Managing Supply in the On-Demand Economy: Flexible Workers, Full-Time Employees, or Both?

Jing Dong
Columbia University, 3022 Broadway, New York, NY 10027 jing.dong@gsb.columbia.edu

Rouba Ibrahim
University College London, 1 Canada Square, London E14 5AB rouba.ibrahim@ucl.ac.uk

There are different workforce models in the "gig" economy. While some on-demand service providers rely strictly on either traditional employees or independent contractors, others rely on a blended workforce which melds a layer of contingent workers with a core of permanent employees. In deciding on the "right number of right people to staff at the right time", managers must appropriately weigh the pertinent tradeoffs. In this paper, we study cost-minimizing staffing decisions in service systems where the manager must decide on how many flexible (contractors) and/or fixed (full-time) agents to staff in order to effectively balance operating costs, varying customer demand patterns, and supply-side uncertainty, while not compromising on the quality of service offered to customers. We consider a queueing-theoretic framework where the number of servers is random because part of the workforce is flexible. Since the staffing problem with a random number of servers is analytically intractable, we formulate two problem relaxations, based on fluid and stochastic-fluid formulations, and establish their accuracies in large systems by relying on an asymptotic, many-server, mode of analysis. We derive the optimal staffing policy, and glean insights into the appropriateness of alternative workforce models in on-demand services. We also shed light on the distinction between demand-side (customer arrival rates) and supply-side (number of servers) uncertainties in queueing systems.

Key words: on-demand workforce; sharing economy; random capacity; many-server queues.

1. Introduction

The gig or on-demand economy has gradually become an integral part of the global economy, and it is projected to continue to grow in the coming years (PWC 2017). Naturally, not all on-demand services are delivered in the same manner. For example, ride-sharing applications, such as Uber (*uber.com*) and Lyft (*lyft.com*), rely solely on independent contractors to fulfill ride requests from customers. In such settings, the supply of workers available in each time period is uncertain because those contractors are self-scheduled, i.e., they are free to set their own work schedules. In contrast, several on-demand startups, such as Instacart (*instacart.com*) and Sprig (*sprig.com*), have recently shifted away from staffing a workforce of independent contractors and rely on full-time employees instead. There are also multiple companies, such as Walmart (*walmart.com*) and Netflix (*netflix.com*), which rely on a blended workforce i.e., they meld, as a deliberate business strategy, a layer of contingent workers with a core of permanent employees (Forbes 2015).

Given that diverse landscape of alternative workforce models, a service provider must decide, as a long-term business strategy in an initial planning stage, on the numbers of flexible (contractors) and/or fixed (employees) agents to staff in order to effectively balance operating costs, varying customer demand patterns, and supply-side uncertainty, while not compromising on the quality of service offered to customers. This is the problem that we address in this paper.

1.1. Modeling Framework

We study a cost-minimizing service provider's staffing problem in the context of a stylized queueing model. We assume that both types of workers, fixed or flexible, have the same processing speeds, i.e., service rates. However, the two types of workers differ in their unit operating costs, required working periods, and show-up probabilities.

Customers are both impatient and delay sensitive. There are multiple working periods, customer demand rates are deterministic yet time-varying, and the agent pool may include either fixed or flexible agents (or both). Hereafter, we will use "agent" and "server" interchangeably. A fixed server is compensated c_{fix} per unit time. If a flexible server is available, then she earns c_{flex} per unit time. That is, c_{fix} and c_{flex} are staffing costs. We assume that the number of fixed servers is constant throughout the horizon[1], while the numbers of flexible servers can vary for different periods. When there are flexible servers in the staffed agent pool, the total number of available servers in that period is random. We provide more details about how we model the randomness in §2.4.

Why is our problem challenging? Since the number of servers in our queueing system is random, we are facing a decision-making problem under parameter uncertainty. Because the optimization problem faced by the system manager is analytically intractable, we rely instead on an asymptotic mode of analysis. In particular, we consider a sequence of queueing systems indexed by the arrival rate, λ, and we allow λ to increase without bound.

At a high level, systems with parameter uncertainty involve two "layers" of variability: (i) stochastic variability, for any given realized value of the underlying uncertain parameter, because interarrival, service, and patience times are random; and (ii) parameter uncertainty, because the parameter itself, here the number of servers, is random. We address our capacity-planning question by considering two alternative problem formulations, which correspond to two regimes, respectively. The first formulation assumes that uncertainty effects dominate stochastic fluctuations. The second formulation assumes that both uncertainty effects and stochastic fluctuations are negligible.

Our modelling approach is close to Bassamboo et al. (2010) who derive optimal staffing policies with uncertain arrival rates. However, it is important to emphasize that *the distinction between*

[1] We also considered an alternative setup where fixed workers are not required to show up in every period and are, instead, subject to a requirement on the minimal number of periods during which they must be available. The main insights that we obtain under either modeling framework are similar.

uncertainty in demand (arrival rates) and uncertainty in supply (number of servers) is not a minor technical point. Indeed, in capacity planning, the appropriate staffing level typically consists of a nominal capacity requirement and an additional capacity hedge against **exogenous** uncertainty. With self-scheduling servers, variability is **endogenous** because the distribution of the random number of available servers depends, itself, on the selected pool size. For example, with endogenous uncertainty, staffing a larger pool could also lead to increased variability and, potentially, a worse service level in the system. Thus, it is unclear, a priori, what the optimal staffing policy should be, and whether it would have a similar structure as with exogenous uncertainty. Indeed, we will demonstrate that the optimal staffing policy in systems with endogenous uncertainty, i.e., in supply, gives rise to **different hedging regimes** than with exogenous uncertainty, i.e., in demand.

1.2. Going Beyond Fluid Approximations

There is an extensive body of queueing-theoretic literature which is devoted to studying the classical capacity-sizing (staffing) problem when all servers are fixed, i.e., they must adhere to given schedules that may change over time; e.g., see Whitt (2007). However, there are only a few papers which consider the staffing problem when some or all servers are flexible, i.e., they are free to choose whether or not to show up to work, so that the system's capacity is uncertain. For the most part, those papers have relied on fluid approximations to study the dynamics of large systems with random numbers of servers; e.g., see Whitt (2006b), Gurvich et al. (2018), and Ibrahim (2018). Thus, a key first-order question to answer is whether there is a need to go beyond such fluid approximations when making capacity-sizing decisions and, if so, then when?

In Figures 1-3, we provide preliminary numerical evidence which illustrates that there is a need to go beyond fluid approximations for systems with a random capacity. We plot (dashed curves) the scaled errors entailed in fluid-based staffing prescriptions, as a function of the arrival rate, for various levels of variability in the number of servers. (Those errors are calculated as the absolute differences between fluid solutions and optimal solutions, divided by the square root of the arrival rate.) The bottom solid curves in the figures correspond to errors entailed in refined stochastic-fluid approximations, which we will describe later on. Clearly, when variability in the number of servers is "large enough", there is a need to go beyond fluid-based prescriptions. The preliminary observations in Figures 1-3 motivate our analysis in the remainder of this paper.

1.3. Main Contributions

At a high level, this paper may be divided into two main parts. In the first part, we consider a system where the agent pool consists solely of flexible servers. And, in the second part, we consider a blended workforce, i.e., where both fixed and flexible servers are allowed. When the workforce consists solely of flexible servers, staffing decisions across the multiple periods may be decoupled so

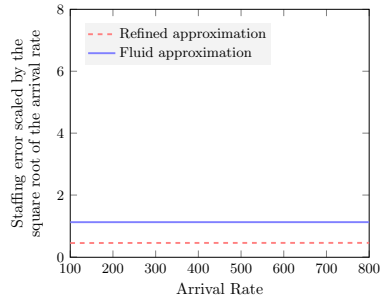

Figure 1 **"Moderate" variability.**

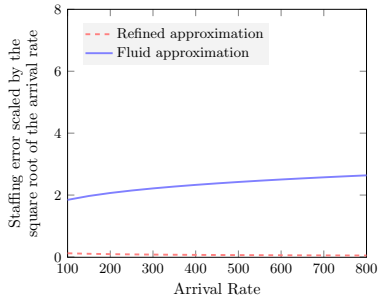

Figure 2 **"High" variability.**

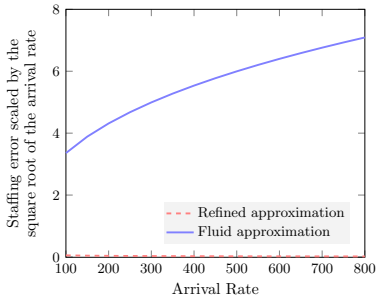

Figure 3 **"Very high" variability.**

that we can focus on systems with a single period instead. In other words, we can make a staffing decision for each period separately. However, when the workforce is blended, we can no longer decouple staffing decisions across the different periods since the fixed servers show up in every period. We consider those two workforce models separately because: (i) they are both prevalent in practice; (ii) focusing on a system with only flexible servers enables us to glean clean insights about the impact of supply-side uncertainty; (iii) the optimal staffing policy with only flexible servers is a key building block in the derivation of the optimal staffing policy with a blended workforce; and (iv) analyzing the blended workforce model provides further insights about the tradeoffs among operating costs, supply-side flexibility (the ability to scale the pool of agents in response to seasonal demand), and supply-side uncertainty. We next summarize our main theoretical and managerial contributions.

