Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

A SEQUENTIAL STATISTICS APPROACH TO DYNAMIC STAFFING UNDER DEMAND UNCERTAINTY

Fatemeh S. Hashemi
Michael R. Taaffe

Department of Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061, USA

ABSTRACT

Service systems are highly dependent on staffing decisions to provide satisfactory quality of service. This paper tackles the problem of decision making under uncertainty pertaining to the source of demand. Regardless of the distribution of the demand, the proposed staffing rule reacts to the requested quality of service to determine the quality of the estimators of the unknown demand-process parameters, as well as making optimal staffing decisions. Theoretical results on the consistency and optimality of the proposed method is illustrated using sequential statistics approaches.

1 INTRODUCTION

Salary-related costs of staffing lead service-system managers to explore methods to guarantee pre-specified Quality-of-Service (QoS) targets without excess staffing. A natural QoS target is to ensure a sufficiently small likelihood of the demand exceeding supply so as to minimize customer abandonments before being served (Whitt 1999). Such staffing problems often assume that beside the QoS constraints, the system parameters such as arrival rate, service rate, and customers' patience are known as well. In real-world applications however, the workforce optimization problems are to be solved in the absence of full knowledge on system parameters, particularly the arrival rates.

In some settings it is reasonable to assume that the stream of arrivals follows a non-homogenous Poisson process. This assumption is advisable when the arrivals to the system act independently of each other, but are functions of either the time-of-day or the day-of-week. When the rate of this process is known, the steady-state fraction of abandonments can be formally calculated and set below the target to meet the QoS constraint. Assuming an unknown arrival rate, forecasting procedures suggest a time-varying point estimate for this rate which can be progressively updated in time. Depending on how volatile the system's environment and the customers' profile are, the amount of error in the estimates would change. Hence in case of such arrival-rate uncertainty, the steady-state fraction of abandonments becomes a random variable and depending on the realization of the arrival rates, it takes different values. Different approaches are presented in the literature to tackle the randomness of QoS constraint. Gurvich, Luedtke, and Tezcan 2010 suggests requiring QoS only for some pre-specified fraction of the arrival rate values, and that leads to a chance-constrained formulation. There the problem of staffing with uncertain demand rates is reformulated to that of finding a solution for a small set of staffing problems with known rates. An alternative formulation is to average the fraction of abandonments over the demand-rate distribution and consider the QoS as to keep its expected value sufficiently small (Bassamboo and Zeevi 2008). Knowledge of the distribution of the forecasting error is also proved to be useful and it greatly simplifies the staffing problem under arrival uncertainty (Chen and Henderson 2001; Whitt 1999; Deslauriers et al. 2007; Maman 2009; Mehrotra, Ozlük, and Saltzman 2010.)

In this paper we tackle the staffing problem assuming that both the arrival rate and the distribution of the error pertaining to the corresponding uncertainty are unknown. In addition we approach this problem in a more dynamic way in the sense that the quality of the point estimates and the amount of sampling effort is determined based on the QoS target and the risk level.

The ensuing section presents motivation captured from methods in the literature. The sequential statistics approach to dynamic staffing and the theoretical results on the consistency and efficiency of the method are provided in Section 3. Section 5 concludes the paper with summary of observations of the proposed method.

2 MOTIVATIONS AND METHODS

Consider the time-varying $GI_t/GI_t/s_t/s_t$ system. Let the system run in $[0,T]$ and assume stationary in $\mathbb{I}_r := [t_r, t_{r+1}]$, $r = 0,1,2,\cdots,q-1$, $t_0 = 0$ and $t_q = T$. This assumption implies that the arrival and service rates are constant within each period. In addition we assume that the length of each period is long enough for the processes to reach steady-state well within the period.

Remark 1 (1) It is beyond the purpose of this paper to determine the length of stationary periods, although there are methods in the literature to tackle heteroskedasticity in staffing problems (e.g. see Whitt 2007 and references therein.)

(2) Letting the service system be a call-center, the time interval $[0,T]$ in our setting could represent a month over which the performance of center is being observed, and each period $\mathbb{I}_r$ may represent a work day.

Suppose that the service processes follow a general distribution with known service rates $\{\mu_1, \mu_2, \cdots, \mu_q\}$. Also assume Poisson arrival processes for which the arrival rates λ_r $(1 \le r \le q)$ are the unknown means of another stochastic processes $\{\Lambda_r : 1 \le r \le q\}$.

