Staffing Service Centers Under Arrival-rate Uncertainty

Jing Zan
Zilliant, Inc., 3815 S. Capital of Texas Highway, Suite 300 Austin, TX 78704,
jzan@utexas.edu

John J. Hasenbein
Graduate Program in Operations Research and Industrial Engineering, Department of Mechanical Engineering, University of Texas at Austin, Austin, TX 78712,
jhas@mail.utexas.edu

David P. Morton
Graduate Program in Operations Research and Industrial Engineering, Department of Mechanical Engineering, University of Texas at Austin, Austin, TX 78712,
morton@mail.utexas.edu

Vijay Mehrotra
University of San Francisco, School of Management, 2130 Fulton Street San Francisco, CA 94117, vmehrotra@usfca.edu

Abstract

We consider the problem of staffing service centers with quality-of-service (QoS) constraints. In our work, we focus on doubly stochastic service center systems; that is, we focus on solving service center staffing problems when the arrival rates are uncertain in addition to the inherent randomness of the system's inter-arrival times and service times. We introduce formulations that handle staffing decisions made over two adjacent decision periods (stages). In our models, we minimize the staffing costs over two decision stages while satisfying a service quality constraint on the second stage operation. A Bayesian update is used to obtain the second-stage posterior arrival-rate distribution based on the first-stage prior arrival-rate distribution and the observations in the first stage. The problem considered in this paper is a single-class single-station service center with random arrival rate. A two-stage stochastic recourse formulation is built to analyze the relationship between the staffing decisions over the two periods. After reformulation, we show that our two-stage model can be rewritten as a newsvendor model. We then provide an algorithm which solves the two-stage staffing problem under several commonly used QoS constraints.

1 Introduction

Service centers, which handle more than 70% (Borst et al. [4]) of the customer-business interactions in the US, have been viewed as the modern business frontier. With estimated annual expenditures exceeding $300 billion (Gilson and Khandelwal [6]), the service center industry has received increased attention from both business and from the operations research community. With 60-80% (Aksin et al. [1]) of the service center operating costs coming from labor costs, service center managers are tempted to reduce the number of servers so as to cut labor costs. However, doing so may risk quality of service, such as making customers wait too long, causing them to either abandon the system or fume over poor service. Such outcomes may incur penalty costs (for third-party providers) or damage the corporate

image. This naturally gives rise to an interest in finding an optimal staffing policy to attain the desired trade-off between service quality and operational efficiency.

Service center systems are stochastic systems because they contain random elements like: the arrival of customers; the time it takes agents to serve customers; and, the time before a customer abandons the system. Queueing models are usually used to represent such stochastic systems. In recent work, researchers have begun to realize the importance of incorporating arrival-rate randomness into the model formulations. Bassamboo et al. [2] consider a fluid approximation for multi-class, multi-type call centers. They use a linear-programming based method to solve for an asymptotically optimal staffing and routing policy that minimizes the staffing cost and abandonment penalty. Bassamboo and Zeevi [3] extend the work in [2], using a data-driven method that provides the optimal staffing level without knowing the probabilistic structure of the arrival rates. Also using a fluid approximation, Gurvich et al. [7] build a chance-constrained formulation, which yields the staffing and routing policy for multi-class, multi-type call centers with arrival rate uncertainty. Their procedure provides a feasible solution that is nearly optimal. In our work, we do not appeal to fluid limits and instead employ queueing models with full short-time-scale stochastic dynamics to formulate service center staffing problems.

In the works we sketch above, a single staffing decision is made in a model that, once the arrival rate uncertainty is revealed, operates in steady state. Such a model does not account for the level of adaptivity that exists in some systems. For this reason, we also introduce a model that allows us to adjust staffing levels in, say, two adjacent four-hour time stages. In doing so, it is important to capture costs incurred for increasing or decreasing the level of the workforce over these time scales. In our work, we apply stochastic programming with recourse to model the staffing decisions over two adjacent time periods. Robbins and Harrison [9] formulate a service center problem as a two-stage mixed integer stochastic program to combine the staffing policy and staff scheduling decision, the decision made after the staffing policy is made, into a single optimization program. Gans et al. [5] propose an approach to include arrival-rate updates, again using a two-stage stochastic program with recourse. In both [5] and [9], the authors focus on the service center scheduling problem or the scheduling problem nested within the staffing problem. In the following of the paper, we first consider the situation where we assume the staffing decision for the first decision stage has been made, and focus on the relationship between the optimal second stage staffing decision and the observations from the first stage. Then, we consider the situation where the first stage staffing decision is not given and needs to be made while taking into consideration the expected second stage staffing cost.