Asymptotic results. In systems with a random number of servers, i.e., where at least part of the capacity consists of flexible workers (this covers both the case with flexible servers only, and the case with a blended workforce), we derive optimal staffing policies based on fluid and stochastic-fluid approximations with multiple periods and time-varying demand rates; see Theorem 3, Lemma 2, and Theorem 6. We also rigorously justify the accuracies of those approximations by quantifying their corresponding optimality gap in large systems. In particular, we demonstrate that stochastic-fluid approximations are "extremely" accurate, especially when the magnitude of uncertainty in supply is large; see Theorems 2 and 5.

Optimal staffing policy with flexible servers only. For the optimal staffing policy in a system with flexible servers only, we distinguish among *four regimes*, depending on the magnitude of variability of the random number of servers. Let n denote the expected number of servers, and $\sigma_n = an^q$, for $a > 0$ and $0 \leq q \leq 1$, its standard deviation. The four regimes that we identify are: (i) variability-dominated, for $0 \leq q \leq 1/2$, where there is no concrete benefit from an uncertainty hedge over the regular square-root staffing hedge; (ii) "moderately" uncertainty-dominated, for $1/2 < q \leq 3/4$, where the uncertainty hedge can be embodied in a simple newsvendor-problem-based solution. In this regime, we can *ignore* the dependence between variability in supply and

staffing prescription, i.e., approximating σ_n by $\sigma_{\lambda/\mu}$, where λ is the arrival rate and μ is the service rate, is extremely accurate; (iii) "strongly" uncertainty-dominated, for $3/4 < q < 1$, where there is additional benefit from using an uncertainty hedge which *accounts* for the dependence between variability in supply and staffing prescription. In this regime, using the simple newsvendor-problem-based solution actually leads to a considerable loss in accuracy; and (iv) "extremely" uncertainty-dominated, for $q = 1$, where the uncertainty hedge is on the order of the mean number of servers (in the first three regimes, it is of a smaller order than the mean), so that the system can be either underloaded or overloaded. In this regime, it is also important to account for the dependence between variability in supply and staffing prescription.

Figures 4 and 5 illustrate the regime-dependent optimal staffing policy and demonstrate the difference between exogenous and endogenous uncertainty. Our objective for now is to convey key insights, so we keep our exposition here at a high level. In Figure 4, we consider a queueing system with a random number of servers and a single period. We plot three curves: the optimal staffing policy, i.e., the cost-minimizing prescription for the staffing problem specified in §2, the simple newsvendor-based approximation which ignores the dependence between variability in supply and staffing prescription, and the refined approximation which accounts for that dependence. Figure 4 illustrates that, while the newsvendor-based and refined solutions are almost indistinguishable in the variability-dominated and moderately-uncertainty-dominated regimes, they are considerably different when the level of uncertainty in supply is sufficiently large. For example, for the choice of parameters in the figure and in the strongly uncertainty-dominated regime, the percent error for the newsvendor solution, relative to the optimal solution, is over 20%, while for the refined solution, it is less than 1%. Naturally, the staffing levels in the figures depend on our specific choice of parameters. Nevertheless, we deliberately include those numbers in the figures to illustrate the considerable differences between the alternative staffing prescriptions. In other words, we see that there can be significant loss in accuracy when ignoring the dependence between variability in supply and capacity prescription. However, this is not the case when considering staffing decisions in queueing systems with other forms of parameter uncertainty, e.g., random arrival rates, as we illustrate in Figure 5. In Figure 5, we keep the same parameters as in Figure 4, and consider a random arrival rate which has the same distribution as the random number of servers in Figure 4, while the number of servers is deterministic (nonrandom). It is shown in Bassamboo et al. (2010) that the staffing policy with a random arrival rate gives rise to *two regimes* (instead of the four regimes above), variability-dominated, for $0 \le q \le 1/2$, and uncertainty-dominated, for $1/2 < q \le 1$. In the uncertainty-dominated regime, a simple newsvendor-problem-based capacity prescription is extremely accurate for all values of q; moreover, it is increasingly accurate for more variable demand, i.e., as q increases. These two regimes are illustrated in Figure 5.

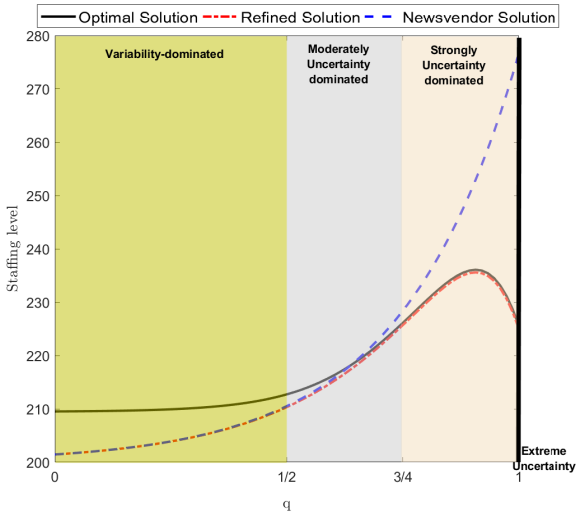

Figure 4 **Optimal staffing policy with a random number of servers $N(n)$ with $n = \mathbb{E}[N(n)]$ and standard deviation $\sigma_n = n^q$.**

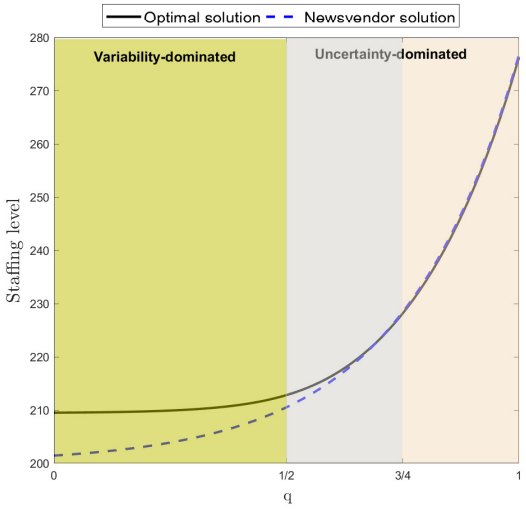

Figure 5 **Optimal staffing policy with a random arrival rate $\Lambda(\lambda)$ with $\lambda = \mathbb{E}[\Lambda(\lambda)]$ and standard deviation $\sigma_\lambda = \lambda^q$.**

Optimal staffing policy with a blended workforce. For the optimal staffing policy with both flexible and fixed servers, we show that when the fixed servers are cheaper than the flexible servers, the optimal staffing policy is to rely solely on the fixed resource in the low-demand periods, and to blend in the high-demand periods. The fixed resource is used to match all or part of the demand, but *not* to hedge against supply-side uncertainty in the system. In contrast, the flexible resource is used to both match the remaining demand in blended periods *and* to hedge against variability in capacity. At a high level, the optimal staffing policy with a blended workforce may be derived by first determining the level of the fixed resource. Then, we consider a staffing problem with flexible servers only, albeit one where the arrival rate in each period is reduced by the service capacity of the fixed resource in that period. Thus, the analysis reduces, in the second step, to determining the staffing policy in a system with flexible servers only.

When supply-side uncertainty is not extreme, which corresponds to $q < 1$ for $\sigma_n = an^q$, the staffing level of the fixed resource may be derived based on a crude fluid approximation of the problem. When supply-side uncertainty is extreme, which corresponds to $q = 1$ for $\sigma_n = an^q$, the level of the fixed resource may differ substantially from its fluid solution. Nevertheless, in this case as well, we can begin by solving for the optimal number of fixed servers, and can then use that solution to leverage our results from the setting with flexible servers only.

Benefits of blending the workforce. To gain a deeper understanding into the benefits of a blended workforce, we investigate, numerically, the impact of blending the workforce on both the cost incurred by the firm and on the quality of service offered to customers. We do so by comparing a blended system to systems where the manager relies strictly on one of the two

resources. We find that blending the workforce usually leads to significant cost reduction for the firm, and to a more "balanced" level of service.

The rest of this paper is organized as follows. We conclude this section with a brief review of the literature. In §2, we present this paper's modeling framework. In §3, we consider the setting with flexible servers only, and derive the corresponding optimal staffing policy. In §4, we study the staffing problem with a blended workforce. We derive the optimal staffing policy and present supporting numerics on the advantage of a blended workforce. In §5, we consider more general (not exponential) distributions for patience time. In §6, we draw conclusions. We relegate all proofs to the Appendices.

1.4. Related Literature and Organization

Our paper is part of the literature on staffing queueing systems under parameter uncertainty; e.g., see Harrison and Zeevi (2005) and Bassamboo et al. (2010). Our paper is also broadly related to the extensive literature analyzing asymptotics of many-server queueing systems with impatient customers (e.g., see Garnett et al. (2002), Zeltyn and Mandelbaum (2005), Whitt (2006a), Bassamboo and Randhawa (2010)), and to the extenstive literature on optimal staffing decisions in service systems (e.g., see Maglaras and Zeevi (2003), Borst et al. (2004), Bassamboo et al. (2005)). However, none of those papers considers a random number of servers. Whitt (2006b) studies staffing decisions in many-server queues with an uncertain arrival rate, an uncertain number of servers, and a single period. Here, we go beyond the fluid approximation of that paper (We have illustrated the importance of doing so in Figures 1-3 above), and study optimal staffing policies in the multi-period setting with both fixed and flexible servers. Atar (2008) derives a diffusion limit for the number of customers in the system with a random number of servers and random service rates. However, the staffing question is not addressed there.