In this setting, the goal of staffing problem is to set the staffing level for each stationary period, in order to ensure QoS targets throughout the whole time interval $[0,T]$ except for the initial transient portion of each period. Accordingly we consider to solve the staffing problem within each stationary period. For ease of notation, dismiss the subscripts r, and let $a(s,\lambda)$ be the long run fraction of abandoning customers for $GI/GI/s/s$ system with arrival rate λ. When λ is fixed, letting $R = \frac{\lambda}{\mu}$ denote the offered load to the system, Erlang's loss probability gives the steady-state fraction of abandoning customers as follows.

$$a(s,\lambda) = \frac{R^s/s!}{\sum_{i=0}^{s} R^i/i!}.$$

Let the QoS target be α, and hence it is required to ensure $a(s,\lambda) \le \alpha$. With known λ, square-root staffing rule (Tijms 2003) suggests that the optimal staffing level to guarantee this QoS target is $R + k_\alpha\sqrt{R}$. Here k_α satisfies $\frac{k_\alpha \Phi(k_\alpha)}{\phi(k_\alpha)} = \frac{1-\alpha}{\alpha}$, where Φ and ϕ are the cdf and pdf of standard normal distribution respectively.

When λ is random, different realization of this random variable, yields different abandonment fraction $a(s,\lambda)$, and hence seeking to hold the constraint $a(s,\lambda) \le \alpha$ means to hold the constraint for all realizations. Depending on the distribution of λ this might be impossible or extremely costly. An alternative formulation is to average $a(s,\lambda)$ over the λ-distribution and ensure that the expected value is small enough (Bassamboo and Zeevi 2008). Such average constraint formulation is as follows

$$s^* = \min\{s \in \mathbb{Z}^+ : \mathbb{E}_\lambda[\lambda a(s,\lambda)] \le \alpha \mathbb{E}_\lambda[\lambda]\}. \tag{1}$$

$\mathbb{E}_\lambda$ in (1) denotes the expected value with respect to the distribution of λ, that is, letting $F_\lambda(\cdot)$ be the cumulative distribution function of λ,

$$\mathbb{E}_\lambda[\lambda a(s,\lambda)] = \int_0^\infty \lambda a(s,\lambda) dF_\lambda. \tag{2}$$

In the chance-constrained formulation proposed in (Gurvich, Luedtke, and Tezcan 2010) , given a pre-specified risk level δ for the probability that the constraint $a(s,\lambda) \leq \alpha$ is violated, we set

$$s^* = \min\{s \in \mathbb{Z}^+ : \mathbb{P}_\lambda(\{a(s,\lambda) \leq \alpha\}) \geq 1-\delta\}, \tag{3}$$

where $\mathbb{P}_\lambda(\{a(s,\lambda) \leq \alpha\}) = \int_0^\infty \mathbb{I}\{a(s,\lambda) \leq \alpha\} dF_\lambda$. Letting $\lambda^* = \inf\{x \in \mathbb{Z}^+ : \mathbb{P}\{\lambda \leq x\} \geq 1-\delta\}$, and

$$s(\lambda^*) = \min\{s \in \mathbb{Z}^+ : a(s,\lambda^*) \leq \alpha\}, \tag{4}$$

be the minimal staffing level required to satisfy the abandonment constraint when the arrival rate is λ^*. (Gurvich, Luedtke, and Tezcan 2010) show that $s(\lambda^*)$ is the optimal solution for the chance-constrained formulation (3).

Note that in order to calculate (2) for the formulation (1), and (4) for that of (3), the knowledge of F_λ is required. In real-world applications however, this knowledge is often not provided to us. Within forecasting methods, F_λ can only be estimated using historical data, and the "estimated" F_λ, is usually deemed to be the "true" distribution, hence the error corresponding to the estimation of F_λ is not tackled in proof of optimality of staffing rules (see for example Whitt 1999). In the next section we propose a dynamic routine for setting the staffing level that guarantees (3) with unknown F_λ.

3 DYNAMIC STAFFING RULE VIA SEQUENTIAL STATISTICS

Sequential statistics propose methods to estimate the mean of a population with unknown variance (Nádas 1969; Anscombe 1952; Chow and Yu 1981; Starr and Woodroofe 1969). Recently these methods have found great applications in popular stochastic optimization, sample average approximation, and stochastic trust-region schemes (see for example Byrd et al. 2012; Bayraksan and Pierre-Louis 2012; Hashemi, Ghosh, and Pasupathy 2014; Hashemi 2015). In this paper, we employ sequential statistics in a rather different search for optimality, namely for optimizing the staffing level in service systems. We would see in this paper that the power of parameter free analysis in sequential statistics method is a crucial factor for attaining staffing level optimality in the absence of the most common structural assumptions. In what follows we first rigorously layout the proposed method and then discuss "consistency" and "efficiency" throughout the Sections 3.2 and 3.3.

