls according to the policy \( \pi(N, l) \in \Pi \).

After the arrival rate \( l \) realizes, the system operates as a birth and death process with birth rate \( l \pi_n \) and death rates

\[
\mu_n = \min(N, n) + \gamma [n - N]^+. 
\]

It is straightforward to solve the balance equations for the steady-state distribution for the number-in-system process, from which it follows that the following performance measures are well-defined:

\[
P_{\pi}(\text{ab}; N, l) = \text{the probability an entering customer abandons};
\]

\[
P_{\pi}(\text{out}; N, l) = \text{the probability an arriving customer is routed to the outsourcer}. 
\]

The objective of the system manager is to minimize the expected long-run average cost, when there are costs due to customer abandonment, routing customers to the outsourcing vendor, and staffing costs. Every customer that abandons the system before receiving service costs \( a \) and the per call cost of routing to the outside vendor is \( p \) (which can also include indirect costs such as the hidden costs of outsourcing). The long-run average operating cost associated with \( \pi \in \Pi \) when the arrival rate realizes as \( l \) and the staffing level is \( N \) is

\[
z_{\pi} := z_{\pi}(N, l) = plP_{\pi}(\text{out}; N, l) + alP_{\pi}(\text{ab}; N, l). \tag{2.1}
\]

This is expressed as a random variable by replacing the realized arrival rate \( l \) in (2.1) with the random arrival rate \( \Lambda \). For a given realization \( l \) of \( \Lambda \), \( z^{\text{opt}}(N, l) \) denotes the minimum cost, and \( \pi^{\text{opt}}(N, l) \in \Pi \) is a policy that achieves that minimum cost. We let \( z^{\text{opt}}(N, \Lambda) \) denote the random
variable associated with the minimum cost and let \( \pi^{\text{opt}}(N, \Lambda) \) denote an optimal routing policy that may depend on the actual realization \( l \) of \( \Lambda \). The expected long-run average cost under the policy \( u = (N, \pi) = (N, \pi(N, \Lambda)) \), with respect to the random arrival rate \( \Lambda \), is

\[
C(u) = cN + E[z_{\pi}(N, \Lambda)].
\]  

(2.2)

We would like to find a staffing level \( N^{\text{opt}} \) and a routing control policy \( \pi^{\text{opt}}(N, \Lambda) \) that achieves the minimum long-run average cost

\[
C^{\text{opt}} := \inf_u C(u) = \min_{N \in \{0, 1, 2, \ldots \}} cN + E[z^{\text{opt}}(N, \Lambda)].
\]  

(2.3)

Remark 1. (Including waiting costs.) The objective function in (2.2) can be modified to include a customer waiting cost at the in-house call center by modifying (2.1) as follows. Suppose the cost for one customer to wait one time unit is \( w \geq 0 \). Then, (2.1) becomes

\[
z_{\pi} := z_{\pi}(N, l) = plP_{\pi}(\text{out}; N, l) + aP_{\pi}(\text{ab}; N, l) + w(l - P_{\pi}(\text{out}; N, l))\overline{W}_{\pi}(l),
\]

where \( \overline{W}_{\pi}(N, l) \) is the steady-state average waiting time, including both abandoning and served customers. Letting \( \overline{Q}_{\pi}(N, l) \) denote the steady-state average number of customers waiting in queue to be served, it follows from Little’s law that

\[
l(1 - P_{\pi}(\text{out}; N, l))\overline{W}_{\pi}(N, l) = \overline{Q}_{\pi}(N, l),
\]

and so

\[
z_{\pi} = plP_{\pi}(\text{out}; N, l) + aP_{\pi}(\text{ab}; N, l) + w\overline{Q}_{\pi}(N, l).
\]

Also, since the steady-state rate at which abandoning customers arrive must equal the steady-state abandonment rate

\[
lP_{\pi}(\text{ab}; N, l) = \gamma\overline{Q}_{\pi}(N, l),
\]

and so

\[
z_{\pi} = plP_{\pi}(\text{out}; N, l) + \left(a + \frac{w}{\gamma}\right)lP_{\pi}(\text{ab}; N, l).
\]

The analysis in this paper is valid with \( a \) replaced by \( a' := a + w/\gamma \). Therefore, to include a customer waiting cost, the only change is to replace \( a \) in (2.1) by \( a' \).

It makes intuitive sense that if \( c \geq p \) the system manager will not invest in any in-house capacity because serving an arrival is more costly than routing that arrival to the outsourcing vendor (recall that \( \mu = 1 \), so \( c, p \) and \( a \) are comparable). Continuing with such intuitive comparisons (see Table
In-house capacity investment

<table>
<thead>
<tr>
<th>c &lt; min(a, p)</th>
<th>c ≥ min(a, p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a &gt; p</td>
<td>co-sourcing</td>
</tr>
<tr>
<td>a ≤ p</td>
<td>no outsourcing</td>
</tr>
</tbody>
</table>

Table 2.1: Optimal In-house Staffing and Outsourcing Decisions

1), if $c \geq \min(a, p)$, then the system manager will either route every call to the outsourcer ($a > p$) or will let every call abandon ($p \geq a$). This suggests that it is only when $c < \min(a, p)$ that the system manager will invest in in-house capacity. Furthermore, he will not route calls to the outsourcer if $a \leq p$. In summary, we expect the system manager to invest in capacity and route some calls to the outsourcing vendor only if $c < p < a$. The following proposition confirms this observation.

**Proposition 1.** (Characterizing the parameter regimes.)

(i) Suppose that $c \geq \min(a, p)$. Then, $N_{\text{opt}} = 0$ solves (2.3).

   (a) In addition, if $a > p$, then $\pi_{\text{opt}}(N, l) = (0, 0, 0, \ldots)$, so that all calls are routed to the outsourcing vendor.

   (b) Otherwise, if $a \leq p$, then $\pi_{\text{opt}}(N, l) = (1, 1, 1, \ldots)$, so that all calls are left to abandon.

(ii) If $c < \min(a, p)$ and $a \leq p$, then the optimal control policy is $\pi_{\text{opt}}(N, l) = (1, 1, 1, \ldots)$ for any given $N \geq 0$.

From Proposition 1, it follows that the only cases with a non-trivial optimal staffing are when $c < \min(a, p)$. Hence, for the remainder of the paper we assume that $c < \min(a, p)$.

3. Exact Analysis

For a fixed staffing level $N$ and realized arrival rate $l$, the problem of minimizing $z_\pi$ is a Markov decision problem (MDP), having solution $z_{\text{opt}}(N, l)$. This MDP has been solved in Koçağa and Ward (2010) in the context of an admission control problem. It follows from Theorems 3.1, Theorem 3.2, and Theorem 3.3 in Koçağa and Ward (2010) that the optimal policy is a deterministic threshold policy (with a potentially infinite threshold level). Hence we can restrict ourselves to the class of threshold control policies

$$\tau(T) = (\tau_n(T) : n \in \{0, 1, 2, \ldots\}) \in \Pi,$$
defined for threshold level $T := T(N, l) \in [0, \infty]$ as

$$\tau_n(T) := \begin{cases} 
1 & \text{if } n < T, \\
0 & \text{if } n \geq T.
\end{cases}$$