Our work is also related to papers on nurse staffing with absenteeism, such as Green et al. (2013) and Wang and Gupta (2014). It is is also related to the literature on volunteer operations, e.g., harvest gleaning operations, where the the volunteers have uncertain availability; see Ata et al. (2018) and references therein. In this paper, we consider more general show-up behavior, and our asymptotic mode of analysis is different, as well as our consideration of a blended workforce. Azriel et al. (2019) propose a new queueing model, Erlang-S model, for servers which change their availability stochastically. A key difference between our setting and that of Azriel et al. (2019) is the time scale of randomness in supply: In our setting, the randomness is realized at a longer time scale than stochastic fluctuations in the system. In particular, the random number of servers is realized at the beginning of each period, and can be considered to be fixed for that period.

(Stochastic-fluid approximations are extremely accurate, in our setting, because of that difference in time scale.) In contrast, the Erlang-S model assumes that randomness in the number of servers is at the same time scale as stochastic fluctuations.

This paper is most closely related to recent papers on queues with a self-scheduling capacity. Gurvich et al. (2018) were the first to study the operational management of systems with self-scheduling agents. They consider a profit-maximizing firm which has three different levers of agent control at its disposal: the pool size, a cap on the number of allowed agents, and the compensation paid to agents. Ibrahim (2018) studies the capacity-sizing problem with a binomially-distributed number of servers and impatient customers, and proposes using delay announcements as an effective control in such systems. The focus in Gurvich et al. (2018) is different from ours, since that paper does not consider asymptotic accuracy results. On the other hand, Ibrahim (2018) establishes the asymptotic accuracy of fluid approximations, albeit when the number of servers is binomially-distributed. We go beyond both the binomial assumption and fluid approximations in this paper. As we illustrated in Figures 1-3, there may be a need to go beyond fluid approximations, depending on the magnitude of variability in the number of servers. In this paper, we quantify the improvement in accuracy entailed by refining the fluid solution to the stochastic-fluid solution. In §2.2, we demonstrate the importance of going beyond the assumption of a binomially-distributed capacity.

More generally, there is a growing stream of literature on the management of on-demand service platforms, e.g., see Ozkan and Ward (2018), Hu and Zhou (2018), Taylor (2018), and Cachon et al. (2017). Our work compliments that line of literature.

2. Staffing Decisions with a Random Capacity

In this section, we present our modeling framework. We begin by describing our queueing model. Then, we present different models for the distribution of the random number of servers in the system (for now, we do not distinguish between the fixed and flexible resources). We also present empirical evidence illustrating the orders of magnitude of supply-side variability in practice. This evidence motivates our subsequent modeling assumptions and analysis in the paper.

2.1. Queueing Model

When there are flexible servers in the agent pool, the total number of available servers is random. We consider a single-class $M/M/N+M$ queueing model, with a random number of servers N. Service times are independent and identically distributed (i.i.d.) exponential random variables with rate μ. Customers are impatient, and their patience times are i.i.d. exponentially distributed with rate θ. Customers are processed in the order in which they arrive, i.e., we use the first-come-first-served discipline. The total number of servers, N, is a nonnegative integer-valued random variable. It is realized at the beginning of each period. The arrival, service, and abandonment processes are

all mutually independent, also independent of N. Abandonment makes the system stable, even when N is random (Whitt 2006b). Specifically, conditional on a particular realization of N, a proper steady-state distribution always exists. In this paper, we assume that each period is long enough, e.g., relative to the average service time, so that we can focus on steady-state performances. Note that steady-state performance measures can be calculated by conditioning and unconditioning on N.

We assume that there are k periods, and that period i has length T_i. The different periods may correspond to different work shifts in a single day, e.g., morning, afternoon, and evening shifts, or to successive days, weeks, months, etc., depending on the time scale at which the manager decides on her staffing requirements. The arrival rate of the Poisson arrival process in period i is given by λ_i. We fix $\lambda > 0$ and let $\lambda_i \equiv \lambda\xi_i$, where $\xi_i \geq 0$ for each i. We index all relevant quantities by λ, to indicate the dependence on the arrival rates. In our asymptotic analysis, we let λ grow without bound while keeping each ξ_i constant. Note that λ_i's are of the same scale as λ. We also assume, without loss of generality, that the alternative periods are numbered in order of increasing λ_i values, i.e., $\lambda_i \leq \lambda_j$ for $i \leq j$. In other words, we re-index the different periods so that the λ_i values are ordered.

2.2. A Random Number of flexible servers

We assume that the pool sizes of flexible agents may vary across periods, i.e., they can be scaled to meet seasonal demand fluctuations. Let n^i_λ denote the total number of flexible servers scheduled for period i. Let $N_{flex}(n^i_\lambda)$ denote the random number of flexible servers who show up in period i, which depends on n^i_λ. Without loss of generality, we assume that

$$N_{flex}(n^i_\lambda) = \eta_{n^i_\lambda} + \epsilon_{n^i_\lambda}, \tag{1}$$

where $\mathbb{E}[N_{flex}(n^i_\lambda)] = \eta_{n^i_\lambda}$ and $\epsilon_{n^i_\lambda}$ is a random variable with $\mathbb{E}[\epsilon_{n^i_\lambda}] = 0$ and $\mathrm{Var}[\epsilon_{n^i_\lambda}] = \sigma^2_{n^i_\lambda}$. In (1), we ignore the integrality assumptions on n^i_λ and $N_{flex}(n^i_\lambda)$: This is reasonable when the system is large, which is the case of primary interest to us. We also note that the expected queue length expression for Erlang-A queue can be extended to nonnegative real values for the number of servers (Mandelbaum and Zeltyn 2007), i.e. the staffing problem faced by the manager is defined for both integer and non-integer values of n^i_λ.

In this paper, our aim is to characterize the manager's optimal staffing decisions, i.e., what n^i_λ in (1) should be. To be able to do so, we must relate the show-up decisions of flexible workers, in the pool of size n^i_λ, to the distribution of $N_{flex}(n^i_\lambda)$ in (1). Thus, a natural, first-order, question to ask is: which agent show-up model would be appropriate to consider? We hasten to emphasize that there is no single answer to this question, since different agent show-up models may be appropriate

depending on the specific application context in mind. Thus, we do not attempt here to propose a single model to "fit all" settings. Instead, we explore how different agent-participation models, which may emerge in practice, impact the distribution of $N_{flex}(n_\lambda^i)$. We are especially interested in quantifying the order of magnitude for $\sigma^2_{n_\lambda^i}$ as a function of n_λ^i. This is because, as we will demonstrate later, this order of magnitude affects the structure of the optimal staffing policy.

2.3. Binomial Model and Extensions

For a starting point, we rely on the existing literature. With self-scheduling agents, one natural model to consider is the classical Binomial model where agents are assumed to make their joining decisions independently of each other, and with a constant joining probability, e.g., see Ibrahim (2018), Gurvich et al. (2018), and Ata et al. (2018). In this section, we begin by discussing the classical Binomial model and its implications on the variability of the number of available servers. Then, we discuss possible modeling extensions; and justify the need for such extensions by presenting supporting empirical evidence based on data collected from Uber.

For ease of exposition, we focus here on a single period; thus, we drop dependence on i. For $1 \le j \le n_\lambda$, we define the Bernoulli random variable I_j, where $I_j = 1$ if agent j is available for work, and $I_j = 0$ otherwise. Then, $N_{flex}(n_\lambda)$ in (1) can be written as follows:

$$N_{flex}(n_\lambda) = \sum_{j=1}^{n_\lambda} I_j. \tag{2}$$

The classical Binomial model. We begin by assuming that I_j in (2) are i.i.d. Bernoulli random variables with a constant and deterministic success probability p. In this case, it is readily seen that $N_{flex}(n_\lambda)$ has a binomial distribution with $\eta_{n_\lambda} = n_\lambda \cdot p$ and $\sigma^2_{n_\lambda} = n_\lambda \cdot p(1-p)$. In particular, σ_{n_λ} is of order $\sqrt{n_\lambda}$.

In §2.3.1, we present empirical evidence illustrating that variability in practice is typically of a larger order than that implied by the classical Binomial model. Indeed, despite its analytical tractability, that model has several shortcomings; thus, there is a need to consider alternative models as well. For example, it assumes that each agent makes her participation decision independently of other agents. In practice, agent decisions typically exhibit correlations, e.g., because of coordinated joining and leaving decisions facilitated by social-media platforms[2]. Such correlations lead to over-dispersion, i.e., additional variability, compared with the classical Binomial model. For another example, joining probabilities may be neither homogeneous across agents, nor deterministic. Indeed, Chen et al. (2017) provide empirical evidence that each Uber driver faces a hierarchy of random, unforeseen, shocks, e.g., linked to weather conditions or promotional events from competitors. A labor supply decision, for each driver, depends on specific heterogeneous realizations

[2] https://warwick.ac.uk/newsandevents/pressreleases/uber_drivers_are/

of those shocks. Thus, ex-ante, the show-up probability of an agent is, itself, a random variable, which also leads to over-dispersion.

Correlated Bernoulli sequences. We now describe extensions to the classical Binomial model which capture higher orders of variance. We begin by describing a model for capturing correlations between agent joining decisions, i.e., I_j in (2). While modeling correlated Bernoulli sequences is not new, our intention here is not to provide an exhaustive review of the relevant literature. Rather, we show how one intuitively appealing model, the Generalized Binomial model proposed by Drezner and Farnum (1993), could be used to explain different orders of magnitude for $\sigma^2_{n_\lambda}$.