3.1 Algorithm Listing

For each $i = 0,1,2,\cdots$ let $\{\Lambda_n(i)\}_{n\geq 1}$ be a sequence of independent and identically distributed random variables. Assume for any u and v $(u \neq v)$, $\{\Lambda_n(u)\}_{n\geq 1}$ and $\{\Lambda_n(v)\}_{n\geq 1}$ be iid realizations of Λ with unknown mean λ, unknown variance σ^2, and having moment generating function $\psi(\cdot)$, which is assumed to be finite at $c_0 < \infty$. For each i let $S_n(i)$ be the partial sum of $\Lambda_n(i)$s, and set

$$N_i(\alpha) \quad = \quad \min\{n : n < \alpha^2 S_n^2(i)\}. \tag{5}$$

Define the following estimators for λ:

$$\bar{\Lambda}_{N_i(\alpha)} = \frac{S_{N_i(\alpha)}(i)}{N_i(\alpha)},\ i = 0,1,2,\cdots,$$

and consider the staffing rule as follows:

$$S^* \quad = \quad \min\{s : \alpha^{-1} < \frac{\sum_{i=0}^{s} \mathbf{Z}_i}{\mathbf{Z}_s} \mid N_1(\alpha), N_2(\alpha), \cdots, N_s(\alpha)\}, \tag{6}$$

where $\mathbf{Z}_i = \frac{\bar{\Lambda}^i_{N_i(\alpha)}}{\mu^i i!}$.

The algorithm listing is as follows:

Algorithm 1 Dynamic Staffing Algorithm Listing

Given: k_α, α, μ, and number of replications m.

Dynamic Staffing Rule

Set $ES = 0$ ▷ Initialize average staffing size.
Set $i = 1$. ▷ Initialize iteration number.
while $i < m$ **do**
    Set $s = 1$ ▷ Initialize service size.
    Set $\bar{\Lambda} =$ ADAPTIVESAMPLING(α) ▷ Derive the sampled estimator for $s = 1$.
    Set $Z = \frac{\bar{\Lambda}^s}{\mu^s s!}$
    Set $sum = 1 + Z$ ▷ Initialize $a^{-1}(s, \Lambda)$
    while $\alpha^{-1} \geq sum/Z$ **do**
        Set $s = s + 1$ ▷ Update service size
        Set $\bar{\Lambda} =$ ADAPTIVESAMPLESIZE(α);
        Set $Z = \frac{\bar{\Lambda}^s}{\mu^s s!}$
        Set $sum = sum + Z$
    end while
    Set $s^* = s$
    Set $ES = \frac{i-1}{i} ES + \frac{1}{i} s^*$
    Set $i = i + 1$
end while
Return ES

Algorithm 2 Adaptive Sampling Algorithm Listing

Given: Data population $\Xi = \{\xi_1, \xi_2, \cdots\}$ (representing iid observations from Λ)

function ADAPTIVESAMPLING(α).
    Set $n = 1$
    Set $S_n = \xi_1$
    while $n \geq \alpha^2 S_n^2$ **do**
        Set $n = n + 1$
        Set $S_n = S_n + \xi_n$
    end while
    Return $\bar{\Lambda} = S_n/n$
end function

3.2 Consistency

Consistency of the method is to be illustrated through the analysis of both the adaptive sampling part as well as the dynamic staffing portion of Algorithm 1. Therefore we first prove the following lemma that decomposes the behavior of the sample size (5) for estimating λ, in terms of the QoS target α.

Lemma 1 (i) For any i, given $\alpha > 0$, $N_i(\alpha)$ is well-defined, that is $\mathbb{P}(N_i(\alpha) < \infty) = 1$;
(ii) For any i, with probability one we have $\lim_{\alpha \to 0} \alpha^2 \lambda^2 N_i(\alpha) = 1$;
(iii) For any i, $\lim_{\alpha \to 0} \alpha^2 \lambda^2 \mathbb{E}[N_i(\alpha)] = 1$;

Proof. Part (i), suppose for given α, (5) is not satisfied as $n \to \infty$. Then (5) implies that $0 > \alpha^2\mu^2$, which is a contradiction.