Under the threshold policy $\tau(T)$, after the arrival rate realizes, an arriving call will be accepted into the system if and only if the number of customers currently in the system is less than the threshold. Hence, the system operates as an $M/M/N/T + M$ queue and the process tracking the number of customers in the system is a birth-and-death process on $\{0, 1, \ldots, N - 1, N, N + 1, \ldots, T\}$ with birth rate $l$ and death rate in state $n$

$$\mu_n = \min(n, N) + \gamma[n - N]^+. $$

Then, we can solve the balance equations to find the steady-state probabilities

$$\theta_k(N, l) = \left( \prod_{i=1}^{k} \frac{l}{\mu_i} \right) \theta_0(N, l)$$

for

$$\theta_0(N, l) = \frac{1}{\sum_{k=0}^{T} \left( \prod_{i=1}^{k} \frac{l}{\mu_i} \right)},$$

and develop the expressions for the performance measures\footnote{See also Section 7 in Whitt (2005) for exact expressions for other performance measures of interest, such as the expected wait time conditioned on an arrival being served, in a more general model that allows for state-dependent abandonment rates.}

$$P_{\tau(T)}(\text{out}; N, l) = \theta_T(N, l)$$

$$Q_{\tau(T)}(N, l) = \sum_{k=0}^{T} [k - N]^+ \theta_k(N, l).$$

Next, recalling from Remark\footnotemark[1] that $lP_{\tau}(\text{ab}; N, l) = \gamma Q_{\tau}(N, l)$, we can express the long-run average cost in terms of the steady-state probabilities as

$$z_{\tau(T)} = pl\theta_T(N, l) + a\gamma Q_{\tau(T)}(N, l).$$

Hence we can optimize over $T$ to find

$$T^{\text{opt}} := \arg \min_{T \in \{0,1,2,\ldots,\}} z_{\tau(T)},$$

for which

$$z_{\tau(T^{\text{opt}})}(N, l) = z^{\text{opt}}(N, l).$$
Unfortunately, the resulting expression for $z_{\tau(T)}$ is not simple, and so the above minimization over $T$ to find $T^\text{opt}$ must be performed numerically. Furthermore, it still remains to take expectations, and minimize over the staffing level to numerically solve for

$$N^\text{opt} = \arg \min_{N \in \{0, 1, 2, \ldots\}} cN + E \left[ z_{\tau(T^\text{opt})}(N, \Lambda) \right],$$

and the associated minimum cost $C^\text{opt}$. To do this, we must perform an exhaustive search over $N$. The reason an exhaustive search should be performed is that it is very difficult to establish, in general, that the cost in (2.2) is convex in the staffing level $N$. In fact, even for a system where the arrival rate is known and there is no outsourcing option, convexity results are yet to be established when $\mu < \gamma$ (Armony et al. (2009) and Koole and Pot (2011)).

The exhaustive search to find $N^\text{opt}$ involves a numeric integration to calculate the expectation with respect to the arrival rate $\Lambda$, and, for each value $l$ used in the numeric integration, there is another search that must be performed to find $z^\text{opt}(N, l)$. In other words, the exhaustive search algorithm is not a simple line search as it includes three nested layers of enumeration that correspond to the staffing level, the arrival rate used in numeric integration and the outsourcing threshold. The exhaustive search algorithm is formally described as follows:

**Initialization:** Set $N^0 = 0$, $C^0 = C((0, \pi^\text{opt}(0, \Lambda))) = \min(a, p) \lambda$, and $N = 1$. Decide on the maximum possible staffing level to allow, $N_{\text{max}}$.

**Step 1:** Compute $C((N, \pi^\text{opt}(N, \Lambda))) = cN + E \left[ z^\text{opt}(N, \Lambda) \right]$ via numeric integration. For each possible arrival rate realization $l$ in the numeric integration, initialize $T = N$, decide on the stopping criterion$^3$ and then compute $T^\text{opt} = T^\text{opt}(N, l)$ and $z^\text{opt} = z^\text{opt}(N, l)$ as follows.

(A) Solve for $z_{\tau(T)}$ from the steady-state probabilities $\{\theta_n(T) : n = 0, 1, 2, \ldots, T\}$.

(B) If $z_{\tau(T+1)} \geq z_{\tau(T)}$ or if the stopping criterion holds, set $T^\text{opt} = T$, $z_{\tau(T)} = z^\text{opt}$ and stop.

Otherwise, increase $T$ by 1 and go to step (A).

**Step 2:** If $C((N, \pi^\text{opt}(N, \Lambda))) < C^0$, then $N^0 \leftarrow N$ and $C^0 \leftarrow C((N, \pi^\text{opt}(N, \Lambda)))$.

**Step 3:** If $N = N_{\text{max}}$, then set $N^\text{opt} = N^0$ and $C^\text{opt} = C^0$ and stop. Otherwise, increase $N$ by 1 and go to Step 1.

Although the algorithm above can compute the optimal policy numerically, it does not provide any insight with regards to the structure of the optimal policy. Furthermore, the computation

$^2$Lemma 3.2 in Koçaka and Ward (2010) implies $T^\text{opt} > N$.

$^3$For example, Theorem 3.4 in Koçaka and Ward (2010) provides a bound on the difference between the current cost and the minimum cost.

$^4$If $z_{\tau(T+1)} \geq z_{\tau(T)}$, Theorem 3.2 in Koçaka and Ward (2010) implies that $T* = T$. 

11
time to obtain an optimal policy can be several hours for large system sizes. (We implemented the algorithm in Matlab.) Therefore we take the following approach to develop our proposed policy: we evaluate the performance of the family of square root safety staffing policies combined with threshold routing. To do this, we first assume that the form of the arrival rate uncertainty is on the order of the square root of the mean arrival rate, and then show that square root safety staffing combined with threshold routing is asymptotically optimal (Section 4). Second, we propose a universal policy $U$ that is based on that asymptotic optimality result (Section 5), and show numerically that not only its computation time is in the order of seconds, it also has a very good performance even outside of the regime in which we proved its asymptotic optimality (Section 6).

4. Asymptotic Analysis

In this section of the paper only, we assume that the order of uncertainty in the arrival rate is the same as the square-root of the mean of the arrival rate. To do this, we consider a sequence of systems indexed by the mean arrival rate $\lambda$, and let $\lambda \to \infty$. We assume that the random arrival rate $\Lambda$ throughout this sequence of systems can be expressed as

$$\Lambda = \lambda + X \sqrt{\lambda}, \quad (4.1)$$

where $X$ is a random variable with mean zero and has $E|X| < \infty$. For this section only, the expectation operator is with respect to $X$ (instead of $\Lambda$). We note that this form for the arrival rate is a special case of the model assumed in Maman (2009). Our convention is to use the superscript $\lambda$ to denote a process or quantity associated with the system having random arrival rate $\Lambda = \Lambda^\lambda(X)$ given in (4.1). The notation

$$l^\lambda(x) = \lambda + x \sqrt{\lambda}$$

denotes the realized arrival rate in the system having mean arrival rate $\lambda$; the $\lambda$ superscript should remind the reader that $l^\lambda(x) \to \infty$ as $\lambda \to \infty$ for any $x \in (-\infty, \infty)$.