We begin by assuming, without loss of generality, that agent j is the j^{th} agent in the pool of size n_λ to make a joining decision. We define $\bar{I}_j \equiv (1/j)\sum_{k=1}^{j} I_k$ and let $\mathcal{F}_j$ be the σ-field generated by the history $\{I_1, ..., I_j\}$. As in Drezner and Farnum (1993), we assume that

$$\mathbb{P}(I_{j+1} = 1|\mathcal{F}_j) = (1-\alpha)\cdot p + \alpha \cdot \bar{I}_j, \tag{3}$$

for some probability p and $\alpha \in [0,1)$. In other words, (3) assumes that an agent's joining decision is a convex combination of p and the relative frequency of agents who have already joined. In particular, if $\bar{I}_j > p$ $(< p)$, then $\mathbb{P}(I_{j+1} = 1|\mathcal{F}_j) > p$ $(< p)$. That is, the more agents join, the more likely it is that additional agents will join as well. We also note that letting $\alpha = 0$ in (3) allows us to retrieve the classical Binomial model. Heyde (2004) derives asymptotic properties for the variance, $\sigma^2_{n_\lambda}$, implied by (3). In particular, as $n_\lambda \to \infty$, the following is shown to hold[3]:

$$\sigma^2_{n_\lambda} \sim \begin{cases} \frac{p(1-p)n_\lambda}{1-2\alpha} & \text{for } \alpha < 1/2, \\ p(1-p)n_\lambda \log(n_\lambda) & \text{for } \alpha = 1/2, \\ \frac{p(1-p)n_\lambda^{2\alpha}}{(2\alpha-1)\Gamma(2\alpha)} & \text{for } \alpha > 1/2. \end{cases} \tag{4}$$

Based on (4), we note that for $\alpha > 1/2$, σ_{n_λ} is of a larger order of magnitude than $\sqrt{n_\lambda}$, i.e., this model allows for over-dispersion relative to the classical Binomial model. Asymptotic properties of Bernoulli sequences with more general long-range dependence structures can also be found in Romano and Wolf (2000).

A random joining probability. For an alternative extension of the classical Binomial framework, we could also assume that for each period, the joining probability is a random variate drawn from a distribution $\mathcal{P}$ with $\mathbb{E}[\mathcal{P}] = p$. Given $\mathcal{P} = p'$, each agent makes a joining decision with probability p', independently of other agent. In this case, the expected number of agents who are available is $\eta_{n_\lambda} = \mathbb{E}[\mathbb{E}[N_{flex}(n_\lambda)|\mathcal{P}]] = n_\lambda \cdot \mathbb{E}[\mathcal{P}] = n_\lambda p$, and its variance is given by the conditional variance formula:

$$\begin{aligned} \sigma^2_{n_\lambda} = \mathrm{Var}[N_{flex}(n_\lambda)] &= \mathbb{E}[\mathrm{Var}[N_{flex}(n_\lambda)|\mathcal{P}]] + \mathrm{Var}[\mathbb{E}[N_{flex}(n_\lambda)|\mathcal{P}]] \\ &= n_\lambda \cdot \mathbb{E}[\mathcal{P}(1-\mathcal{P})] + n_\lambda^2 \cdot \mathrm{Var}[\mathcal{P}]. \end{aligned}$$

[3] We write $A_n \sim B_n$, if $\frac{A_n}{B_n} \to 1$ as $n \to \infty$.

It is readily seen that if $\text{Var}[\mathcal{P}] > 0$, σ_{n_λ} is on the same order of magnitude as n_λ.

2.3.1. How Much Variability Do We Observe in Practice? We next present some empirical evidence substantiating the orders of magnitude of variability in the random number of servers, that are observed in practice. In particular, we analyze Uber pickup data, which are made publicly available by the website FiveThirtyEight[4], to quantify variability in the numbers of drivers that are on the road at different time epochs.

For the purposes of this section, we use the data set describing daily-aggregated Uber trip statistics in January and February 2015 (59 days). It contains the total numbers of active cars from each base serving New York City, on each day, in that time frame. For each day, we further sum the numbers of active cars across all base codes to obtain the total number of active cars available in the New York City area that day. We treat these numbers as realizations of the random numbers of Uber drivers per day. We note that we do not use more granular data, e.g., at the hourly level, since we do not have access to such detailed data.

In Table 1, we present the average, $\hat{\eta}_n$, and the standard deviation, $\hat{\sigma}_n$, for the number of cars available per weekday (calculated across all data for the same weekday in our sample). We calculate $\hat{q} = \log(\hat{\sigma}_n)/\log(\hat{\eta}_n)$ assuming $\sigma_n = \eta_n^q$. While we focus here on the weekday effect due to data availability, we expect that finer hour-of-day effect would also be quite relevant in describing trends and patterns in the data. Table 1 provides evidence of considerable variability in the data, e.g., the values of $\hat{q}$ are all higher than 0.5, which corresponds to the order implied by the classical Binomial model. This indicates that there is a need to consider larger orders of variability. In this paper, we consider q as high as 1, i.e., the extremely uncertainty-dominated supply. We hypothesize that despite having self-scheduling agents, Uber is probably a conservative example for supply-side variability relative to smaller size and less well controlled platforms.

Weekday	$\hat{\eta}_n$	$\hat{\sigma}_n$	$\hat{q}$
Sunday	7,075	701	0.74
Monday	7,155	707	0.74
Tuesday	7,364	1,639	0.83
Wednesday	8,129	450	0.68
Thursday	8,424	738	0.73
Friday	8,606	1,040	0.77
Saturday	7,976	795	0.74

Table 1 Statistics on the numbers of active Uber cars in New York City in January and February, 2015.

[4] The data are publicly available at `https://github.com/fivethirtyeight/uber-tlc-foil-response`.

2.4. This Paper's Model

In §2.3, we considered alternative agent show-up models which may arise in practice, and illustrated how different modeling assumptions lead to different orders of magnitude for σ_{n_λ}. This is important because, as we will demonstrate in subsequent sections, how σ_{n_λ} scales with η_{n_λ} plays a central role in the optimal staffing policy. On the other hand, the fine detail of those specific agent show-up models are of secondary importance for our purposes. Thus, for subsequent analysis, we consider the simplified model:

$$N_{flex}(n_\lambda^i) = \eta_{n_\lambda^i} + \sigma_{\eta_{n_\lambda^i}} \epsilon_i, \tag{5}$$

for i.i.d. random variables $-1 \le \epsilon_i \le 1$ with $\mathbb{E}[\epsilon_i] = 0$. We assume that ϵ_i has a strictly positive probability density function (pdf), f_ϵ, on $(-1, 1)$. Thus, its cumulative distribution function (cdf), F_ϵ, is invertible on that domain. Note that in this simplified model, the distribution of ϵ_i does not depend on n_λ^i. For the ease of exposition, we also assume the specific form $\sigma_n = an^q$, for some $a > 0$ and $0 < q \le 1$. For $q = 1$, we also impose that $a < 1$ to ensure that $N_{flex}(n_\lambda) \ge 0$. Assume that η_n is strictly increasing in n. Then there is a one-to-one correspondence between n_λ^i and $\eta_{n_\lambda^i}$. This implies that the manager's staffing decision can be equivalently formulated in terms of $\eta_{n_\lambda^i}$. For example, for the three agent models discussed in §2.3, it holds that $\eta_{n_\lambda^i} = n_\lambda^i \cdot p$. By a slight abuse of notation, and for the ease of exposition, we denote hereafter the expected pool size by n_λ^i. That is, we consider the following model, which is equivalent to (5):

$$N_{flex}(n_\lambda^i) = n_\lambda^i + \sigma_{n_\lambda^i} \epsilon_i, \tag{6}$$

where the manager's objective is to determine cost-effective n_λ^i.

3. Capacity Sizing with Flexible Servers

In this section, we consider a workforce with flexible servers only, and study cost-minimizing staffing decisions in this case. The goal is to understand the impact of supply-side uncertainty. In §4, we consider a workforce with both fixed and flexible servers.

3.1. A Long-Term Staffing Problem

In this paper, we consider the long-term staffing question that the manager faces. In practice, managers must make staffing decisions ahead of time to allow for agent training. The timeline of decision-making is as follows: At time zero, i.e., the initial planning stage, the manager makes a staffing decision on the average flexible pool size (the average numbers of flexible agents desired) n^i for each period i. Then, at the beginning of each period i, the staffing level realizes, i.e., we observe a specific realization, $N_{flex}(n^i) = s^i$, which is drawn from the distribution of the random variable $N_{flex}(n^i)$. For the remainder of period i, the system operates like an Erlang-A queue with s^i servers.

We assume that the arrival rate, λ_i, is a deterministic constant for each period. We also assume that λ_i is known, a priori, to the manager. With a deterministic λ_i, the optimal staffing level would remain the same so long as the manager must decide on her staffing level, n^i, before the start of period i [5]. Note that stochastic variation for given model parameters impacts system behavior on the short-time scale, and is thus less important when uncertainty in model parameters is introduced on a longer time scale. This motivates us to look at the stochastic-fluid optimization problem which ignores stochastic variability.