Part (ii), by (5), as $\alpha \to 0$, $N_i(\alpha) \to \infty$ wp1. In addition we have almost surely

$$N_i^{-1}(\alpha) < \alpha^2 (\frac{S_{N_i(\alpha)}}{N_i(\alpha)})^2. \tag{7}$$

Also since (5) is not satisfied with $N_i(\alpha) - 1$ we get

$$N_i^{-1}(\alpha)(\frac{N_i(\alpha)}{N_i(\alpha)-1} \geq \alpha^2 (\frac{S_{N_i(\alpha)-1}}{N_i(\alpha)-1})^2. \tag{8}$$

So part (ii) is concluded from (7) and (8). Part (iii) follows directly by (Nádas 1969). □

Part (i) of Lemma 1 shows that the sample size resulting from (5) is well-defined, meaning that the point-estimate for λ is always computed with finite sample size when α is bounded away from zero. As α approaches zero, part (ii) and (iii) show the growth rate of $N_i(\alpha)$ and its expected value, respectively, in terms of α. These illustrate that the amount of effort to compute the point estimate depends on the required QoS. This is intuitive as with weak quality of service, and hence large value of α, only a course estimate of λ would most likely suffice to find an appropriate staffing level. However when α is close to zero, we must be extremely conservative in order to lower the chance of error pertained to the demand uncertainty.

We now proceed to study the behavior of the algorithm with respect to choosing the optimal staffing level. Noting that $\frac{\mathbf{Z}_s}{\sum_{i=0}^{s}\mathbf{Z}_i}$ is the stochastic analogue of $a(s,\lambda)$ defined in Section 2, we consider the notation of optimality as introduced in (Gurvich, Luedtke, and Tezcan 2010):

Definition 1 Given positive risk level δ, the optimal staffing rule denoted by S^* is the one that satisfies

$$\mathbb{P}(a(S^*,\Lambda) < \alpha) > 1 - \delta. \tag{9}$$

The condition (9) implies that in order for S^* to be an optimal staffing rule, the stochastic analogue of the QoS in terms of the steady-state fraction of abandoning arrivals is "allowed to be violated" in at most a fraction δ of the arrival-rate realizations.

Given this definition of optimality, we also define "consistent staffing rules" as follows.

Definition 2 Under the risk-level δ, a candidate staffing rule S^* is called "consistent" if as $\alpha \to 0$, (9) holds true.

Definition 2 simply states that a "consistent" staffing rule would ensure a sufficiently large staffing level when an extremely high quality of service is required.

In what follows we prove the consistency of the method proposed in this section.

The next theorem shows that using this quality of point estimate, the staffing rule (6) is consistent, where consistency is defined in Definition 2.

Theorem 1 (i) For any $0 < \alpha < 1$, $\mathbb{P}(S^* < \infty) = 1$.

(ii) Consider a risk level δ, and $\lambda > \lambda^* = \inf\{s \in \mathbb{Z}^+ : \Phi(s) \geq 1 - \delta\}$, where Φ denotes the cdf function of a normal random variable. Then as $\alpha \to 0$, (6) is consistent.

Proof. For part (i), suppose for given $0 < \alpha < 1$, $S^* \to \infty$. This means that almost surely we have

$$\alpha^{-1}\mathbf{Z}_\infty \geq \sum_{i=0}^{\infty} \mathbf{Z}_i. \tag{10}$$

Since $Z_i > 0$ for all is, $\sum_{i=0}^{\infty}\mathbf{Z}_i$ can't be finite unless we have a convergent sum with the limit $a > 0$, and $\lim_{i\to\infty}\mathbf{Z}_i = 0$ almost surely. By (10), that means that we must have $a \leq 0$, which is a contradiction to the assumption that $\mathbf{Z}_i$s are positive.