2 Two-stage Staffing Problem with Given First-stage Staffing Decision

We consider the problem of staffing a service center with a single class of customers and a single type of agent under a quality-of-service (QoS) constraint. The queueing model we use to represent such a service staffing problem is an $M/M/n$ model. We further assume the system we study has a stochastic *arrival rate*. That is, we assume that arrivals to the system occur according to a doubly stochastic Poisson process. In operating the service center over two time periods (stages), we assume that: (i) the distribution of the arrival rate for the first stage is known or has been previously estimated; (ii) the staffing level for the first stage, x_1, is given at the beginning of the first stage; and, (iii) the number of customers who arrive during the stage, n, is observed. We update the distribution of the arrival rate for the second stage based on n and then pick the staffing level, x_2, for stage two based on the updated distribution. Figure 1 illustrates these time dynamics.

The service center's manager has two competing concerns. First the manager is concerned with the staffing cost for the second stage (we do not consider the cost for the first stage here, since the staffing

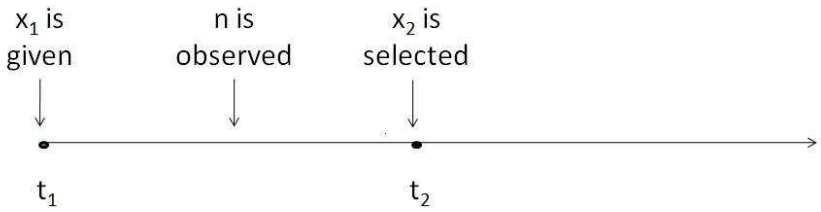

Figure 1: Time Dynamics of the Problem when x_1 is Given

level for the first stage is given), and hence would tend to hire as few servers in the second stage as possible. Second, the manager is concerned with service quality, which will be poor if an insufficient number of servers are hired. In this section, we use the function $\alpha(x_2,\lambda)$ to represent any quality service metric which depends on x_2 and λ, for example, this function could be the probability that a customer must wait, under a second period staffing level x_2 given arrival rate λ. We use Λ to denote the arrival rate as a random variable, and use λ to denote a deterministic value. Without loss of generality we assume that each server has unit service rate.

Let c be the unit staffing cost, c^+ be the unit staffing cost for additional service agents, c^- be the unit salvage cost for sending unneeded service agents home and ε, which takes a value between the minimal and maximal possible values of service quality, be the service quality level threshold. Let $F_\Lambda(\lambda)$ be the CDF of the random arrival rate Λ, and $\alpha(x_2,\lambda)$ be the value of the QoS metric, conditioned on $\Lambda=\lambda$. The optimization model that minimizes staffing costs subject to the QoS constraint is then:

$$\min_{x_2\geq 0} \quad cx_1 + c^+(x_2-x_1)^+ - c^-(x_1-x_2)^+ \tag{1a}$$

$$\text{s.t.} \quad \int_0^\infty \alpha(x_2,\lambda)dF_\Lambda(\lambda) \leq \varepsilon. \tag{1b}$$

The integral in the QoS constraint in (1) simply gives the unconditional value of this QoS metric.

In our call volume forecasting model we assume that the prior distribution for Λ is $gamma(\alpha,\beta)$. The first period calls are then observed and used to produce an updated estimate for the distribution of Λ, i.e., the posterior distribution which is used in the second period. Since gamma is a conjugate prior when a Poisson likelihood function is used, the posterior distribution for Λ is also gamma. In particular, assume the prior distribution for the call volume Λ is $gamma(\alpha,\beta)$ with probability density function

$$g_1(\lambda_1;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_1^{\alpha-1}e^{-\beta\lambda_1} \quad \text{for} \quad \lambda_1 \geq 0.$$

After observing n arrivals over $l \in \mathbb{R}_+$ minutes in the first stage, we obtain the estimated arrival rate distribution for the second stage (the posterior distribution), which is $gamma(\alpha+n,\beta+l)$ with density function

$$g_2(\lambda_2;n,\alpha,\beta,l) = \frac{(\beta+l)^{(\alpha+n)}}{\Gamma(\alpha+n)}\lambda_2^{\alpha+n-1}e^{-(\beta+l)\lambda_2} \quad \text{for} \quad \lambda_2 \geq 0.$$

To focus on the dependency of the second stage optimal staffing level on the number of observed arrivals in the first stage, in our problem, we assume l is fixed. Thus, to simplify the notation, we eliminate l

from the parameter set of the posterior distribution, and denote its density function as $g_2(\lambda_2; n, \alpha, \beta)$ In this case, model (1) can be written as:

$$\min_{x_2 \geq 0} \quad cx_1 + c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{2a}$$

$$\text{s.t.} \quad \int_0^\infty g_2(\lambda; n, \alpha, \beta)\alpha(x_2, \lambda)d\lambda \leq \varepsilon. \tag{2b}$$