An admissible policy $u = (N, \pi) := \{(N^\lambda, \pi^\lambda) : \lambda \geq 0\}$ refers to an entire sequence that specifies an admissible policy for each $\lambda$. In particular, $N^\lambda$ is a non-negative integer and $\pi^\lambda = \pi^\lambda(N^\lambda, l^\lambda(x)) \in \Pi$ for each $\lambda$ and any realization $x$ of $X$ (so that the system arrival rate is $l^\lambda(x)$). The notation $z^\lambda_\pi(N, l^\lambda(x))$ is the long-run average operating cost, as defined in (2.1), for the system with realized arrival rate $l^\lambda(x)$, and $z^\lambda_\pi(N, \Lambda^\lambda(X))$ is the associated random variable. Similarly, $P^\lambda_\pi(\text{out}; N, l^\lambda(x))$ ($P^\lambda_\pi(\text{ab}; N, l^\lambda(x))$) is the steady-state probability an arriving customer is
routed to the outsourcer (probability of abandonment) when the realized arrival rate is $I^\lambda(x)$, and $P^\lambda_{\pi}(\text{out}; N, \Lambda^\lambda(X))$ ($P^\lambda_{\pi}(\text{ab}; N, \Lambda^\lambda(X))$) is the associated random variable.

In this section, we first define what we mean by asymptotic optimality (Section 4.1). Then, we perform an asymptotic analysis in order to understand the behavior of the family of square-root staffing policies combined with threshold routing (Section 4.2). Finally, we optimize over the aforementioned policy class to obtain our proposed policy (Section 4.3), and we establish its asymptotic optimality.

### 4.1 The Asymptotic Optimality Definition

Our definition of asymptotic optimality is motivated by first observing that the lowest achievable cost on fluid scale is $c\lambda + o(\lambda)$, where the notation $f^\lambda = o(g^\lambda)$ means that $\lim_{\lambda \to \infty} f^\lambda / g^\lambda = 0$.

**Proposition 2.** (Fluid-scaled cost.) Under the assumption (4.1) we have that:

(i) Any admissible policy $u = (N, \pi)$ has

$$\liminf_{\lambda \to \infty} \frac{cN^\lambda + E \left[ z^\lambda_{\pi}(N, \Lambda^\lambda(X)) \right]}{\lambda} \geq c.$$  

(ii) If $N^\lambda = \lambda + \beta \sqrt{\lambda} + o(\lambda)$, then, under the routing policy $\tau(\infty)$ that outsources no customers,

$$\lim_{\lambda \to \infty} \frac{cN^\lambda + E \left[ z^\lambda_{\tau(\infty)}(N, \Lambda^\lambda(X)) \right]}{\lambda} = c.$$  

The following refined and diffusion scaled cost function (defined for the any admissible policy $u = (N, \pi)$)

$$\hat{C}^\lambda(u) := \sqrt{\lambda} \left( \frac{cN^\lambda + E \left[ z^\lambda_{\pi}(N, \Lambda^\lambda(X)) \right]}{\lambda} - c \right) \geq 0$$  

(4.2) captures both the cost of additional staffing (above the offered load level $\lambda$) and the cost of the routing control.

**Definition 1.** (Asymptotic optimality.) An admissible policy $u^* = (N^*, \pi^*) = \{(N^\lambda^*, \pi^\lambda^*(N^\lambda^*, \Lambda^\lambda(X)) : \lambda \geq 0\}$ is asymptotically optimal if

$$\limsup_{\lambda \to \infty} \hat{C}^\lambda(u) < \infty$$  

and

$$\limsup_{\lambda \to \infty} \hat{C}^\lambda(u^*) \leq \liminf_{\lambda \to \infty} \hat{C}^\lambda(u),$$

for any admissible policy $u$.  

4.2 The Asymptotic Behavior of Square Root Safety Staffing Combined with Threshold Routing

It has been shown in the extensive literature on staffing in large-scale service systems (e.g. Halflin and Whitt (1981); Borst et al. (2004); Mandelbaum and Zeltyn (2009)) that when the arrival rate is deterministic, square root safety staffing performs extremely well in minimizing both the staffing plus delay costs as well as staffing costs subject to performance constraints. When the arrival rate $\lambda$ is large, under square root safety staffing, the waiting times are small (at the order of $1/\sqrt{\lambda}$), so that the percentage of customers that should be routed to the outsourcer (Koçağa and Ward (2010)), as well as the percentage of customers that abandon (Garnett et al. (2002)) are both small. This suggests that square root safety staffing should be also relevant when the arrival rate is random. Similarly to Koçağa and Ward (2010), to route calls, we use a threshold routing policy, $\tau = \{\tau(T^\lambda) : \lambda \geq 0\}$, as defined in Section 3. The threshold level $T^\lambda = T^\lambda(N^\lambda, \Lambda^\lambda(X))$ is determined after the arrival rate realizes as $l^\lambda(x)$.

The following lemma establishes the asymptotic behavior of square root safety staffing combined with threshold routing for a fixed realization $x$ of $X$. Let $\phi$ and $\Phi$ be the standard normal pdf and cdf, respectively.

**Lemma 1.** (Asymptotic behavior with deterministic arrival rate.) Suppose the random variable $X$ realizes as the value $x \in (-\infty, \infty)$. Assume the policy $u = (N, \tau)$ is such that

\[
N^\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda}) \tag{4.3}
\]

\[
T^\lambda = N^\lambda + \hat{T}\sqrt{l^\lambda(x)}, \text{ where } \hat{T} := \hat{T}(\beta, x) \in [0, \infty). \tag{4.4}
\]

Suppose the initial number of customers in the system $Y^\lambda_0$ is such that $Y^\lambda_0 - N^\lambda \Rightarrow \hat{Y}(0)$ as $\lambda \to \infty$, for some random variable $\hat{Y}(0)$ that is finite with probability 1. Then,

\[
\frac{1}{\sqrt{\lambda}} z^\lambda(N, l^\lambda(x)) \to \hat{z}(x - \beta, \hat{T}), \text{ as } \lambda \to \infty,
\]

where

\[
\hat{z}(m, \hat{T}) := \frac{A(m, \hat{T})}{B(m, \hat{T})} \tag{4.5}
\]

for

\[
A(m, \hat{T}) := p\phi\left(\hat{T} + \frac{m}{\gamma}\right) + \left(a + \frac{w}{\gamma}\right) \left[\phi\left(\frac{m}{\gamma}\right) - \phi\left(\hat{T} + \frac{m}{\gamma}\right)\right] + \frac{m}{\sqrt{\gamma}} \left[\Phi\left(\frac{m}{\sqrt{\gamma}}\right) - \Phi\left(\hat{T} + \frac{m}{\gamma}\right)\right]
\]

\[
B(m, \hat{T}) := \frac{\phi\left(\frac{m}{\sqrt{\gamma}}\right)}{\phi(m)} \Phi(m) + \frac{1}{\sqrt{\gamma}} \left[\Phi\left(\hat{T} + \frac{m}{\gamma}\right) - \Phi\left(\frac{m}{\sqrt{\gamma}}\right)\right]
\]

14
The appearance of $\beta - x$ as an argument in $\hat{z}$ in Lemma 1 occurs because under (4.1) the staffing $N^\lambda$ in (4.3) is such that the system operates in the QED regime regardless of the realization $x$ of $X$; in particular,
\[
\frac{N^\lambda - l^\lambda(x)}{\sqrt{\lambda}} = \frac{N^\lambda - \lambda}{\sqrt{\lambda}} - x \to \beta - x \text{ as } \lambda \to \infty.
\]
Furthermore, the following Corollary to Lemma 1 highlights that the dependence of the threshold level on the realized arrival rate is through the definition of $\hat{T}$, and not through its multiplier (which is always of order $\sqrt{\lambda}$ under the assumption (4.1)).