For part (ii), by part (ii) of Lemma 1, as $\alpha \to 0$, for given i, $N_i(\alpha) \to \infty$. Since $\{\Lambda_{N_i(\alpha)}(i)\}_{i\geq 1}$ are iid realizations of Λ, $\{\bar{\Lambda}_{N_i(\alpha)}(i)\}_{i\geq 1}$ are iid as well. Also for given i, by the CLT when $N_i(\alpha) \to \infty$, $\bar{\Lambda}_{N_i}$ is normally distributed. So for small enough α we can consider $\{\bar{\Lambda}_{N_i(\alpha)}\}_{i\geq 1}$ as iid observations of a normal random variable $\bar{\Lambda}$. Recall the notation $a(s,\bar{\Lambda}) = \frac{\mathbf{Z}_s}{\sum_{i=0}^{s}\mathbf{Z}_i}$. As $N_i(\alpha) \to \infty$, $a(s,\bar{\Lambda}) \to \frac{R^s/s!}{\sum_{i=0}^{s}R^i/i!}$ almost surely. This abandonment rate is increasing in the arrival rate, and by $\lambda > \lambda^*$, $a(s,\lambda^*) \leq \alpha$. Hence for small enough α we get

$$\begin{aligned}\mathbb{P}_{\bar{\Lambda}}(\{a(S^*,\bar{\Lambda}) \leq \alpha\}) &= \int_0^\infty \mathbb{I}\{a(S^*,\bar{\Lambda}) \leq \alpha\} dF_{\bar{\Lambda}}, \\ &\geq 1-\delta.\end{aligned}$$

where the last inequality follows by definition of λ^*. Hence the result follows by Definition 2.

□

3.3 Efficiency

In this section we provide an upperbound on the rate at which the expected value of S^* in (6) grows when $\alpha \to 0$.

Theorem 2 Given $N_1(\alpha)$, let $\alpha < \mu c_0$ and $x_0 = \min\{x : (\frac{\mu}{\lambda})^x x! > \mu\sqrt{N_1}\}$. Then

$$\mathbb{E}S^* \leq x_0 + (\psi(c_0))^{\alpha^{-2}}(\frac{e\mu}{\sqrt{\alpha}}). \tag{11}$$

Proof.

$$\mathbb{E}S^* = \sum_{x=0}^{\infty}\mathbb{P}(S > x) = \sum_{x=0}^{\infty}\mathbb{P}(\alpha^{-1}\mathbf{Z}_x > \sum_{i=0}^{x}\mathbf{Z}_i).$$

We have

$$\begin{aligned}\mathbb{P}(\alpha^{-1}\mathbf{Z}_x > \sum_{i=0}^{x}\mathbf{Z}_i) &\leq \mathbb{P}(\alpha^{-1}\mathbf{Z}_x > \mathbf{Z}_1), \\ &= \mathbb{P}(\alpha^{-1}(\frac{\bar{\Lambda}_{N_x}^x}{\mu^x x!}) > (\frac{\bar{\Lambda}_{N_1}}{\mu})), \\ &\leq \mathbb{P}(\alpha^{-1}(\frac{\bar{\Lambda}_{N_x}^x}{\mu^x x!}) > (\alpha\mu\sqrt{N_1})^{-1}, \\ &= \mathbb{P}(\bar{\Lambda}_{N_x} > (\frac{\mu^{x-1}x!}{\sqrt{N_1}})^{\frac{1}{x}}),\end{aligned} \tag{12}$$

For all $x > x_0$, $(\frac{\mu}{\lambda})^x x! > \mu\sqrt{N_1}$, and letting $a_x := (\frac{\mu^{x-1}x!}{\sqrt{N_1}})^{\frac{1}{x}}$, $a_x > \lambda$. Therefore by the large deviation bound (Bucklew 1990) we get

$$\mathbb{P}(\bar{\Lambda}_{N_x} > a) \leq (\phi_x(c_0))^{N_x}\exp(-c_0 a_x N_x). \tag{13}$$

Also by Lemma 1 we have $\limsup_x N_x = \liminf_x N_x = \alpha^{-1}$. Hence by (12) and (13) we get

$$\begin{aligned}
\mathbb{P}(\alpha^{-1}\mathbf{Z}_x > \sum_{i=0}^{x}\mathbf{Z}_i) \leq \mathbb{P}(\bar{\Lambda}_{N_x} > a) &\leq (\psi(c_0))^{N_x}(c_0 a_x N_x)^{-x}, \\
&= (\psi(c_0))^{N_x}(c_0 N_x)^{-x}(\frac{\sqrt{N_1}}{\mu^{x-1}x!}), \\
&= (\psi(c_0))^{N_x}(\frac{\mu\sqrt{N_1}}{(c_0 N_x \mu)^x x!}), \\
&\leq (\psi(c_0))^{\alpha^{-2}}(\frac{\mu\alpha^{-1}}{(c_0\alpha^{-1}\mu)^x x!}).
\end{aligned}$$