Numerical Examples. To investigate the properties of the second-stage optimal solution, we solve the problem using various parameters in the prior distribution. Let $x_2^*(n; \alpha, \beta)$ denote the optimal second-stage staffing level as a function of n for the parameter set (α, β). In the experiments, we use the probability that a customer must wait for service as the service quality measurement. That is we assume

$$\alpha(x_2, \lambda_2) = \mathbb{P}(wait > 0 \mid x_2, \Lambda_2 = \lambda_2).$$

We use the Jagers-van Doorn continuous extension of the Erlang-C formula [8], that is

$$\alpha(x_2, \lambda_2) = \left[\lambda_2 \int_0^\infty te^{-\lambda_2 t}(1+t)^{x_2-1}dt\right]^{-1}.$$

The prior distribution for Λ is such that $\mathbb{E}\Lambda = \alpha/\beta$ and Var $\Lambda = \alpha/\beta^2$. In the experiments, when we vary α and β, we want them vary in such a way that α/β is fixed while α/β^2 is varied. The prior distribution is more concentrated about the mean $\mathbb{E}\Lambda = \alpha/\beta$ as the variance α/β^2 shrinks. Figure 2 shows the plot of $x_2^*(n; \alpha, \beta)$ versus n for different sets of (α, β). The figure depicts the solutions of (2) for parameter sets $(\alpha, \beta) = (2.5, 0.5)$, $(5, 1)$, $(10, 2)$, $(25, 5)$. All the experiments in section 2 are performed on a PC with Intel Core Due CPU P9600 processors at 2.66GHz and 2.67GHz, and 2.00 GB of RAM. We summarize our observations on the numerical results shown in Figure 2 in the propositions and conjecture in the following paragraph.

Characterizing Solutions. Define A as the subset of $\mathbb{R}_+^2$, on which the queueing system is stable. In most applications, an unstable system does not satisfy any reasonable QoS constraint. For example, suppose we consider the problem for a $M/M/n$ system and $\alpha(x, \lambda)$ is the probability a customer waits for service, then the system is only stable when $x > \lambda$. If $x < \lambda$, the stationary waiting time is infinite. Thus for the $M/M/n$ system, we consider quality measurement functions on set $A = \{(x, \lambda) \in \mathbb{R}_+^2 \bigcap \{x > \lambda > 0\}\}$. Before we state our results, we first give some conditions on the service quality measurement function $\alpha(x, \lambda) : A \to \mathbb{R}_+$,

(A1) $\alpha(x, \lambda)$ is a continuous function on A, and

$$\lim_{x \to \infty} \alpha(x, \lambda) = 0, \forall \lambda > 0,$$

and

$$\lim_{\lambda \to 0} \alpha(x, \lambda) = 0, \forall x > 0.$$

(A2) $\alpha(x, \lambda)$ is a continuous function on A, and $\alpha(x, \lambda)$ is strictly decreasing in x for any $\lambda > 0$ and strictly increasing in λ for any $x > 0$.

(A3) $\alpha(x, \lambda)$ is a continuous function on A, and $\alpha(x, \lambda)$ is differentiable in λ on A. $\frac{\partial \alpha(x,\lambda)}{\partial \lambda}$ is strictly decreasing in x for any $\lambda > 0$.

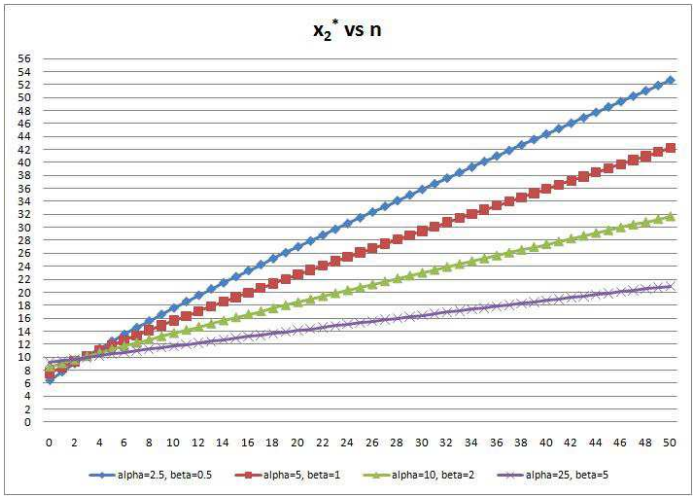

Figure 2: Function $x_2^*(n)$ for Gamma Prior Distribution

(A4) For any service quality level threshold ε,

$$\sup_{x>0} \alpha(x,\lambda) > \varepsilon, \quad \forall \lambda > 0.$$

(A5) The distribution of Λ satisfies $\int_A dF_\Lambda(\lambda) > 0$.