**Corollary 1.** Lemma 1 continues to hold when $T^\lambda$ in (4.4) is re-defined as
\[
T^\lambda = N^\lambda + \hat{T}\sqrt{\lambda}.
\]

The issue is that in order to analyze the performance of square root safety staffing combined with threshold routing, we require that Lemma 1 and Corollary 1 hold when the fixed value $x$ is replaced by the random variable $X$.

**Theorem 1.** *(Asymptotic cost convergence.)* Assume the policy $u = (N, \tau)$ is as defined by the equations (4.3) and (4.4). Then, under the conditions of Lemma 1,
\[
\hat{C}^\lambda(u) \to \hat{C}(u) := c\beta + E \left[ \hat{z}(\beta - X, \hat{T}) \right], \text{ as } \lambda \to \infty.
\]

### 4.3 The Proposed Policy

It is sensible to set the parameters $\beta$ and $\hat{T}$ of Lemma 1 in order to minimize the limiting cost $\hat{C}(u)$ of Theorem 1. The first step is to observe that, for $p < a$ and any given $\beta$, Proposition 4.1 in Koçaga and Ward (2010) shows that for the realized arrival rate $l^\lambda(x)$, the unique $\hat{T}^* = \hat{T}^*(\beta - x) < \infty$ that solves
\[
(a - p)\gamma\hat{T} - \hat{z}(\beta - x, \hat{T}) = p(\beta - x)
\]
has the property that
\[
\hat{z}(\beta - x, \hat{T}^*) \leq \hat{z}(\beta - x, \hat{T})
\]
for any other $\hat{T} \geq 0$. Otherwise, for $a \leq p$, $\hat{T}^* = \infty$ and
\[
\hat{z}(\beta - x, \infty) := \lim_{\hat{T} \to \infty} \hat{z}(\beta - x, \hat{T}) \leq \hat{z}(\beta - x, \hat{T}_0)
\]
for any finite $\hat{T}_0 \geq 0$. The second step is to plug $\hat{T}^*$ into the limiting expression in Theorem 1 and to optimize over $\beta$ to find
\[
\beta^* := \arg \min_\beta \left\{ c\beta + E \left[ \hat{z}(\beta - X, \hat{T}^*(\beta - X)) \right] \right\}.
\]
It is important to observe that $\beta^*$ is well-defined in the sense that $\beta^*$ is finite and

$$c\beta^* + E\left[\hat{z}(\beta^* - X, \hat{T}^*(\beta^* - X))\right] = \inf_{\beta \in (-\infty, \infty)} c\beta + E\left[\hat{z}(\beta - X, \hat{T}^*(\beta - X))\right] < \infty.$$ 

This follows from the next two propositions.

**Proposition 3.** For any $\beta \in (-\infty, \infty)$, $c\beta + E\left[\hat{z}(\beta - X, \hat{T}^*(\beta - X))\right] < \infty$.

**Proposition 4.** The infimum in $\inf_{\beta \in (-\infty, \infty)} c\beta + E\left[\hat{z}(\beta - X, \hat{T}^*(\beta - X))\right]$ is attained by a finite $\beta \in (-\infty, \infty)$.

We are now in a position to define our proposed policy: We let

$$u^* = (N^*, \tau^*) := \{(N^\lambda, \tau(T^\lambda)) : \lambda \geq 0\} \quad (4.10)$$

that has the staffing level

$$N^\lambda = \lambda + \beta^* \sqrt{\lambda} \quad (4.11)$$

and sets the threshold level $T^\lambda = T^\lambda(N^\lambda, \Lambda^\lambda(X))$ when the arrival rate realizes as $l^\lambda(x)$ as

$$T^\lambda = N^\lambda + \hat{T}^*(\beta^* - x) \times \sqrt{l^\lambda(x)}, \quad (4.12)$$

for $\hat{T}^*(\beta^* - x)$ defined by (4.6) with $\beta^*$ replacing $\beta$.

Theorem 1 is valid for $u^*$, and so

$$\hat{C}^\lambda(u^*) \to \hat{C}^* := c\beta^* + E\left[\hat{z}(\beta^* - X, \hat{T}^*(\beta^* - X))\right]$$

Our next result confirms that $C^*$ is the minimum achievable cost, meaning that the policy $u^*$ is asymptotically optimal.

**Theorem 2.** (Asymptotic optimality of our proposed policy.) The policy $u^*$, defined through (4.9), (4.10), (4.11), and (4.12) is asymptotically optimal under (4.1); i.e., under any other admissible policy $u$

$$\lim_{\lambda \to \infty} \inf \hat{C}^\lambda(u) \geq \hat{C}^*.$$ 

Furthermore, it follows that our proposed policy has associated cost that is $o(\sqrt{\lambda})$ higher than the minimum achievable cost for a given $\lambda$; i.e., that

$$\frac{cN^\lambda + E\left[z^\lambda_*, (N^*, \Lambda^\lambda(X))\right] - C^\lambda_{\text{opt}}}{\sqrt{\lambda}} \to 0, \quad \text{as } \lambda \to \infty,$$

where $C^\lambda_{\text{opt}} := C_{\text{opt}}$ for $C_{\text{opt}}$ defined in (2.3) for the system with mean arrival rate $\lambda$. 

16
Remark 2. (Performance under the optimal threshold.) Another asymptotically optimal policy

\[(N^*, \pi^{opt}) := \{(N^{\lambda^*, \pi^{\lambda^*, opt}}) : \lambda \geq 0\}\]

has staffing levels \(N^{\lambda^*}\) defined in (4.11), and, after the arrival rate \(\Lambda^\lambda(X)\) realizes as \(l^\lambda(x)\), solves the relevant Markov decision problem for the routing control policy \(\pi^{\lambda^*, opt} = \pi^{\lambda^*, opt}(N^{\lambda^*}, l^\lambda(x))\) that achieves the minimum long-run average operating cost \(z^{\lambda^*, opt}(N^*, l^\lambda(x))\). To see this, it is enough to observe that

\[z_{\tau^*}(N^*, l^\lambda(x)) \geq z^{\lambda^*, opt}(N^*, l^\lambda(x)),\]

for every \(\lambda\) and any realization \(x\) of \(X\).

In the following, \(f^\lambda = O(g^\lambda)\) means that \(\limsup_{\lambda \to \infty} |f^\lambda / g^\lambda| < \infty\).

Remark 3. (Comparison to Bassamboo et al. (2010).) When \(a < p\) (in addition to our assumption that \(c < \min(a, p)\)), it follows from Proposition 1 part (ii) that the optimal control policy does not outsource any calls. Then, the cost minimization problem (2.2) is a pure staffing problem (instead of a joint staffing and routing problem), which is equivalent to the problem solved in Bassamboo et al. (2010). Theorem 1 part (c) of that paper, adapted to our setting, shows that a policy based on a newsvendor prescription can have associated cost that is \(O(\sqrt{\lambda})\) higher than the minimum achievable cost for a given \(\lambda\). In comparison, our proposed policy has associated cost that is \(o(\sqrt{\lambda})\) higher than the minimum achievable cost for a given \(\lambda\) by Theorem 2. Hence we expect our policy to provide significant improvements over that of Bassamboo et al. (2010), as the arrival rate uncertainty decreases.