Consistently with Bassamboo and Randhawa (2010) and Bassamboo et al. (2010), we consider two customer-related costs: (i) A delay cost, h, per customer for each unit of time that this customer spends waiting to be served, and (ii) an abandonment penalty cost, r, incurred per customer who abandons before being served. Recall that the per unit time staffing cost is given by c_{flex} for a flexible server. Throughout this paper, we make the following assumption:

ASSUMPTION 1. *We assume that* $c_{flex} < (h/\theta + r)\mu$.

This assumption ensures that the flexible resource is cheap enough to avoid pathological cases where the system manager would not staff from this resource.

We let $Q^i_\lambda(n^i_\lambda)$ and $\xi^i_\lambda(n^i_\lambda)$ denote the steady-state queue length and steady-state rate of customer abandonment in period i. We let $X^i_\lambda(n^i_\lambda)$ denote the steady-state number of customers in the system, in period i, so that:

$$Q^i_\lambda(n^i_\lambda) = (X^i_\lambda(n^i_\lambda) - N_{flex}(n^i_\lambda))^+,$$

where $x^+ \equiv \max\{x, 0\}$. With exponentially-distributed patience times, it is also well known that:

$$\xi^i_\lambda(n^i_\lambda) = \theta \cdot \mathbb{E}[Q^i_\lambda(n^i_\lambda)],$$

where θ is the rate of the patience-time distribution (Mandelbaum and Zeltyn 2007). Letting $\mathbf{n}_\lambda \equiv (n^1_\lambda, \ldots, n^k_\lambda)$, the system manager's staffing problem is given by:

$$\begin{aligned} \min_{\mathbf{n}_\lambda} \quad & \Pi_\lambda(\mathbf{n}_\lambda) \\ & \equiv \sum_{i=1}^{k} T_i \left(c_{flex} n^i_\lambda + h \cdot \mathbb{E}[Q^i_\lambda(n^i_\lambda)] + r \cdot \xi_\lambda(n^i_\lambda) \right), \\ & = \sum_{i=1}^{k} T_i \left(c_{flex} n^i_\lambda + (h + r\theta) \mathbb{E}[Q^i_\lambda(n^i_\lambda)] \right), \end{aligned} \tag{7}$$

[5] In some applications, we could also consider the case where additional information could be gathered to yield improved demand forecasts as we get closer to the start of the period. In that case, we may want to update our staffing decision over time. As flexible servers may be more flexible to call upon in the last minute, these servers will bring an extra layer of benefit. This case is beyond the scope of the current paper.

When employing flexible servers only, the multi-period staffing problem (7) can be fully decompose in to k single-period staffing problems. Particularly, for period i, $1 \leq i \leq k$, the system manager is solving

$$\min_{n_\lambda \geq 0} \Pi^i_\lambda(n_\lambda) \equiv c_{flex} n_\lambda + (h + r\theta)\mathbb{E}[Q^i(n_\lambda)]. \tag{8}$$

For simplicity of exposition, we shall drop the dependence on the period index, i, for the rest of the discussion in this section. We write n^*_λ as the optimal solution of (8).

The problem formulation in (8) is prohibitively difficult to solve in closed form, because our choice of n_λ affects the distribution of the number of servers which, in turn, affects the distributions of the queue-length, $Q_\lambda(n_\lambda)$, and the abandonment rate, $\xi_\lambda(n_\lambda)$. Thus, we first formulate a fluid (ignoring both stochastic variability and parameter uncertainty) relaxation, and then a stochastic-fluid (ignoring stochastic variability only) relaxation of (8). We analyze the structure of the optimal staffing rules based on these two problem relaxations.

3.2. Fluid Approximation

For the fluid relaxation of our problem, we ignore both uncertainty effects and stochastic fluctuations in the system. In particular, the fluid abandonment rate in our problem is given by $(\lambda - n\mu)^+$, which is obtained by substituting the random number of servers, $N_{flex}(n)$, by its expected value, n. This leads to the following fluid relaxation:

$$\min_n \bar{\Pi}_\lambda(n) \equiv c_{flex} n + \left(\frac{h}{\theta} + r\right) \mu \left(\lambda/\mu - n\right)^+. \tag{9}$$

We write

$$\beta \equiv \left(\frac{h}{\theta} + r\right) \mu, \tag{10}$$

which can be interpreted as the performance cost. We denote $\bar{n}_\lambda$ as the optimal solution to (9). As per Assumption 1, we have that $c_{flex} < \beta$. Thus, $\bar{n}_\lambda = \lambda/\mu$, i.e., it is most cost-effective to match the mean supply with the mean demand. Given its simple form, the fluid approximation in (9) is appealing, provided that it does not entail a significant loss in accuracy. We next characterize the optimality gap of (9), in a regime where the arrival rate, λ, is large. To facilitate the asymptotic analysis, we first introduce a few definitions.

Definition 1. Let f and g be two functions defined on some subset of $\mathbb{R}$. Then, as $n \to \infty$,

(a) $f(n) = \mathcal{O}(g(n))$ if there exists $M > 0$ and $C > 0$ such that $|f(n)| \leq M|g(n)|$ for $n \geq C$;

(b) $f(n) = o(g(n))$ if for any $\xi > 0$, there exists $N(\xi)$ such that $|f(n)| \leq \xi|g(n)|$ for all $n \geq N(\xi)$;

(c) $f(n) = \Theta(g(n))$ if there exist $M > 0$, $L > 0$ and $C > 0$ such that $L|g(n)| \leq |f(n)| \leq M|g(n)|$ for $n \geq C$.

We are now ready to state the optimality gap of the fluid approximation.

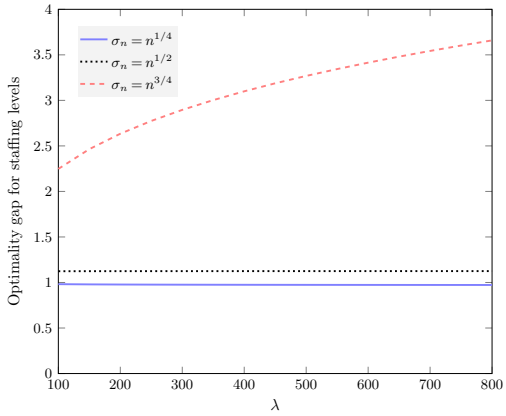

Figure 6 **Errors for optimal fluid staffing levels, $|\bar{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ.**

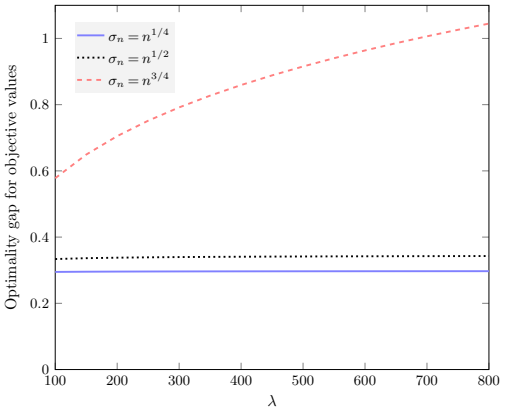

Figure 7 **Optimality gap in objective values, $|\Pi_\lambda(\bar{n}_\lambda) - \Pi_\lambda(n^*_\lambda)|/\sqrt{\lambda}$, as a function of λ.**

THEOREM 1. *For large λ,*

$$\Pi_\lambda(\bar{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}\left(\max\left\{\sigma_\lambda, \sqrt{\lambda}\right\}\right).$$

Theorem 1 shows that the accuracy of a first-order fluid approximation degrades as the uncertainty in the number of servers increases. In particular, when σ_λ is "small", i.e., of an order which is smaller than the order of stochastic fluctuations in the system, the optimality gap for the fluid solution is relatively small, i.e., $\mathcal{O}(\sqrt{\lambda})$. However, when σ_λ is "large", i.e., of an order which is larger than the order of stochastic fluctuations in the system, fluid approximations may lead to a considerable loss in accuracy.

Numerical example. We illustrate the results of Theorem 1 in Figures 6 and 7. In these figures, we let $N_{flex}(n_\lambda) = n_\lambda + \sigma_{n_\lambda}\epsilon$ and assume that ϵ has a uniform distribution, $\epsilon \sim U(-1,1)$. We consider $\sigma_n = n^q$ for $q = 1/4, 1/2$, and $3/4$. We let $c_{flex} = 1/3$, and $p = h = \mu = \theta = 1$. In Figure 6, we plot the scaled staffing-level errors, between the fluid and original solutions, $|n^*_\lambda - \bar{n}_\lambda|/\sqrt{\lambda}$, as λ increases. In Figure 7, we plot the corresponding scaled errors in the objectives, $|\Pi_\lambda(\bar{n}_\lambda) - \Pi_\lambda(n^*_\lambda)|/\sqrt{\lambda}$. Figure 6 illustrates the orders of magnitude for the asymptotic accuracy of fluid prescriptions. In particular, the fluid approximation's accuracy degrades as the level of uncertainty in the number of servers increases. Figure 7 illustrates the orders of magnitude of gaps in Theorem 1: the fluid approximation is considerably worse for larger values of q, which correspond to the uppermost curve in the figure.