Remark 1. *Notice that $\alpha(x,\lambda)$ represents a QoS metric at arrival rate λ when we have x service agents. Our problem is a bi-criteria problem, the more service agents we have, the higher the staffing cost would be, and the lower the service quality would be. In our model, to control the service quality, we require $\alpha(x,\lambda)$ to be less than some pre-assigned threshold value ε in the constraint. Condition **(A1)** implies that when the arrival rate approaches 0, or when we have a large number of service agents, the service quality approaches the ideal level. Condition **(A2)** indicates that the service quality improves as the number of service agents increases, and deteriorates as the arrival rate increases. Thus, for most commonly used service quality measurements, such as the utilization and the continuous version of the probability a customer waits (given in [8]), conditions **(A1)** and **(A2)** hold. Condition **(A4)** further guarantees the existence of the solution to model* (1).

Remark 2. *When condition **(A2)** holds, the function $\alpha(x,\lambda)$ is strictly increasing in λ for any $x > 0$. Condition (A3) is indicates that as more service agents are added, increased call volumes have a decreasing detrimental effect on the quality of service.*

Proposition 1. *Consider model* (1) *except replace the objective function with $C_{x_1}(x_2)$, where $C_{x_1}(x_2)$ is strictly increasing in x_2, and assume conditions **(A2)**, **(A4)** and **(A5)** hold for $\alpha(x_2,\lambda)$. Then there exists a unique solution to the associated model, denoted as x_2^*, where x_2^* solves $\int_0^\infty \alpha(x_2,\lambda) dF_\Lambda(\lambda) = \varepsilon$.*

Proof. Let $h(x_2) = \int_0^\infty \alpha(x_2,\lambda) dF_\Lambda(\lambda)$. We have $\alpha(x_2,\lambda)$ is strictly decreasing in x_2 for any $\lambda > 0$ on A. Thus **(A5)** implies that $h(x_2)$ is strictly decreasing in x_2. **(A4)** and the continuity of $\alpha(\cdot,\cdot)$ imply the existence of x_2^*. Since C_{x_1} is strictly increasing in x_2 and $\alpha(x_2,\lambda)$ is continuous, the solution to

the optimization model is achieved at the boundary of the feasible region, that is, x_2^* is the solution to $\int_0^\infty \alpha(x_2,\lambda)dF_\Lambda(\lambda) = \varepsilon$. Also x_2^* is unique, since $h(x_2)$ is strictly monotone in x_2. □

Remark 3. *Note that by Proposition 1, x_2^* solves equation*

$$\int_0^\infty \alpha(x_2,\lambda)dF_\Lambda(\lambda) = \varepsilon$$

and hence does not depend on x_1.

Proposition 1 above can be applied to model (1) where function $\alpha(x,\lambda)$ may represent any QoS satisfying **(A2)**, **(A4)** and **(A5)**, and the arrival rate distribution need not be gamma. Proposition 2 and Conjecture 3 below are only for the specified model (2) in this section.

Proposition 2. *Let $x_2^*(n;\alpha,\beta)$ denote the optimal solution to model (2) for the parameter set (α,β), given that n customers are observed in stage 1. Assume **(A1)** - **(A5)** hold for $\alpha(x_2,\lambda)$ and the shape parameter α, in the prior gamma distribution, is a positive integer. Then the optimal solution $x_2^*(n;\alpha,\beta)$ is a strictly increasing function of n for any fixed (α,β).*

Proof. From Proposition 1, given fixed α, β, and n, $x_2^*(n;\alpha,\beta)$ solves

$$\int_0^\infty \alpha(x_2,\lambda)dF_\Lambda(\lambda;n,\alpha,\beta) = \varepsilon,$$

and is unique. Also, we have

$$\begin{aligned}
\int_0^\infty \alpha(x_2,\lambda)dF_\Lambda(\lambda;n,\alpha,\beta) &= \int_0^\infty \alpha(x_2,\lambda)d\mathbb{P}(\Lambda\le\lambda \mid n,\alpha,\beta) \\
&= \alpha(x_2,\lambda)\mathbb{P}(\Lambda\le\lambda \mid n,\alpha,\beta)|_0^\infty \\
&\quad - \int_0^\infty \mathbb{P}(\Lambda\le\lambda)d(\alpha(x_2,\lambda)) \\
&= \int_0^\infty \frac{\partial\alpha(x_2,\lambda)}{\partial\lambda}d\lambda \\
&\quad - \int_0^\infty \frac{\partial\alpha(x_2,\lambda)}{\partial\lambda}\mathbb{P}(\Lambda\le\lambda \mid n,\alpha,\beta)d\lambda \\
&= \int_0^\infty \frac{\partial\alpha(x_2,\lambda)}{\partial\lambda}\mathbb{P}(\Lambda>\lambda \mid n,\alpha,\beta)d\lambda.
\end{aligned}$$