Now we have

$$\begin{aligned}
\mathbb{E}S^* &= \sum_{x=0}^{\infty}\mathbb{P}(S > x), \\
&= \sum_{x=0}^{x_0}\mathbb{P}(\alpha^{-1}\mathbf{Z}_x > \sum_{i=0}^{x}\mathbf{Z}_i) + \sum_{x=x_0+1}^{\infty}\mathbb{P}(\alpha^{-1}\mathbf{Z}_x > \sum_{i=0}^{x}\mathbf{Z}_i),
\end{aligned}$$

Therefore, by $\alpha < \mu c_0$

$$\begin{aligned}
\mathbb{E}S^* &\leq x_0 + \sum_{x=x_0+1}^{\infty}(\psi(c_0))^{\alpha^{-2}}(\frac{\mu\alpha^{-1}}{(c_0\alpha^{-1}\mu)^x x!}), \\
&\leq x_0 + \sum_{x=0}^{\infty}(\psi(c_0))^{\alpha^{-2}}(\frac{\mu\alpha^{-1}}{x!}), \\
&= x_0 + (\psi(c_0))^{\alpha^{-2}}(\frac{e\mu}{\sqrt{\alpha}}).
\end{aligned}$$

□

Theorem 2 provides a crude upper bound on the growth rate of the expected value of S^* in terms of α^{-1}. This upper bound is achieved with only assuming that the moment generating function of Λ is finite at c_0. We expect that we can get much tighter upper bounds in the presence of stronger conditions on the distribution of Λ, so as to get $\mathbb{E}S^*$ growing more slowly than geometric in α^{-1}. In particular, we expect that for staffing rules under demand uncertainty, similar to the algorithm proposed in this paper, the efficiency criteria can be identified when compared with best available approximations using known arrival rates. As such, it is appealing to obtain comparable results to that of Square Root Staffing Rules (henceforth SRSR) discussed in Section 2. In the following section we implement Algorithm 1 and in comparison with SRSR, we study the optimality of our method in various settings.

4 Numerical Study

As the algorithm consists of two separate modules for calculating the sample size in terms of QoS and the staffing level, we are interested in analyzing the behavior of the algorithm in each module. Let $f(\alpha) = \alpha^2\lambda^2 N_i(\alpha)$ denote the limiting factor in Lemma 1. As depicted in Figure 1, the deviation of $f(\alpha)$ observations from one, as well as the estimated variance of $f(\alpha)$ approaches zero as α gets close to zero. The mathematical interpretation of such behavior is that $\lim_{\alpha\to 0} f(\alpha) = 1$, almost surely, as is also proved rigorously within Lemma 1.

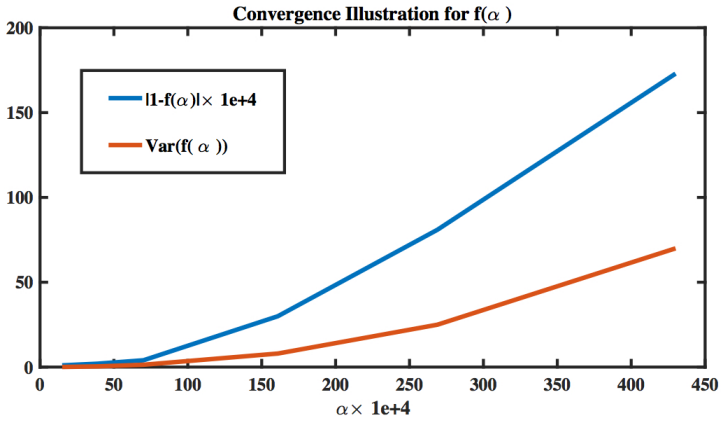

Figure 1: Convergence Illustration for $f(\alpha)$: $f(\alpha)$ goes to one almost surely, as α goes to zero.

Augmented with this sampling behavior, the algorithm computes the sampled estimators for λ and uses it to obtain the staffing level. Accordingly it is appealing to see how close the resulted staffing level is to that of methods under complete information on λ, e.g. SRSR.

Given known arrival and service rates λ and μ, SRSR suggests $s^* = R + k_\alpha \sqrt{R}$. For fixed R, the third column in Table 1 lists the SRSR outputs for s^* across different values of α. As observed in this table, lower values of α require higher staffing level as they dictate higher QoS to the system. Assuming unknown λ, the fourth column of Table 1 shows the resulting values for estimated $\mathbb{E}S^*$ in Theorem 2. Similar to SRSR, the required staffing level grows higher as α decreases. In addition this growth in the values of $\mathbb{E}S^*$ seems to be very slow, and as mentioned in Section 3.3, much tighter upper bounds on the rate of increase is expected than that of (11).