3.3. Stochastic-Fluid Approximation

For the stochastic-fluid relaxation, we ignore stochastic fluctuations in the system. In particular, customers arrive at the rate of λ per unit of time. The processing capacity is $N_{flex}(n_\lambda)\mu$ and, by con-

servation of flow, the resulting stochastic-fluid abandonment rate is given by $\mathbb{E}[(\lambda - N_{flex}(n_\lambda)\mu)^+]$. Thus, the resulting stochastic-fluid approximation to (8) is:

$$\begin{aligned}\min_{n_\lambda \geq 0} \tilde{\Pi}_\lambda(n_\lambda) &\equiv c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda/\mu - N_{flex}(n_\lambda))^+] \\ &= c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda/\mu - n_\lambda - \sigma_{n_\lambda}\epsilon)^+] \end{aligned} \tag{11}$$

We denote $\tilde{n}_\lambda$ as the optimal solution to (11).

Paralleling Theorem 1, we first study the accuracy of the stochastic-fluid staffing prescription in (11), in a regime where the arrival rate, λ, is large.

THEOREM 2. *For large λ,*

$$\Pi_\lambda(\tilde{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}\left(\min\left\{\lambda/\sigma_\lambda, \sqrt{\lambda}\right\}\right).$$

Theorem 2 shows that when σ_λ is "large", i.e., of an order which is larger than the order of stochastic fluctuations in the system, stochastic-fluid approximations are remarkably accurate. Indeed, the stochastic-fluid approximation becomes *increasingly* accurate as the variability in the number of servers increases. On the other hand, when σ_λ is "small", i.e., of an order which is smaller than the order of stochastic fluctuations in the system, the optimality gap for the stochastic-fluid solution is on the order of stochastic fluctuations in the system, i.e., $\mathcal{O}(\sqrt{\lambda})$. In other words, when the variability in the number of servers is small, there is no distinct advantage from using stochastic-fluid approximations over fluid approximations to the system (cf. Theorem 1).

Numerical example. We illustrate the asymptotic results of Theorem 2 in Figures 8 and 9. In these figures, we compare the optimal stochastic-fluid solution, $\tilde{n}_\lambda$, to the optimal solution of the original problem, n^*_λ. We consider the same system parameters as in Figures 6 and 7. In contrast with the fluid solution, Figures 8 and 9 illustrate the improvement in accuracy for $\tilde{n}_\lambda$ as the uncertainty in the number of servers increases. Indeed, for $\sigma_n = n^{3/4}$ (bottom curve in the plots), n^*_λ and $\tilde{n}_\lambda$ are practically indistinguishable. For $\sigma_n = n^q$ and $q \leq 1/2$, comparing Figures 6 and 8 reveals that the improvement in accuracy entailed in refining the fluid solution by relying on the stochastic-fluid solution, is asymptotically negligible. For example, while the curve corresponding to $\sigma_n = n^{1/4}$ is monotonically increasing in 6, it is constant in Figure 8 which indicates that the order of magnitude of errors in the former case is larger than in the latter since we are scaling by the same quantity in both cases, namely $\sqrt{\lambda}$, and studying monotonicity as λ grows.

3.3.1. Optimal Solution. While the stochastic fluid optimal solution, $\tilde{n}_\lambda$, achieves very small optimality gap (Theorem 2), it can only be solved numerically. In this section, we study the structural property of $\tilde{n}_\lambda$ in order to gain additional insights. We also derive, when possible, simpler closed-form staffing prescriptions that achieve the same optimality gap as $\tilde{n}_\lambda$. As we will explain,

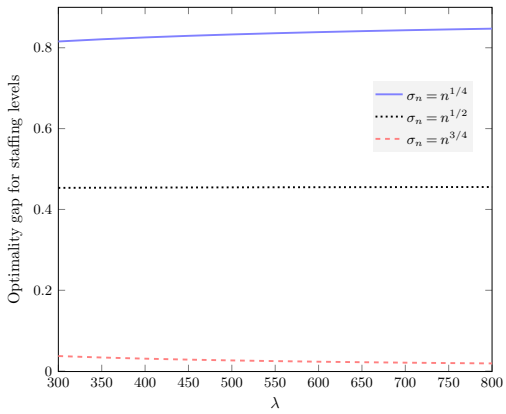

Figure 8 **Errors for optimal stochastic-fluid levels, $|\bar{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ.**

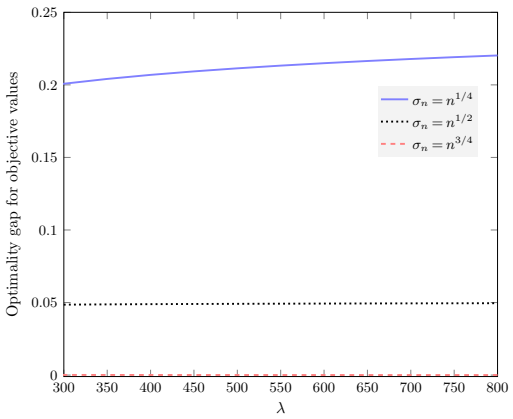

Figure 9 **Optimality gap in objective values, $|\Pi_\lambda(\bar{n}_\lambda) - \Pi_\lambda(n^*_\lambda)|/\sqrt{\lambda}$, as a function of λ.**

this rely on a detailed study of the endogeneity of uncertainty, i.e., the dependence of variability on the staffing prescription itself.

We begin by verifying that this problem's solution exists and is unique. To this end, we analyze the first two derivatives of $\tilde{\Pi}_\lambda(n_\lambda)$.

When $n_\lambda + \sigma_{n_\lambda} \le \lambda/\mu$:

$$\tilde{\Pi}'_\lambda(n_\lambda) = c_{flex} - \beta < 0.$$

Thus, in this region, we can reduce the objective cost by increasing n_λ. When $n_\lambda - \sigma_{n_\lambda} \ge \lambda/\mu$,

$$\tilde{\Pi}'_\lambda(n_\lambda) = c_{flex} > 0,$$

Thus, in this region, we can reduce the objective cost by decreasing n_λ. The two cases combined imply that the optimal $\tilde{n}_\lambda$ is achieved in the following region:

$$\Omega_\lambda \equiv \{n_\lambda : n_\lambda - \sigma_{n_\lambda} < \lambda/\mu < n_\lambda + \sigma_{n_\lambda}\},$$

which implies that $\tilde{n}_\lambda = \Theta(\lambda)$. Now, when $n_\lambda \in \Omega_\lambda$,

$$\tilde{\Pi}'_\lambda(n_\lambda) = c_{flex} - \beta F_\epsilon\left(\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}\right) - \beta\sigma'_{n_\lambda}\int_{-1}^{\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}} x f_\epsilon(x)dx \tag{12}$$

$$\tilde{\Pi}''_\lambda(n_\lambda) = \underbrace{\beta f_\epsilon\left(\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}\right)\frac{1}{\sigma_{n_\lambda}}\left(1 + \sigma'_{n_\lambda}\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}\right)^2}_{(A)} \underbrace{- \beta\sigma''_{n_\lambda}\int_{-1}^{\frac{\lambda/\mu - n_\lambda}{\sigma_{n_\lambda}}} x f_\epsilon(x)dx}_{(B)} \tag{13}$$

We first notice that both (A) and (B) are positive in (13). Thus, it is not clear a priori whether (13) is positive, i.e., that the objective is strictly convex. However, we also notice that for $n_\lambda \in \Omega_\lambda$, $\sigma_{n_\lambda}\sigma''_{n_\lambda} = o(1)$ and $\sigma'_{n_\lambda} = o(1)$; see Definition 1. This suggests that, for λ large enough, $\tilde{\Pi}''_\lambda(n_\lambda) \ge 0$

which, in turn, implies that $\tilde{\Pi}'_\lambda(n_\lambda)=0$ in (12) has a unique solution, which is the minimizer of (11), i.e., $\tilde{n}_\lambda$. Next, we analyze the properties of $\tilde{n}_\lambda$. To do so, it will be helpful to define

$$\hat{n}_\lambda \equiv \lambda/\mu - \gamma\sigma_{\lambda/\mu}, \quad \text{for} \quad \gamma \equiv F_\epsilon^{-1}(c_{flex}/\beta). \tag{14}$$

It is important to explain why we introduce $\hat{n}_\lambda$ in (14): Recall that a fundamental distinction between uncertainty in demand and in supply is that, in the latter case, variability is endogenous because the distribution of the number of servers depends itself on the staffing prescription. Indeed, when the uncertain parameter is the arrival rate, as in Bassamboo et al. (2010), there is no endogeneity between the staffing decision and the underlying randomness, and the optimal staffing rule takes the form of $\hat{n}_\lambda$ in (14). We also note that if $\sigma'_{n_\lambda}=0$, $\tilde{\Pi}''_\lambda(n_\lambda)\geq 0$ and $\hat{n}_\lambda$ is the solution of $\tilde{\Pi}'_\lambda(n_\lambda)=0$. However, when $\sigma'_{n_\lambda}>0$ as in our model, the endogeneity arises. A natural question to ask, then, does the endogeneity matter or when would it be appropriate to ignore the dependence between staffing prescription and that underlying randomness, i.e., σ_{n_λ} can be approximated by $\sigma_{\lambda/\mu}$, without losing much in optimality? Lemma 1 provides an answer to this question by quantifying the gap between $\tilde{n}_\lambda$ and $\hat{n}_\lambda$.