The second-stage arrival rate follows gamma distribution with shape parameter $\alpha+n$ and scale parameter $\beta+l$. Since we assume α is an integer and n is the total number of arrivals in the first stage, which is also an integer, the shape parameter in the posterior distribution is still an integer. Let $G(\cdot \mid \alpha,\beta)$ be the CDF of gamma distribution. When the shape parameter can only take integer values, we have

$$G(\lambda \mid \alpha_1,\beta) > G(\lambda \mid \alpha_2,\beta), \quad \forall\lambda>0,\alpha_2>\alpha_1>0,\beta>0.$$

This implies that in the posterior distribution, $\mathbb{P}(\Lambda>\lambda \mid n,\alpha,\beta)$ is strictly increasing in n for any $\lambda>0$, that is

$$\mathbb{P}(\Lambda>\lambda \mid n_1,\alpha,\beta) < \mathbb{P}(\Lambda>\lambda \mid n_2,\alpha,\beta), \quad \forall n_1<n_2\in\mathbb{Z}_+.$$

Together with condition **(A3)**, we have that for $\forall n_1<n_2\in\mathbb{Z}_+$, $x_2^1<x_2^2$, where x_2^i satisfies

$$\int_0^\infty \frac{\partial\alpha(x_2^i,\lambda)}{\partial\lambda}\mathbb{P}(\Lambda>\lambda \mid n_i,\alpha,\beta)d\lambda = \varepsilon, \quad i=1,2.$$

This implies our result. □

Remark 4. *Notice that in Proposition 2, we require that the shape parameter of the prior distribution, α, take only integer values. We need this to get the dominance condition of the CDF of the posterior gamma distribution. In our application, the meaning of the shape parameter is the number of arrivals observed. Thus, it makes practical sense to assume that the initial shape parameter is a positive integer.*

Conjecture 3. *For any parameter sets (α_1,β_1) and (α_2,β_2), if $\frac{\alpha_1}{\beta_1}=\frac{\alpha_2}{\beta_2}$ and $\frac{\alpha_1}{\beta_1^2}<\frac{\alpha_2}{\beta_2^2}$, then $\frac{\partial x_2^*(n;\alpha_1,\beta_1)}{\partial n}<\frac{\partial x_2^*(n;\alpha_2,\beta_2)}{\partial n}$ for any $n>0$.*

Remark 5. *If we fix the mean of the prior distribution while letting the variance of the prior distribution decrease, the prior distribution is then more concentrated around its mean. Conjecture 3 indicates that if this is the case, then the prior has more weight in the second stage staffing decision.*

We assume the doubly stochastic Poisson process is governed by a gamma distribution. One may be tempted to simplify the problem by using a discrete distribution to model the arrival rate, so as to make the problem easier to solve. However, discretizing the distribution may result in badly behaved solutions, as demonstrated in [10].

3 Two-stage Staffing Problem

Now we start to consider the true two-stage problem. We extend the problem considered in the above section to a two-stage problem, in which the first stage staffing decision, x_1, is also a decision variable. Similar to the problem previous section, we still consider the problem of staffing a service center with a single class of customers and a single type of agents under a quality-of-service (QoS) constraint. Again we assume that arrivals to the system occur according to a doubly stochastic Poisson process and the queueing model we use to represent the staffing problem is an $M/M/n$ model. Considering operating the service center over two time periods, we assume that: (i) the distribution of the arrival rate for the first stage is known or has been previously estimated; (ii) the staffing level for the first-stage, x_1, is selected at the beginning of the first stage; and, (iii) the number of customers who arrive during the first stage, n, is observed. We update the distribution of the arrival rate for the second stage based on n and then pick the staffing level, x_2, for stage two based on the updated distribution. Figure 3 illustrates these time dynamics.

3.1 Model Formulation

We start with the following general two-stage model. Let N denote the number of arrivals in the first stage, and let n represent a realization of N. Then the two-stage model is as follows:

$$\min_{x_1\geq 0} \quad cx_1+\mathbb{E}_N h(x_1,N), \tag{3a}$$

$$\text{where} \quad h(x_1,N|_{N=n}) = \min_{x_2\geq 0} \quad c^+(x_2-x_1)^+ - c^-(x_1-x_2)^+ \tag{3b}$$

$$\text{s.t.} \quad \textit{QoS constraint}. \tag{3c}$$

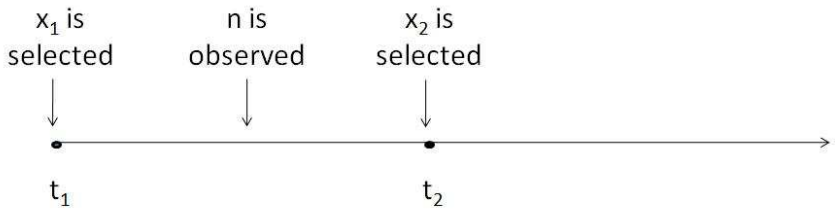

Figure 3: Time Dynamics of the Problem when x_1 is Optimized

It is obvious that the optimal second-stage staffing level x_2^* does not depend on the first-stage staffing level x_1, as long as the QoS constraint is only on the second-stage service quality. The second-stage optimal staffing level x_2^* is affected by the observation from the first-stage, N, since the posterior distribution of the arrival rate depends on N. So x_2^* is a function of N, and thus x_2^* is a random variable, which we denote by $x_2^*(N)$. The specific of function $x_2^*(N)$ is determined by the QoS constraint. The optimal second-stage cost, on the other hand, depends on the value of x_1. This means the optimal first-stage staffing level x_1^* is determined by the distribution of the optimal second-stage staffing level $x_2^*(N)$.