5. The Proposed Universal Policy

For models that do not assume uncertain arrival rates, square root safety staffing is known in the literature to be very robust. For an \(M/M/N\) queue with no abandonments, no dynamic routing decisions, and known arrival rate, Borst et al. (2004) show that square root safety staffing performs extremely well, both inside and outside of the parameter regime (linear staffing and waiting costs) in which they prove it to be asymptotically optimal (see their numerical experiments in Section 10). In a more recent paper, Gurvich et al. (2014) prove that performance approximations that are based on the premise that the staffing is of a square-root safety form are asymptotically universally accurate, as the arrival rate becomes large. This latter paper is also limited to the case of deterministic arrival rates and no dynamic control.
This leads us to propose the universal policy $U$ when there are no restrictions on the form of the arrival rate uncertainty, as in (4.1). We define $U$ for the model as specified in Section 2 and analyzed exactly in Section 3 without considering a sequence of systems as in Section 4. To do this, we begin with the non-negative random variable $\Lambda$ that represents the system arrival rate and has mean $E[\Lambda] = \lambda$. Then, we make the transformation

$$X := \frac{\Lambda - \lambda}{\sqrt{\lambda}}$$

and use the random variable $X$ to define $U$

$$U = (N_U, \pi_U(N_U, \Lambda)).$$

The proposed staffing level is

$$N_U = \left[ \lambda + \beta^* \sqrt{\lambda} \right],$$

for $\beta^*$ that satisfies (4.9), with $X$ in that expression defined by (5.1), and the function $\lceil \cdot \rceil$ rounds the expression inside the brackets to the nearest integer. The proposed routing policy when the arrival rate $\Lambda$ realizes as $l$ is the threshold routing policy

$$\pi_U(N, l) = \tau(T_U)$$

for

$$T_U = N_U + \hat{T}^* \sqrt{l},$$

and $\hat{T}^*$ defined by (4.6), with $x$ in that expression replaced by $(l - \lambda)/\sqrt{\lambda}$.

The universal $U$ policy “pretends” that the magnitude of the uncertainty in the arrival rate $\Lambda$ is on the order of $\sqrt{\lambda}$, as in (4.1), and sets $\beta^*$ and $\hat{T}^*$ accordingly. In contrast to the policy defined in Section 4.3 under assumption (4.1), the magnitude of the second order term appearing in the definitions of $N_U$ and $T_U$ may not be of order $\sqrt{\lambda}$. In particular, depending on the distribution of $\Lambda$, the value of $\beta^*$ may end up being of the same order of $\sqrt{\lambda}$, so that the second term in $N_U$ is of order $\lambda$ (see discussion in Section 6.5). This flexibility suggests that $U$ may perform well, even outside of the regime in which it is proved to be asymptotically optimal, as we indeed observe in the next section.

6. Numerical Evaluation of the Proposed Policy

Theorems 1 and 2 establish when the order of uncertainty in the arrival rate is the same as the square-root of the mean arrival rate, so that (4.1) holds, $U$ staffs and routes in a way that achieves
minimum cost for large enough $\lambda$. However, Theorems 1 and 2 do not provide guidance on: how large $\lambda$ must be, what happens when (4.1) does not hold, or how $U$ performs in comparison to alternative benchmark policies. In this section we show that $U$ generally achieves within 0.1% of the minimum cost even when these assumptions are relaxed, and its robustness in comparison to two benchmark policies is the highest. To do this, we first vary the system size expressed by the mean arrival rate $\lambda$ (Section 6.1) and the level of uncertainty in $\Lambda$ (Section 6.2) to gain an initial conclusion that $U$ performs remarkably well. Then, we show that this conclusion is, to a large degree, insensitive to changes in the cost parameters (Section 6.3) and the asymmetry of the arrival rate distribution (Section 6.4). In summary, $U$ performs extremely well, even when the system is far away from the regime in which it is proved to be asymptotically optimal.

Throughout our numerical examples, we set the mean service time and the mean patience time equal to 1 and fix the cost parameters at $c = 0.1$, $p = 1$ and $a = 5$ unless specified otherwise. It follows from Proposition 1 that our choice of cost parameters is such that it is optimal for the system manager to set a nonzero staffing level and routes some calls to the outsourcer.

6.1 Finite System Size

Having established that $U$ is asymptotically optimal as the system size grows without bound, under the form of uncertainty in $\Lambda$ as in (4.1), we proceed to evaluate its performance for finite size systems. This evaluation is done by comparing $U$ to the numerically computed optimal staffing policy. We compute the optimal staffing level $N^{\text{opt}}$ via an exhaustive search, as described in Section 3.

Table 6.1 illustrates the performance of our proposed staffing policy with respect to the optimal staffing level by varying the system size and letting the distribution of the arrival rate $\Lambda$ be in accordance with (4.1). Specifically, we assume that $X$ follows a Uniform distribution on $[-1, 1]$, and increase the mean arrival rate $\lambda$ from 1 to 1600. This implies that $\Lambda$ follows a Uniform distribution with its support interval increasing from $[0, 2]$ to $[1560, 1640]$. This is consistent with our assumption in Theorems 1 and 2 that prove asymptotic optimality of $U$ as $\lambda$ becomes large under (4.1). The first and second columns in Table 6.1 show the resulting distribution for $\Lambda$. The third and fourth columns in Table 6.1 show the optimal staffing level and the associated optimal average cost, while columns five and six show our proposed approximate staffing policy, along with its average cost. Column seven displays the staffing error which is the difference between the optimal staffing level and our approximate staffing level. Finally, column eight displays the percentage cost error with respect to the optimal policy.

We see from Table 6.1 that $U$ performs extremely well for all system sizes, that are consistent
with the assumption that the uncertainty in the arrival rate is of the same order as the square-root of the mean arrival rate. Notice that the percentage cost error may be nonzero even when the staffing error equals zero. This is because $U$ sets the threshold level according to (4.12) which may not equal to the optimal threshold. In light of Theorems 1 and 2 that establish asymptotic optimality, it is not surprising that our policy performs extremely well for large $\lambda$. The less expected numerical insight is that $U$ also performs extremely well for small $\lambda$. (We note that there is a chance that the rounding can go the wrong way, and subsequently may cause a large cost error in extremely small system size. However such small systems sizes are not realistic for most call center applications).

In summary, $U$ is very robust to system size, provided the order of uncertainty in the arrival rate is as assumed in (4.1).

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Distribution of $\Lambda$</th>
<th>Optimal Policy</th>
<th>$U$</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$U[0,2]$</td>
<td>3</td>
<td>0.4149</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>$U[6,12]$</td>
<td>16</td>
<td>1.7702</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>$U[20,30]$</td>
<td>36</td>
<td>3.8979</td>
<td>36</td>
</tr>
<tr>
<td>100</td>
<td>$U[90,110]$</td>
<td>121</td>
<td>12.7131</td>
<td>121</td>
</tr>
<tr>
<td>226</td>
<td>$U[210,240]$</td>
<td>257</td>
<td>26.5227</td>
<td>257</td>
</tr>
<tr>
<td>400</td>
<td>$U[380,420]$</td>
<td>443</td>
<td>45.3338</td>
<td>442</td>
</tr>
<tr>
<td>625</td>
<td>$U[600,650]$</td>
<td>678</td>
<td>69.1435</td>
<td>678</td>
</tr>
<tr>
<td>900</td>
<td>$U[870,930]$</td>
<td>964</td>
<td>97.9536</td>
<td>963</td>
</tr>
<tr>
<td>1600</td>
<td>$U[1560,1640]$</td>
<td>1685</td>
<td>170.5732</td>
<td>1684</td>
</tr>
</tbody>
</table>

Table 6.1: Performance of $U$: increasing system size.