LEMMA 1. *For γ in (14), assume that there exists $\delta>0$ such that $f_\epsilon(x)>0$ for $x\in(\gamma-\delta,\gamma+\delta)$. Then, for λ large enough,*

$$|\hat{n}_\lambda - \tilde{n}_\lambda| = \mathcal{O}\left(\sigma_\lambda^2/\lambda\right),$$

and

$$\bar{\Pi}_\lambda(\hat{n}_\lambda) = \bar{\Pi}_\lambda(\tilde{n}_\lambda) + \mathcal{O}\left(\sigma_\lambda^3/\lambda^2\right). \tag{15}$$

To understand the significance of Lemma 1, we recall the optimality gap in Theorem 2 and let $\sigma_n = an^q$. We begin by noting that, based on Theorem 2: For $q>1/2$, $|\Pi_\lambda(\tilde{n}_\lambda)-\Pi_\lambda(n^*_\lambda)|=\mathcal{O}(\lambda/\sigma_\lambda)$. Also, it holds that for $q\leq 3/4$, $\sigma_\lambda^3/\lambda^2\leq\lambda/\sigma_\lambda$. Thus, based on (15), we see that for $1/2<q\leq 3/4$, we can ignore the dependence between the uncertainty and our staffing decision, and use the simpler staffing rule $\hat{n}_\lambda=\lambda/\mu-\gamma\sigma_{\lambda/\mu}$, as defined in (14), to achieve the $\mathcal{O}(\lambda/\sigma_\lambda)$ optimality gap as in Theorem 2. The next theorem is the main result of this section. It summarizes the gaps in approximating the optimal staffing level, n^*_λ, based on the fluid or stochastic-fluid relaxations. When the magnitude of uncertainty is not very large, we derive closed-from staffing prescriptions that achieve the same order of optimality gap as $\tilde{n}_\lambda$ (this corresponds to cases I and II in Theorem 3).

THEOREM 3. *Let $\sigma_n=an^q$, for $a>0$ and $0\leq q\leq 1$, and distinguish among four cases:*

*(I) **[Variability-dominated.]** If $0\leq q\leq 1/2$, we set $n_\lambda=\bar{n}_\lambda=\lambda/\mu$. In this case, we have*

$$\Pi_\lambda(n_\lambda)=\Pi_\lambda(n^*_\lambda)+\mathcal{O}(\sqrt{\lambda}).$$

(II) ***[Moderately uncertainty-dominated.]*** *If* $1/2 < q \leq 3/4$, *we set* $n_\lambda = \hat{n}_\lambda = \lambda/\mu - \gamma\sigma_{\lambda/\mu}$. *In this case, we have*

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda).$$

(III) ***[Strongly uncertainty-dominated.]*** *If* $3/4 < q < 1$, *set* $n_\lambda = \tilde{n}_\lambda$. *In this case,*

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda).$$

(IV) ***[Extremely uncertainty-dominated.]*** *If* $q = 1$ *and* $0 < a < 1$, *then set* $n_\lambda = \tilde{n}_\lambda = \frac{\lambda}{\mu}\eta$, *where* η *denotes the solution to*

$$c_{flex} + \beta a \int_{-1}^{1/(a\eta)-1/a} F_\epsilon(u)du - \frac{\beta}{\eta} F_\epsilon\left(\frac{1}{a\eta} - \frac{1}{a}\right) = 0. \tag{16}$$

In this case,

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda) = \Pi_\lambda(n_\lambda^*) + \mathcal{O}(1).$$

We emphasize that in cases I and II, n_λ is *not* the optimal solution to (11), but rather a simplified closed-form solution which we prove yields, asymptotically, the same order of accuracy as the actual optimal solution (cf. Theorem 2). In contrast, n_λ in cases III and IV is the actual solution to (11). For case III and IV, simplifying as we did in the former cases can lead to substantial errors as demonstrated in Lemma 1. We also note that in case III, $\tilde{n}_\lambda$ is the solution to an implicit equation since the dependence between staffing prescription and parameter uncertainty cannot be ignored in this case.

It is also interesting to note that in case IV of Theorem 3, i.e., with $q = 1$, it may be beneficial to underload or overload the system. For example, we obtain from (16) that $\eta \geq 1$, i.e., we underload, if $c_{flex} \leq \beta F_\epsilon(0) - \beta a \int_{-1}^{0} F_\epsilon(u)du$, i.e., capacity is cheap. This lies in contrast with all cases where $q < 1$. There the uncertainty hedge is of a smaller order than the expected number of servers.

Numerical evidence. We present numerical evidence substantiating the results of Theorem 3 in Figures 10 and 11. The dashed curves in those figures correspond to the percent error in using the solution from case (II) of Theorem 3 relative to the optimal solution of (8); the solid curves correspond to using the solution from case (III) instead. We let $\sigma_n = n^q$ for $0 < q \leq 0.95$, i.e., we exclude case (IV) in the theorem, and assume that ϵ in (5) is uniformly distributed over $(-1, 1)$. In both figures, we assume that $\lambda = 200$, $c_{fix} = 1/2$, $c_{flex} = 1/4$, and $r = h = \mu = 1$. In Figure 10 we let $\theta = 1$, and in Figure 11 we let $\theta = 2$, which gives a smaller β as defined in (10).

Clearly, there is considerable loss in accuracy if ignoring the dependence between the variability in supply and the staffing prescription when there is considerable variability, i.e., q is large. This loss is also exacerbated for more impatient customers (Figure 11). However, when the variability is not too large ($q < 3/4$ in the figures, as in Theorem 3), there is no asymptotically discernable gain from taking that dependence into account.

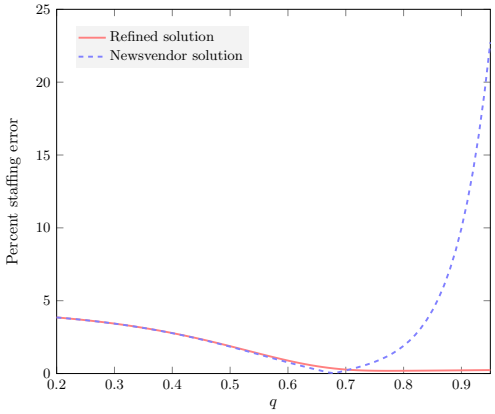

Figure 10 **Stochastic-fluid solution: Percent relative errors in staffing for $\theta = 1$.**

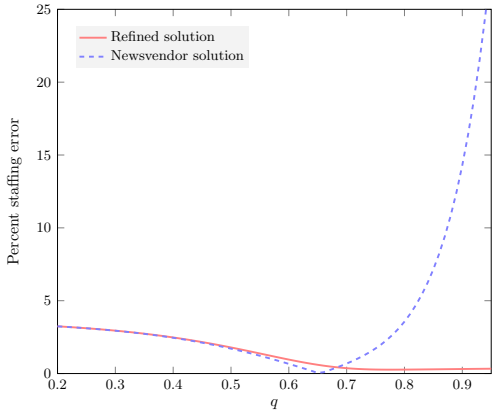

Figure 11 **Stochastic-fluid solution: Percent relative errors in staffing for $\theta = 2$.**

$p = 0.4$ and $\alpha = 0.5$ in (3)				
λ	$\tilde{n}$	n_{Bin}	$\Pi(\tilde{n})$	$\Pi(n_{Bin})$
100	295	270	45.1	47.1
200	579	525	83.5	87.9
300	841	778	121	127
400	1,109	1,036	158	164
500	1,353	1,293	195	201

$p = 0.4$ and $\alpha = 0.8$ in (3)				
λ	$\tilde{n}$	n_{Bin}	$\Pi(\tilde{n})$	$\Pi(n_{Bin})$
100	349	270	63.0	70.4
200	689	525	113	129
300	1,016	778	163	187
400	1,315	1,036	209	241
500	1,675	1,293	255	291

Table 2 **Implications on staffing levels and costs: Classical Binomial versus Correlated Bernoulli models.**

3.3.2. Revisiting Extensions of the Binomial Model: Staffing Implications. In Table 2, we revisit the classical Binomial model and the correlated Bernoulli model introduced in §2.3. Our objective is to quantify the implications on optimal staffing prescriptions when wrongfully relying on the classical Binomial model. In particular, we let $r = h = \mu = \theta = 1$, and $c_{flex} = 1/3$. We assume that the actual distribution of the number of available servers is according to the correlated Binomial model in (3). In that model, we fix $r = 0.4$, and consider two values of α: 0.5 and 0.8. We recall that the order of variability is $\sigma_n = \mathcal{O}(n^\alpha)$ for $\alpha > 1/2$, as given in (4). We vary the value of λ, and report the values of the optimal staffing prescription, $\tilde{n}$, to problem (11); see the second column in the tables. To quantify the error obtained in approximating the correlated Bernoulli model by the classical Binomial model, we also report the optimal staffing prescriptions according to the the classical Binomial model, n_{Bin}; see the third column in the tables. Finally, we report the objective costs under each staffing prescription. Table 2 illustrates that, depending on the value of α, the increase in costs by wrongfully relying on the classical Binomial model can be great. For example, for $\alpha = 0.8$, the percent increase in cost between $\Pi(\tilde{n})$ and $\Pi(n_{Bin})$ ranges between 12% for $\lambda = 100$ and 14% for $\lambda = 500$. Moreover, as expected, for $\alpha = 0.5$, the difference between the two costs is almost negligible: It is about 2% for $\lambda = 500$.

4. Capacity Sizing with a Blended Workforce

In this section, we study the optimal staffing level with a blended workforce. We focus here on analyzing the tradeoffs between the staffing cost, the flexibility to scale up or down supply with seasonal demand, and supply-side uncertainty. We also highlight how the staffing policy with a blended workforce "builds on" the staffing policy with flexible servers only, developed in the previous section.