3.2 Two-stage Model with Constraint on Utilization

Now we describe in detail our two-stage model using utilization as the metric in the QoS constraint. In this case, the model is

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N h(x_1, N), \tag{4a}$$

$$\text{where} \quad h(x_1, N|_{N=n}) = \min_{x_2 \geq 0} \quad c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{4b}$$

$$\text{s.t.} \quad \mathbb{P}_{\Lambda_2|_{N=n}}\left(\frac{\Lambda_2}{x_2} < \delta\right) \geq 1 - \varepsilon. \tag{4c}$$

Here, ε and δ are some pre-selected values between 0 and 1.

It is obvious that in (4), $x_2^*(N)|_{N=n}$ is determined only by the second-stage constraint and we have

$$x_2^*(N)|_{N=n} \in \arg\min\left\{x \geq 0 : \mathbb{P}_{\Lambda_2|_{N=n}}\left(\frac{\Lambda_2}{x} < \delta\right) \geq 1 - \varepsilon\right\}.$$

As before, we assume that $\Lambda_1 \sim gamma(\alpha, \beta)$, and we use a Bayesian update to obtain the distribution of Λ_2 after observing N. That is, after observing n arrivals over $l \in \mathbb{R}_+$ minutes in the first stage, we have $\Lambda_2 \sim gamma(\alpha + n, \beta + l)$. Thus, using $F_{\Lambda_2|_{N=n}}(\cdot)$ to represent the CDF of the gamma distribution for Λ_2 given $N = n$, (3.2) becomes

$$x_2^*(N)|_{N=n} \in \arg\min\left\{x : F_{\Lambda_2|_{N=n}}(\delta x) = \frac{\gamma(\alpha + n, (\beta + l)\delta x)}{\Gamma(\alpha + n)} \geq 1 - \varepsilon, x \geq 0\right\}.$$

Let $G_n(\cdot)$ be the CDF of a gamma distribution with parameters $\alpha+n$ and $(\beta+l)\delta$, then we have

$$F_{\Lambda_2|_{N=n}}(\delta x) = G_n(x)$$

and

$$x_2^*(N)|_{N=n} = \lceil G_n^{-1}(1-\varepsilon)\rceil. \tag{5}$$

We re-write (4) with the optimal second stage staffing level, x_2^*:

$$\min_{x_1\geq 0} \quad cx_1 + \mathbb{E}_N[c^+(x_2^*(N)-x_1)^+ - c^-(x_1-x_2^*(N))^+]. \tag{6}$$

Model (6) can be re-written as:

$$\max_{x_1\geq 0} \quad \mathbb{E}_N[-cx_2^*(N) - (c^+-c)(x_2^*(N)-x_1)^+ - (c-c^-)(x_1-x_2^*(N))^+]. \tag{7}$$

Model (7) has the form of a standard newsvendor's problem. Therefore, the solution is given by:

$$x_1^* \in \arg\min\left\{x\geq 0 : \mathbb{P}_N(x_2^*(N)\leq x) \geq \frac{c^+-c}{c^+-c^-}\right\}.$$

Now, we discuss the distribution of $x_2^*(N)$. As mentioned before, x_2^* is a function of N. Thus to obtain the distribution of x_2^*, we need to obtain the distribution of N. Under our assumptions, $\Lambda_1 \sim gamma(\alpha,\beta)$, and $N\sim Poisson(\Lambda_1)$. Use $g(\lambda;\alpha,\beta)$ to stand for the PDF of a gamma distribution with parameters α and β, we have

$$\begin{aligned}
\mathbb{P}(N=n) &= \int_0^\infty g(\lambda;\alpha,\beta)\mathbb{P}(N=n|\Lambda_1=\lambda)d\lambda \\
&= \int_0^\infty g(\lambda;\alpha,\beta)\frac{\lambda^n e^{-\lambda}}{n!}d\lambda \\
&= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\frac{\lambda^n e^{-\lambda}}{n!}d\lambda \\
&= \frac{\beta^\alpha\Gamma(n+\alpha)}{\Gamma(\alpha)n!(\beta+1)^{\alpha+n}}\int_0^\infty \frac{\lambda^{\alpha+n-1}e^{-(\beta+1)\lambda}(\beta+1)^{\alpha+n}}{\Gamma(n+\alpha)}d\lambda \\
&= \frac{\beta^\alpha\Gamma(n+\alpha)}{\Gamma(\alpha)n!(\beta+1)^{\alpha+n}}.
\end{aligned}$$