### 6.2 Varying Arrival Rate Uncertainty

Next, we evaluate the robustness of $U$ with respect to changes in the level of uncertainty in the arrival rate. This is important because the proof of asymptotic optimality of $U$ requires the assumption that the level of uncertainty in the arrival rate is of the same order as the square-root of the mean arrival rate (i.e., that (4.1) holds). We measure the level of uncertainty in the arrival rate through its coefficient of variation $CV := CV_\Lambda = \frac{\sqrt{\text{Var}[\Lambda]}}{\text{E}[\Lambda]}$. We are interested in both cases where the level of uncertainty in the arrival rate is lower than that assumed in (4.1) and where it is higher.

In this subsection, we keep the mean arrival rate fixed at $\lambda = 100$, and we assume that $\Lambda$ follows a Uniform distribution with support $[a, b]$, so that

$$CV = \frac{1}{\sqrt{3}} \frac{b - a}{a + b} \leq \frac{1}{\sqrt{3}} = 0.5774.$$
In comparison, under assumption (4.1), when $X$ follows a uniform distribution with support $[-1, 1]$ as in Section 6.1, the coefficient of variation of the arrival rate $\Lambda = \Lambda(X)$ is

$$CV_{\Lambda(X)} = \frac{\sqrt{Var[X]}}{\sqrt{\lambda}} = \frac{1}{10\sqrt{3}} = 0.0577. \quad (6.1)$$

By varying $CV$ from 0 to approximately $1/2$, we cover both the cases where the level of uncertainty in the arrival rate is lower that (4.1) and higher.

It is sensible to compare the performance of $U$ to two other possible staffing policies: one that is expected to perform well when the level of uncertainty in the arrival rate is low and the other that is expected to perform well when the level of uncertainty in the arrival rate is high. The first alternative policy we consider is $D$, a square root safety staffing policy that has the same form as (2), but chooses the coefficient of $\sqrt{\lambda}$ differently by assuming that the arrival rate is deterministic and fixed at the mean arrival rate $\lambda = 100$. Specifically when the mean arrival rate is $\lambda$, $D$ staffs

$$N_D := \left[ \lambda + \beta^*_1 \sqrt{\lambda} \right],$$

where

$$\beta^*_1 := \arg\min_\beta c\beta + \hat{z}(\beta, \hat{T}^*(\beta))$$

for $\hat{z}$ as defined in (4.5) and $\hat{T}(\beta)$ that satisfies (4.6). Note that $D$ is exactly the proposed policy $U$ in the case $P(X = 0) = 1$. It is intuitive to expect that the performance of $D$ deteriorates significantly as $CV$ increases.

The second alternative policy we consider is $NV$, a newsvendor based prescription that is a modification of the policy proposed in Bassamboo et al. (2010) to include co-sourcing. The $NV$ policy follows a fluid approximation which ignores stochastic queueing effects and, as a result, when $a > p$, the abandonment cost is irrelevant. That is, in the fluid scale, all customers who cannot be served in-house immediately upon arrival will be outsourced. Similarly, when $a \leq p$ no calls will be outsourced. In particular, in newsvendor terminology the overage cost is $c$ (because of extra staffing) and the underage cost is $\min\{a, p\} - c$ (because we incur the cost of routing or cost of abandonment but do not incur the cost of an additional person for staffing). Then, the critical ratio is $\frac{\min\{a, p\} - c}{\min\{a, p\}}$, and the newsvendor based staffing prescription is

$$N_{NV} := \left[ F_X^{-1} \left( \frac{\min\{a, p\} - c}{\min\{a, p\}} \right) \right].$$

We observe that when (4.1) holds, $N_{NV}$ can also be written as $N_{NV} := \left[ \lambda + \beta^*_2 \sqrt{\lambda} \right]$, where $\beta^*_2 := F_X^{-1} \left( \frac{p-c}{p} \right)$ and $F_X$ is the cumulative distribution function of $X$. Notice that, in sharp contrast to
D, which disregards the uncertainty in the arrival rate, NV disregards the inherent stochasticity of the system that produces queueing. Hence, we expect the performance of the NV policy to deteriorate when CV decreases.

We have specified the staffing rules, \( N_D \) and \( N_{NV} \), of two alternative policies D and NV. There is still the question of what should be the routing policy. For this, we recall that after the arrival rate realizes, the optimal routing policy can be found by solving the relevant MDP (see Koçaña and Ward (2010)). Hence, in our numerical experiments, after the arrival rate realizes, we operate both comparison policies under the optimal routing policy. The U policy follows the threshold routing policy \( \tau(T^{\lambda,\star}) \) where \( T^{\lambda,\star} = T^{\lambda,\star}(N^{\lambda,\star}, \Lambda(X)) \) is as defined in Section 5 (although we observe that the performance of the diffusion based threshold routing policy and the exact solution to the relevant MDP are almost indistinguishable).

Figure 6.1 plots the relative percentage cost error and staffing error of U, NV and D. The staffing error and percentage cost error for U is defined as in columns seven and eight in Table 6.1 and is defined similarly for D and NV. Table B.1 in EC contains further details regarding this study, such as the exact costs and staffing levels. We see from Figure 6.1 that U staffs very close to the optimal staffing level and therefore performs well even for very high CV values. We also see that U outperforms NV for lower CV values and outperforms D for higher CV values. Furthermore, in both cases, the staffing and percentage cost error can be arbitrarily large. Hence we conclude that U is robust and performs extremely well even in parameter settings beyond which it has been proven to be asymptotically optimal.

Figure 6.1 is a first step in concluding that U is very robust, and performs extremely well over a large range of parameter settings much beyond where Theorems 1 and 2 establish its asymptotic optimality. The next step in establishing the aforementioned conclusion is to explore the effect of

\[
\begin{align*}
\text{(a) % cost error} & \\
\text{(b) staffing error (i.e., } N^{\text{opt}} - \Delta \text{ where } \Delta \in \{N_U, N_D, N_{NV}\})
\end{align*}
\]
varying other parameters, for example the staffing cost.

6.3 Varying Staffing Costs

Next, we explore the effect of the staffing cost, which determines the associated critical ratio of the newsvendor policy. To do this, we change the staffing cost $c$ while holding the other parameters constant. We perform three separate studies by fixing the arrival rate distribution at three separate levels of uncertainty: low $CV$, moderate $CV$ and high $CV$. In particular, we assume $\Lambda \sim \mathcal{U}[90, 110]$ to produce low $CV$ (Figures 6.2a and 6.3a), $\Lambda \sim \mathcal{U}[50, 150]$ to produce moderate $CV$ (Figures 6.2b and 6.3b), and $\Lambda \sim \mathcal{U}[10, 190]$ to produce high $CV$ (Figures 6.2c and 6.3c). We plot the percentage cost errors in Figure 6.2 and the staffing errors in Figure 6.3 for $U$, $D$ and $NV$. We refer the reader to Tables B.2-B.4 in EC for further details (exact costs and staffing levels).