Given the observed deviation between values in the third and fourth columns, the proposed algorithm is shown to perform very close to the optimal s^* (SRSR staffing level solution), while the arrival rate remains unknown. We further support this observation by tracking the deviation between $\mathbb{E}S^*$ and s^*, across different values of R. For $\lambda = 3$ and μ ranges from 0.5 to 2.4, the x-axis in Figure 2 represents $R = \frac{\lambda}{\mu}$. Over different values of R we observe that $\mathbb{E}S^*$ does not deviate from s^* by more than 2.5 while the optimal staffing size stands above 10. Hence we conclude that while the algorithm tends to increase staffing size for small values of α, this increase is not dramatic and the algorithm stays very close to optimal.

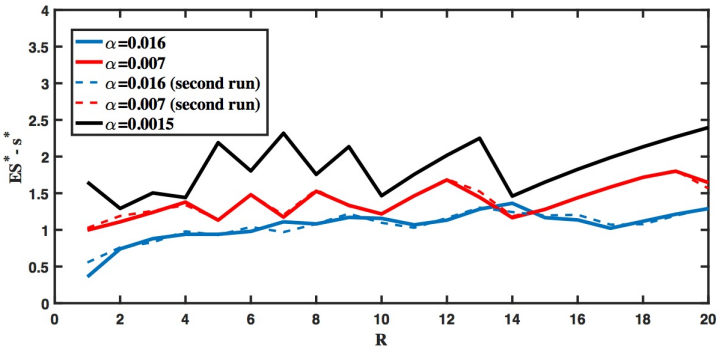

Figure 2: $\mathbb{E}S^* - s^*$ vs the offered load R.

Table 1: $\mathbb{E}S^*$ vs s^* across different values of α.

QoS Parameters			
k_α	α	s^*	$\mathbb{E}S^*$
3	0.0015	13.35	15
2.7	0.0039	12.61	13.98
2.5	0.007	12.12	13.11
2.2	0.0161	11.39	11.91
2	0.0269	10.89	10.79
1.8	0.043	10.4	9.33

5 CONCLUSIONS

This paper introduces a two stage sequential stopping rule to return the number of servers that would satisfy specific QoS target (α) under appropriate risk level (δ). Under the proposed staffing rule, demand rate is estimated "perfectly" only when a high QoS is required; the algorithm accepts coarse estimates only when α is large. This is reached by dynamically balancing the quality of the sampled estimator with the quality of service, and the standard deviation of the estimates go to zero when a high QoS is required. We have shown that when $\alpha \to 0$ the staffing rule remains consistent, by which we mean that the QoS is satisfied with high probability. We have also shown that the upper bound on the expected service size goes to infinity as $\alpha \to 0$, which means that the staffing rule algorithm allows for a large enough number of servers when high quality of service is required. Throughout the results on consistency and efficiency, distribution of the demand need not be known which makes the algorithm a good fit for real-world applications. Empirical simulation studies show that the deviation of the staffing level computed by our method from that of best available method with known arrival rate (square-root staffing rule) is small. For future research we will further explore the efficiency of our method by theoretically deriving the exact rate (rather than an upper-bound) of the expected staffing level in terms of the QoS target, and compare it to the best results in deterministic settings.

REFERENCES

Anscombe, F. J. 1952. "Large-Sample Theory of Sequential Estimation". In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 48, 600–607. Cambridge Univ Press.

Bassamboo, A., and A. Zeevi. 2008. "Staffing Telephone Call Centers Subject to Service-Level Constraints: An Approximate Approach via Constraint Dualization". In *Call Center Forum: Contact Center Management, Philadelphia, PA*.

Bayraksan, G., and P. Pierre-Louis. 2012. "Fixed-Width Sequential Stopping Rules for a Class of Stochastic Programs". *SIAM Journal on Optimization* 22 (4): 1518–1548.

Bucklew, J. A. 1990. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley New York.

Byrd, R. H., G. M. Chin, J. Nocedal, and Y. Wu. 2012. "Sample Size Selection in Optimization Methods for Machine Learning". *Mathematical programming* 134 (1): 127–155.

Chen, B. P., and S. G. Henderson. 2001. "Two Issues in Setting Call Centre Staffing Levels". *Annals of operations research* 108 (1): 175–192.

Chow, Y. S., and K. F. Yu. 1981. "The Performance of a Sequential Procedure for the Estimation of the Mean". *The Annals of Statistics*:184–189.