Notice that in the above formula, if α is a positive integer, then

$$\mathbb{P}(N=n) = \binom{n+\alpha-1}{n}\left(\frac{1}{\beta+1}\right)^n\left(\frac{\beta}{\beta+1}\right)^\alpha.$$

This implies that N has a negative binomial distribution with parameters α and $\frac{\beta}{\beta+1}$, when α is an integer. That is $N\sim NegBin(\alpha,\frac{\beta}{\beta+1})$. There are a couple variations of the negative binomial distribution. Here, we are using the version of the negative binomial distribution that counts the number of failures before the αth success. With this version, the PMF of the negative binomial distribution is

$$\mathbb{P}(K=k|p,\alpha) = \binom{\alpha+k-1}{k}p^\alpha(1-p)^k, \quad k\in\mathbb{Z}_+.$$

It is possible to extend the definition of the negative binomial distribution to the case of a positive real parameter α. The PMF for this extended negative binomial distribution is

$$\mathbb{P}(K=k|p,\alpha)=\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)k!}p^{\alpha}(1-p)^{k},\quad k\in\mathbb{Z}_{+}.$$

Denote the CDF of the extended negative binomial distribution as

$$H(k;\alpha,p)=\mathbb{P}(K\leq k|p,\alpha)=\sum_{i=0}^{k}\frac{\Gamma(i+\alpha)}{\Gamma(\alpha)i!}p^{\alpha}(1-p)^{i},\quad k\in\mathbb{Z}_{+}.$$

We have that $N\sim H(\cdot;\alpha,\frac{\beta}{\beta+1})$. We now provide the algorithm for solving the two-stage model (4) by summarizing the content in this section.

---

step 1 **(find n^* corresponding to x_1^*)**
select $n^*\in\arg\min\left\{k\in\mathbb{Z}_+ : H(k,\alpha,\frac{\beta}{\beta+1})\geq\frac{c^+-c}{c^+-c^-}\right\}$;
step 2 **(obtain $x_1^*=x_2^*(N)|_{N=n^*}$)**
select $x_2^*(N)|_{N=n^*}\in\arg\min\left\{x : F_{\Lambda_2|_{N=n^*}}(\delta x)=\frac{\gamma(\alpha+n^*,(\beta+l)\delta x)}{\Gamma(\alpha+n^*)}\geq 1-\varepsilon, x\geq 0\right\}$;

Now, we discuss about the above algorithm in detail. Let $F_{x_2^*(N)}(\cdot)$ be the CDF of $x_2^*(N)$, which is a càdlàg function. We define the generalized inverse of $F_{x_2^*(N)}(\cdot)$. Let $F^{-1}_{x_2^*(N)}(y)=\inf_{x\in\mathbb{R}}\left\{F_{x_2^*(N)}(x)\geq y\right\}$. We want to obtain the smallest x_1^* that satisfies

$$x_1^*\geq F^{-1}_{x_2^*(N)}\left(\frac{c^+-c}{c^+-c^-}\right),$$

or equivalently

$$F_{x_2^*(N)}(x_1^*)\geq\frac{c^+-c}{c^+-c^-}.$$

That is

$$\mathbb{P}(x_2^*(N)\leq x_1^*)\geq\frac{c^+-c}{c^+-c^-},$$

or equivalently

$$\mathbb{P}(N\leq x_2^{*-1}(x_1^*))\geq\frac{c^+-c}{c^+-c^-}.$$

Denote $x_2^{*-1}(x_1^*)$ as n^*. In step 1, we solve for this n^*, and in step 2, we find x_1^* by evaluating function $x_2^*(n^*)$.

The experiments described in this paragraph show the results of solving (4) with various value of α and β using the algorithm described above. In the experiments, we fix α to be 900, and let β decrease from 45 to 10 with a unit decrement. In such a way, the coefficient of variation of the first-stage arrival rate, $\frac{\sqrt{var(\Lambda_1)}}{mean(\Lambda_1)}$, is fixed, while the mean of the first-stage arrival rate, $mean(\Lambda_1)$, varies from 20 to 90. The service quality threshold value, ε, is set to be 0.05. We conducted the experiments in MATLAB 7.11 (64 bit), and it took 0.22 seconds for MATLAB to finish the experiments. All the experiments in section 3 are performed on a PC with Intel Core i7-980 processors at 3.88GHz, and 24.00 GB of RAM.