![Figure 6.2: % cost error for changing staffing costs at three levels of arrival rate uncertainty](image1)

![Figure 6.3: Staffing error (i.e., $N_{\text{opt}} - \Delta$ where $\Delta \in \{N_U, N_D, N_{NV}\}$) for changing staffing costs at three levels of arrival rate uncertainty](image2)

We first observe that for all staffing costs and across all levels of $CV$, $U$ staffs very close to the optimal policy and thus performs extremely well. On the other hand, $NV$ performs poorly when the staffing cost is low although the effect gets less pronounced for higher $CV$ values. This is because $NV$ tends to understaff for low staffing costs when the $CV$ is low. As a result, when $c \ll p < a$, $NV$ incurs higher routing control and abandonment costs. As the in-house staffing
cost $c$ increases to $p = 1$, $NV$ tends to overstaff, although the adverse effects of overstaffing are not as detrimental. We see that $D$ performs very poorly as the $CV$ level increases because it fails to capture the effect of randomness. Overall, we see that $U$ is robust and performs well across different critical ratio and $CV$ combinations while the alternative policies can perform arbitrarily poorly.

### 6.4 Effect of distribution asymmetry

Our numerical result thus far have assumed symmetric Uniform arrival rate distributions. Next, we generalize our results by considering arrival rate distributions that are asymmetric and follow a Beta distribution to study the effect of skewness on the performance of $U$ and the other policies. Specifically, we assume that $\Lambda \sim Beta(\alpha_1, \alpha_2, \lambda - b\sqrt{\lambda}, \lambda + b\sqrt{\lambda})$, where the first two arguments are the scale parameters of the distribution and the last two arguments are the lower and upper bounds of the support. We let $b$ and $\lambda$ be arbitrarily large so that the arrival rate may not realize in the QED regime (i.e., the assumption 4.1 is not necessarily satisfied). Our proposed policy $U$ is defined for $X \sim Beta(\alpha_1, \alpha_2, b, \lambda)$ from (5.1).

We keep the mean arrival rate fixed at $E[\Lambda] = \lambda = 100$ (i.e., $E[X] = 0$) throughout this section and we consider three cases where we keep the variance of the arrival rate fixed at three levels: The low $CV$ case keeps the variance of $\Lambda$ fixed and equal to that of a $U[90, 110]$ random variable (i.e., $Var(X) = Var(U[-1, 1])$), the moderate $CV$ case keeps the variance of $\Lambda$ fixed and equal to that of a $U[50, 150]$ random variable (i.e., $Var(X) = Var(U[-5, 5])$), and the high $CV$ case keeps the variance of $\Lambda$ fixed and equal to that of a $U[10, 190]$ random variable (i.e., $Var(X) = Var(U[-9, 9])$).

We study the effect of asymmetry by changing the skewness of the Beta distribution through its scale parameters $\alpha_1$ and $\alpha_2$. In particular, we set $E[X] = \frac{\alpha_1 b + \alpha_2 b}{\alpha_1 + \alpha_2} = 0$ and $Var(X) = \frac{\alpha_1 \alpha_2 (\beta - \lambda)^2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)} = \sigma^2$, where $\sigma^2$ denotes the variance of the associated $CV$ level, and we choose $\alpha_1$ and $\alpha_2$ such that $\alpha_1 + \alpha_2 = 2$. We start with a negative-skewed (left-skewed) Beta distribution with scale parameters $\alpha_1 = 1.5$ and $\alpha_2 = 0.5$ with a corresponding skewness of $-1$. Then, we decrease $\alpha_1$ and increase $\alpha_2$ so that the skewness of the Beta distribution increases. The midpoint where $\alpha_1 = \alpha_2 = 1$ and so the skewness equals 0, corresponds to the symmetric Uniform distribution. After $\alpha_1 = \alpha_2 = 1$, the distribution becomes positive-skewed (right-skewed) as we

---

5 Note that setting $E[X] = 0$ yields $\frac{\beta}{\lambda} = -\frac{\alpha_2}{\alpha_1}$, which together with $Var(X) = \sigma^2$ yields $\beta^2 = \frac{\sigma^2 \alpha_1 (\alpha_1 + \alpha_2 + 1)}{\alpha_2}$. Hence the values of $\alpha_1$, $\alpha_2$, $b$, and $\lambda$ are not fully determined and we arbitrarily set $\alpha_1 + \alpha_2 = 2$ and change $\alpha_1$ and $\alpha_2$ accordingly, which also changes $b$ and $\lambda$.

6 Recall that the skewness of Beta distribution is given by $\frac{2(\alpha_2 - \alpha_1)(\sqrt{\alpha_1 + \alpha_2})}{(\alpha_1 + \alpha_2)(\sqrt{\alpha_1 + \alpha_2})}$. 

---

24
decrease $\alpha_1$ and increase $\alpha_2$ and we continue until $\alpha_1 = 0.5$ and $\alpha_2 = 1.5$ which corresponds to a skewness of $+1$. We plot the percentage cost error and staffing errors of $U$ and the other policies in Figure 6.4 and Figure 6.5 respectively. Tables B.6-B.8 in EC provide further details (the exact costs and staffing levels as well as the shape parameters of the Beta distribution).

![Figure 6.4](image-url)  
(a) low CV ($CV_\Lambda = 0.0144$)  
(b) moderate CV ($CV_\Lambda = 0.0722$)  
(c) high CV ($CV_\Lambda = 0.1299$)

Figure 6.4: % cost error for varying skewness levels at three levels of arrival rate uncertainty

![Figure 6.5](image-url)  
(a) low CV ($CV_\Lambda = 0.0144$)  
(b) moderate CV ($CV_\Lambda = 0.0722$)  
(c) high CV ($CV_\Lambda = 0.1299$)

Figure 6.5: staffing error (i.e., $N^{opt} - \triangle$ where $\triangle \in \{N_U, N_D, N_{NV}\}$) for varying skewness at three levels of arrival rate uncertainty

From Figure 6.4 and Figure 6.5, we first see that $U$ continues to perform well under asymmetric arrival rate distributions and across varying levels of skewness. In line with our observations for the uniform distribution, $D$ does not perform well except for low CV values. On the other hand, the newsvendor policy performs well for high levels of variability while its performance deteriorates for lower levels of variability and in particular left-skewed distribution. This is because the newsvendor staffing is given by $N_{NV} := \lambda + \beta_2^* \sqrt{\lambda}$ for $\beta_2^* := F_X^{-1}(\frac{p-c_p}{p})$, which decreases as the skewness decreases. Hence, the newsvendor staffs less as skewness decreases and the distribution gets more left-skewed, as seen in Figure 6.5. Therefore the newsvendor performs worse when the distribution is left-skewed because its understaffing is more severe, resulting in higher abandonment and routing control costs.

### 6.5 Discussion

Our numerical results in Sections 6.1 - 6.4 show that $U$ is extremely robust, and achieves close to minimum cost over a large range of parameters and assumptions on the amount of the arrival rate
uncertainty, and its distribution form. In fact, in virtually all of our experiments U outperformed D and NV and achieved a cost that was very close to the true optimal cost. To better understand the reason for the extremely robust performance of U, we compare the actual expected cost to the diffusion approximation of the expected cost, for a wide range of staffing levels. We keep the staffing level N fixed, and approximate the cost using the expression that appears in the limit in Theorem 1. However, we do not assume that the form of the arrival rate uncertainty is consistent with (4.1) (as assumed by Theorem 1). Specifically, for 

\[ \beta = \frac{N - \lambda}{\sqrt{\lambda}} \text{ and } X = \frac{\Lambda - \lambda}{\sqrt{\lambda}}, \]

it follows that the limiting diffusion cost \( c_{\beta} + E \left[ \hat{z}(\beta - X, \hat{T}^*(\beta - X)) \right] \) when re-scaled gives the following approximation for the actual cost:

\[
c_{\lambda} + E \left[ z^{opt}(N, \Lambda) \right] \approx c_{\lambda} + \sqrt{\lambda} \left( c_{\beta} + E \left[ \hat{z}(\beta - X, \hat{T}^*(\beta - X)) \right] \right) = c_{\lambda} + \sqrt{\lambda} E \left[ \hat{z}(\beta - X, \hat{T}^*(\beta - X)) \right]. \tag{6.2}
\]

Figure 6.6 demonstrates numerically that the approximation (6.2) is very accurate, far beyond what is proven in Theorem 1. Specifically, Figure 6.6 plots the actual expected cost (the left-hand side of (6.2)) and the re-scaled diffusion cost (the right-hand side of (6.2)), and the difference between the two. It is clear that the approximation in (6.2) is extremely accurate, over a wide range of staffing levels. This helps explain the robustness in the performance of U.