Deslauriers, A., P. LEcuyer, J. Pichitlamken, A. Ingolfsson, and A. N. Avramidis. 2007. "Markov Chain Models of a Telephone Call Center with Call Blending". *Computers & Operations Research* 34 (6): 1616–1645.

Gurvich, I., J. Luedtke, and T. Tezcan. 2010. "Staffing Call Centers with Uncertain Demand Forecasts: A Chance-Constrained Optimization Approach". *Management Science* 56 (7): 1093–1115.

Hashemi, F. S. 2015. *Sampling Controlled Stochastic Recursions: Applications to Simulation Optimization and Stochastic Root Finding*. Ph.D. thesis, Virginia Tech. Available via https://vtechworks.lib.vt.edu/.

Hashemi, F. S., S. Ghosh, and R. Pasupathy. 2014. "On Adaptive Sampling Rules for Stochastic Recursions". In *Simulation Conference (WSC), 2014 Winter*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 3959–3970. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Maman, S. 2009. *Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments*. Ph.D. thesis, Technion-Israel Institute of Technology. Available via https://ie.technion.ac.il/~serveng/course2004/References/.

Mehrotra, V., O. Ozlük, and R. Saltzman. 2010. "Intelligent Procedures for Intra-day Updating of Call Center Agent Schedules". *Production and Operations Management* 19 (3): 353–367.

Nádas, A. 1969. "An Extension of a Theorem of Chow and Robbins on Sequential Confidence Intervals for the Mean". *The Annals of Mathematical Statistics* 40 (2): 667–671.

Starr, N., and M. B. Woodroofe. 1969. "Remarks on Sequential Point Estimation". *Proceedings of the National Academy of Sciences* 63 (2): 285–288.

Tijms, H. C. 2003. *A First Course in Stochastic Models*. John Wiley and sons.

Whitt, W. 1999. "Dynamic Staffing in a Telephone Call Center Aiming to Immediately Answer All Calls". *Operations Research Letters* 24 (5): 205–212.

Whitt, W. 2007. "What You Should Know about Queueing Models to Set Staffing Requirements in Service Systems". *Naval Research Logistics (NRL)* 54 (5): 476–484.

AUTHOR BIOGRAPHIES

FATEMEH S. HASHEMI is a postdoctoral associate in the Grado Department of Industrial and Systems Engineering at Virginia Tech under supervision of Prof. Michael R. Taaffe. Her research interests are queuing theory, sequential control, and Monte Carlo simulation and optimization. Fatemeh's research finds application in large scale machine learning, stochastic adaptive controls, and system design for dynamical systems. During her graduate studies, she has completed two co-op work terms at IBM Thomas J. Watson Research Center in 2014. She is the recipient of 2014 INFORMS-SIM best student paper prize, 2014 Graduate Research and Development Award at Virginia Tech, 2013 INFORMS-DGWGOR finalist award, 2013 Best Score Award in Research Symposium at Virginia Tech, 2013 ACM-SIGSIM Student Travel Award, and Virginia Tech Research Travel Grant Awards in 2013 & 2014. Her email address is fatemeh@vt.edu and her website is at https://sites.google.com/site/fatemeh1hashemi/.

MICHAEL R. TAAFFE is an Associate professor and former Graduate Program Director in the Grado Department of Industrial and Systems Engineering at the Virginia Polytechnic Institute and State University. He is an active participant in both the applied probability and simulation research communities and has served as President of the *INFORMS College on Simulation* (1998-2000), *Director of the Operations Research Division of IIE* (1997-1998), as well as a *Council Member* and *Newsletter* editor for the (now-named) *INFORMS Applied Probability Society* (1989-1992). Professor Taaffe has been an active referee for many stochastic operations research and systems engineering journals, *NSF* panelist, and associate/area editor for *Operations Research*, *Operations Research Letters*, *IIE Transactions*, and *INFORMS Journal of Computation*. Professor Taaffe served as judge for the *INFORMS Nicholson Prize*, the *George B. Dantzig Prize*, and the *IIE Pritsker Prize* several times. Previous academic and visiting academic appointments were held at Purdue University, Northwestern University, Georgia Tech, the University of Rhode Island, and the University of Minnesota, where he also served as the graduate program director in the Operations and Management Science Department. His Ph.D (1982), is in Industrial and Systems Engineering from The Ohio State University, Columbus, OH. His email address is taaffe@vt.edu.