3.3 Two-stage Model with Constraint on Probability of Waiting

As mentioned before, the QoS constraint can be of any type. When we use constraints other than the utilization constraint appearing in (4), step 1 is the same. However, in step 2, the function $x_2^*(\cdot)$, which is determined by the second-stage constraint in the model, is different, and the level of difficulty in solving the problem with other kinds of QoS constraints depends on the level of difficulty in evaluating the function $x_2^*(\cdot)$.

To illustrate the complexity introduced by applying other types of QoS constraints, we apply model (3) to an $M/M/n$ queueing system with a QoS constraint on the probability of waiting. In particular, we have

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N h(x_1, N), \tag{8a}$$

$$\text{where} \quad h(x_1, N|_{N=n}) = \min_{x_2 \geq 0} \quad c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{8b}$$

$$\text{s.t.} \quad \mathbb{P}_{\Lambda_2|_{N=n}}(\mathbb{P}(wait > 0 | x_2, \Lambda_2) < \delta) \geq 1 - \varepsilon. \tag{8c}$$

In solving (8), the only difference from solving (4) is the function of x_2^* of N. In (8), function $x_2^*(N)$ is determined by finding

$$x_2^*(N)|_{N=n} \in \arg\min \left\{ x : \mathbb{P}_{\Lambda_2|_{N=n}}(\mathbb{P}(wait > 0 \mid x_2, \Lambda_2) < \delta) \geq 1 - \varepsilon, x \geq 0 \right\}.$$

Using the Jagers-van Doorn continuous extension of the Erlang-C formula [8] for $\mathbb{P}(wait > 0 \mid x_2, \Lambda_2)$, we have

$$x_2^*(N)|_{N=n} \in \arg\min \{ x \geq 0 : \mathbb{P}_{\Lambda_2|_{N=n}} \left(\left[\Lambda_2 \int_0^\infty te^{-\Lambda_2 t}(1+t)^{x-1} dt \right]^{-1} < \delta \right) \geq 1 - \varepsilon \}.$$

Because of the complexity of the formula for $\mathbb{P}(wait > 0 \mid x_2, \Lambda_2)$, once we obtain n^* from step 1, it is not as easy as it is for the utilization constraint model to evaluate $x_2^*(n^*)$. Instead of the relativity explicit formula appearing in (5), one needs to apply a line search to perform this evaluation.

In the next set of experiments, we solved (8) with the same set of values on α and β as in the experiments for solving (4). The ε and δ in (8) are both set to be 0.05. As we mentioned above, line searches on x_2 are needed in step 2 of the algorithm. The lower and upper bounds of the line search are set to be 1 and 120, and the tolerance level of the line search is set to be 0.01. We conducted the experiments again in MATLAB 7.11 (64 bit), and it took 1054.91 seconds for MATLAB to finish the experiments. These experiments demonstrate that our algorithm can efficiently solve the two-stage problem for more complex QoS measures using line searches. Although the solution times are obviously much greater than the times needed with the simple utilization metric, they are still quite reasonable.

4 Conclusion

In this work, we build a two-stage stochastic program with recourse to analyze the relationship between the staffing decisions over two adjacent time periods. A Bayesian update is applied to the arrival rate in the second time period once the new observations arrive during the first time period. The model integrates arrival-rate updates and dependence in staffing decisions over two contiguous time periods. The model minimizes the first stage staffing cost and the expected second stage staffing cost while satisfying

a service quality constraint on the second stage operation. The Bayesian update yields the second-stage arrival-rate distribution based on the first-stage arrival-rate distribution and the observations in the first stage. The second-stage distribution is used in the constraint on the second stage service quality. After reformulation, we show that we can rewrite our two-stage model as a newsvendor model. We provide an algorithm that solves the two-stage staffing problem under some commonly used QoS constraints.

References

[1] Z. Aksin, M. Armony, and V. Mehrotra. The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

[2] A. Bassamboo, J. M. Harrison, and A. Zeevi. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3-4):249–285, 2005.

[3] A. Bassamboo and A. Zeevi. On a data-driven method for staffing large call centers. *Operations Research*, 57(3):714–726, 2009.

[4] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

[5] N. Gans, H. Shen, Y. P. Zhou, N. Korolev, A. McCord, and H. Ristock. Parametric stochastic programming models for call-center workforce scheduling. 2009. Working paper.

[6] K. A. Gilson and D. K. Khandelwal. Getting more from call centers. *The McKinsey Quarterly, http://www.mckinseyquarterly.com*, 2005.

[7] I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7):1093–1115, 2010.

[8] A. A. Jagers and E. A. Van Doorn. On the continued Erlang loss function. *Operations Research Letters*, 5(1):43–46, 1986.

[9] T. R. Robbins and T. P. Harrison. A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research*, 207(3):1608–1619, 2010.

[10] J. Zan. *Staffing service centers under arrival-rate uncertainty*. Ph.D. dissertation, The University of Texas at Austin, 2012.