Figure 6.6 suggests that the performance of square root safety staffing policies can be approximated well without making special assumptions on the limit regime; that is, there is a “universal” approximation. This is because we do not restrict \( \beta \) values to a particular range which can also be evidenced from the \( \beta^* \) values that we observed in our numerical studies in Sections 6.1-6.4. In particular, the \( \beta^* \) values in Figures 2-4 come from a wide range from -12 to 10 (see Table ?? for details). Recalling that \( \sqrt{\lambda} = 10 \) in Figures 2-4, we observe that such extremely low or high values of \( \beta^* \) that are essentially of the same order of magnitude as \( \sqrt{\lambda} \) (so that the resulting safety staffing is in fact of order \( \lambda \)) allow us to capture heavily overloaded or underloaded systems, and thus allows us to approximate parameter regimes beyond what is assumed in our asymptotic analysis in Section 4.

Our observation is also consistent with the universal approximation result of Gurvich et al. (2014) for a M/M/N+M model with deterministic arrival rate and no routing control. Although it is tempting to think that Gurvich et al. (2014) can be used to explain Figure 6.6 the modeling
generalization from a deterministic to a random arrival rate is not immediate, even if we do not allow for outsourcing. Our goal is to show that

$$E \left[ z_{\tau(\infty)}(N, \Lambda) \right] - \sqrt{\lambda} E \left[ \hat{z} \left( \frac{N - \Lambda}{\sqrt{\lambda}}, \infty \right) \right] = O(1), \quad (6.3)$$

where $\hat{z}(m, \infty)$ is as defined in Lemma 1 when the limit as $\hat{T} \to \infty$ is taken. Under the policy $\tau(\infty)$ that outsources no customers (and so besides staffing only incurs costs through customer abandonment), then $z_{\tau(\infty)}(N, l) = a\gamma_{\tau(\infty)}(l)$, and so it follows from Corollary 1 in Gurvich et al. (2014) and algebraic manipulation that

$$z_{\tau(\infty)}(N, l) - \sqrt{l} \hat{z} \left( \frac{N - l}{\sqrt{l}}, \infty \right) = O(1). \quad (6.4)$$

Although it seems reasonable that a technical argument would enable us to show that (6.4) implies

$$E \left[ z_{\tau(\infty)}(N, \Lambda) \right] - E \left[ \sqrt{\Lambda} \hat{z} \left( \frac{N - \Lambda}{\sqrt{\Lambda}}, \infty \right) \right] = O(1),$$

it is not clear how to establish that

$$\sqrt{\lambda} E \left[ \hat{z} \left( \frac{N - \Lambda}{\sqrt{\lambda}}, \infty \right) \right] - E \left[ \sqrt{\Lambda} \hat{z} \left( \frac{N - \Lambda}{\sqrt{\Lambda}}, \infty \right) \right] = O(1), \quad (6.5)$$

Intuitively, we expect something like (6.5) to be true because as the system becomes more understaffed $\hat{z}(\cdot, \infty)$ starts to look linear and as the system becomes more overstaffed $\hat{z}(\cdot, \infty)$ should be negligible (and when the staffing is neither severely understaffed or overstaffed we can appeal to Theorem 1). However, the analytic argument is not straightforward, and remains as an open question.

![Figure 6.6: Expected Cost Difference: Actual vs Diffusion. $[\Lambda \sim \mathcal{U}[20, 780], \lambda = 400, c = 0.1, p = 1$ and $a = 5].$](image)

### 7. Conclusions

This paper deals with the joint staffing and dynamic outsourcing decisions for a co-sourced customer call center. The key feature in our model is the time scales differentiation, where staffing decisions
are made upfront when the arrival rate is uncertain, while the outsourcing decisions are made dynamically in real-time with full knowledge of the arrival rate realization as well as the system state. We propose the policy $U$, which staffs the call center with the mean offered load plus a safety staffing that is of the order of the square root of this mean offered load. The policy also outsources calls whenever the number of customers in line exceeds a threshold. Because the policy $U$ is analytically tractable, it can be used to help call centers negotiate outsourcing contracts; in particular, our analysis shows the implications of various values of the per call outsourcing cost.

When the order of magnitude of arrival rate randomness is the same order as the inherent system fluctuations in the queue length (which is on the order of the square root of the realized arrival rate), we show that our proposed policy is asymptotically optimal as the mean arrival rate grows large. Then, we perform an extensive numerical experiment to study the performance of $U$ beyond the regime for which it is proved to be asymptotically optimal. In all of our numerical experiments in which the system has more than a few servers we did not encounter even one case in which the performance of $U$ was not superb. In contrast, our two benchmark policies each have parameter regimes in which their performance can be arbitrarily bad. It is, therefore, the robustness of the performance of $U$ that we would like to stress as the main takeaway from this paper. One does not need to identify the “right” operating regime in order to determine which policy to use. The $U$ policy appears to be a “one-size-fits-all”.

Several important extensions are worth pursuing. One is to model the staffing decisions of the outside vendor more explicitly, as in Gurvich and Perry (2012) in the case of known arrival rate. In our model we assume that a call outsourced incurs a fixed cost no matter how many calls are sent to the outside vendor. This is consistent with the assumption made in Akşin et al. (2008), and is equivalent to assuming either that the outside service provider has ample service capacity or that it pools demand from a large enough client base so that the calls the company sends to the outside vendor do not have much impact. For the case where the outsourcing is not preferred or does not exist, we see that the performance of $U$ remains superb. (We do not report these results due to space limitations.) However, if the outsourcing capacity is positive and limited we do not know how $U$ will perform, or how it would need to be modified to maintain this superior performance.

Beyond the co-sourcing application considered in this paper, one might use a similar framework to examine joint staffing and control decisions with other system topologies and other types of control such as the problems considered in Gurvich et al. (2008), Dai and Tezcan (2008), Tezcan and Dai (2010), Gurvich and Whitt (2010), and Armony and Mandelbaum (2011). In particular, it will sometimes be possible to incorporate arrival rate uncertainty without changing the control
that has been proven to be asymptotically optimal when the arrival rate is known. Ideally, the robustness of the policy performance with respect to the assumptions on the arrival rate uncertainty in this model will be true in much more generality.

Finally, our numerical results with respect to the remarkable robustness of $U$ suggest that there might be an underlying theoretical justification to this robustness, in the spirit of the universal approximation of [Gurvich et al. 2014]. In Section 6.5 we have discussed why their results in a framework that assumes a known arrival rate and no dynamic control may not be readily applied to our framework. But one wonders whether similar universal approximation principles apply when a random arrival rate and dynamic control are incorporated into the model.

References