4.1. Staffing Problem and Approximations

We assume that the number of fixed servers is fixed throughout the time horizon, but the number of flexible servers can vary for different periods. Let m_λ denote the number of fixed servers, and $\mathbf{n}_\lambda = (n_\lambda^1, \ldots, n_\lambda^k)$ denote the number of flexible servers. The total number of servers in period i is given by:

$$N(m_\lambda, n_\lambda^i) = m_\lambda + N_{flex}(n_\lambda^i) = m_\lambda + n_\lambda^i + \sigma_{n_\lambda^i}\epsilon_i, \tag{17}$$

where $N_{flex}(n_\lambda^i)$ is given in (6). At the initial planning stage, the manager must decide on the numbers of fixed and flexible servers. Recall that we denote c_{fix} as the per unit time staffing cost of a fixed server, and c_{flex} is the per unit time staffing cost for a flexible server. In addition to Assumption 1, we make the following parallel assumption on c_{fix}, to avoid pathological cases.

ASSUMPTION 2. *We assume that* $c_{fix} < (h/\theta + r)\mu$.

We are now ready to formulate the staffing problem with a blended workforce, paralleling (7):

$$\begin{aligned} \min_{m_\lambda, \mathbf{n}_\lambda} \quad & \Pi_\lambda(m_\lambda, \mathbf{n}_\lambda) \\ \equiv & \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + h\mathbb{E}[Q^i(m_\lambda, n_\lambda^i)] + r\xi(m_\lambda, n_\lambda^i) \right), \\ = & \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + (h + r\theta)\mathbb{E}[Q^i(m_\lambda, n_\lambda^i)] \right), \end{aligned} \tag{18}$$

A fundamental difference between the problem formulations in (7), with flexible servers only, and in (18), with a blended workforce, is that (18) may no longer be decomposed into k single-period problems due to the fixed servers.

Solving (18) is challenging as discussed in §3. We thus consider two relaxations: the fluid approximation and the stochastic-fluid approximation. The fluid approximation takes the form:

$$\min_{m_\lambda, \mathbf{n}_\lambda} \bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \equiv \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \beta \left(\lambda_i/\mu - m_\lambda - n_\lambda^i \right)^+ \right). \tag{19}$$

The stochastic-fluid approximation takes the form:

$$\min_{m_\lambda, \mathbf{n}_\lambda} \tilde{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \equiv \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \beta\mathbb{E}\left[\left(\lambda_i/\mu - N(m_\lambda, n_\lambda^i) \right)^+ \right] \right). \tag{20}$$

4.2. Asymptotic Accuracy

In this section, we derive the asymptotic orders of accuracy of the fluid and stochastic-fluid approximations with a blended workforce. Let $n_\lambda^{(m)} \equiv \min \mathbf{n}_\lambda$ and $n_\lambda^{(M)} \equiv \max \mathbf{n}_\lambda$. Denote m_λ^* and $\mathbf{n}_\lambda^*$ as the optimal solution of (18), $\bar{m}_\lambda$ and $\bar{\mathbf{n}}_\lambda$ as the optimal solution of (19), and $\tilde{m}_\lambda$ and $\tilde{\mathbf{n}}_\lambda$ as the optimal solution of (20). Theorem 4 parallels Theorem 1 and corresponds to the fluid approximation, (19).

THEOREM 4. *For large λ,*

$$\Pi_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) = \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}\left(\max\left\{\sigma_{\bar{n}_\lambda^{(M)}}, \sqrt{\lambda}\right\}\right);$$

i.e., if $\bar{n}_\lambda^{(M)} = \Theta(\lambda)$, then $\Pi_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) = \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}\left(\max\left\{\sigma_\lambda, \sqrt{\lambda}\right\}\right)$.

Theorem 5 parallels Theorem 2 and corresponds to the stochastic-fluid approximation, (20).

THEOREM 5. *For large λ,*

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda) = \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}\left(\sqrt{\lambda}\right).$$

Moreover, if $\tilde{n}_\lambda^{(m)} = \Theta(\lambda)$, then $\Pi_\lambda(\bar{m}_\lambda, \tilde{\mathbf{n}}_\lambda) = \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}\left(\min\left\{\lambda/\sigma_\lambda, \sqrt{\lambda}\right\}\right)$.

4.3. Optimal Staffing Policies

In this section, we derive the optimal staffing policy. We shall start by deriving the optimal staffing policy for the fluid problem in (19). We then derive the optimal staffing policy for the stochastic-fluid problem in (20).

Solution to the fluid problem. Since the fluid approximation ignores parameter uncertainty, the optimal staffing policy captures the tradeoff between the costs of staffing and the flexibility of scaling the number of flexible servers to meet seasonality in demand. In contrast, we notice that if we decide to staff enough fixed servers to meet the demand in a given period, h, then we must staff these servers for all other periods as well. Thus, we define the time "modified" cost, c_{fix}^h, which will be useful in describing the fluid-optimal solution:

$$c_{fix}^h \equiv c_{fix} \cdot \frac{\sum_{i=1}^k T_i}{\sum_{i=h}^k T_i} \text{ for } h \geq 1. \tag{21}$$

We summarize the fluid-optimal staffing policy in the following lemma.

LEMMA 2. *The solution to the fluid problem in (19), with a blended workforce, is given by*

$$\begin{cases} \bar{m}_\lambda = 0, & \text{for } k_0 = 0, \\ \bar{m}_\lambda = \frac{\lambda_{k_0}}{\mu} & \text{for } k_0 > 0, \\ \bar{n}_\lambda^i = 0, & \text{for } 1 \leq i \leq k_0, \\ \bar{n}_\lambda^i = \frac{\lambda_i - \lambda_{k_0}}{\mu}, & \text{for } k_0 < i \leq k, \end{cases}$$

where k_0 is defined as follows:

$$k_0 = \begin{cases} 0, & \text{if } c_{flex} < c_{fix}, \\ \max\{1 \leq h \leq k : c_{flex} \geq c_{fix}^h\}, & \text{otherwise.} \end{cases}$$

As the pool of flexible agents can be dynamically adjusted to meet seasonality in customer demand, the only case where the manager would staff fixed servers is if $c_{flex} \geq c_{fix}$, i.e., they are cheaper. However, even in this case, she may still staff the more expensive flexible servers, i.e., she would blend her workforce, unless fixed servers are "very" cheap, e.g., if $c_{fix} < \frac{T_k}{\sum_{i=1}^{k} T_i} c_{flex}$. The general form of the optimal staffing policy in Lemma 2 shows that a manager who blends should rely solely on the fixed resource in the low-demand periods, and blend in the high-demand periods.

It is useful to comment on how the solution in Lemma 2 relates to the optimal solution when no flexible servers are allowed in the pool. We notice that when there are no flexible agents, the manager will set $m_\lambda = \lambda_{k_\beta}/\mu$ where β is defined in (10), and the index k_β is given by the following equation:

$$k_\beta \equiv \max\{1 \leq h \leq k : \beta \geq c_{fix}^h\}.$$

Recalling that, by Assumption 1, $c_{flex} < \beta$, we see that, given the option of a blended workforce, the manager will use less fixed servers, i.e., we have that $k_\beta > k_0$. In addition, the smaller c_{flex} is, the smaller k_0 is, implying that the manager will staff fewer fixed servers.

Solution to the stochastic-fluid problem. While the optimal solution to the fluid staffing problem is readily obtainable and easy to interpret, Theorem 4 shows that it may not be reliable when uncertainty in the number of available servers is large. Thus, there is a need to consider a stochastic-fluid refinement. Compared with the fluid approximation, the stochastic-fluid approximation does take parameter uncertainty into account. Thus, in solving the stochastic-fluid optimal staffing policy, we can capture the tradeoffs among three factors: the staffing cost, the flexibility in scaling the workforce to meet seasonality in demand, and supply-side uncertainty. However, as our main result in this section (Theorem 6) will show, for $\sigma_n = n^q$, when $q < 1$ (Case 1), the first-order factors are only the staffing cost and the scaling flexibility. Particularly, the optimal staffing level follows closely the fluid-based prescription in Lemma 2. Supply-side uncertainty only affects the magnitude of the hedge that we add to the fluid-based prescription. When $q = 1$ (Case 2), the uncertainty plays a first-order role. In this case, we need to add a "risk premium" on the cost of the flexible servers, due to that large magnitude of uncertainty. Thus, the optimal staffing rule may differ substantially from the fluid-based prescription.

We note that there are only two cases in Theorem 6, because the structure of the optimal staffing policy is the same in the variability-dominated, moderately and strongly uncertainty-dominated regimes. Thus, we group these three regimes into only one case, namely Case 1 of Theorem 6. Case 2 in Theorem 6 is reserved to the case of "extreme" uncertainty.

We define:

$$g(c) \equiv c + caF_\epsilon^{-1}(c/\beta) - \beta a \int_{-1}^{F_\epsilon^{-1}(c/\beta)} F_\epsilon(u)du. \tag{22}$$

THEOREM 6. *For the solution of the stochastic-fluid problem in (20), there are two cases:*

- ***Case 1. Variability-dominated, moderately and strongly uncertainty-dominated regimes.*** *If $\sigma_n = an^q$ for $0 < q < 1$ and $a > 0$, then for $\bar{m}_\lambda$ and k_0 as given by Lemma 2, the optimal solution is:*

  — $\tilde{m}_\lambda = \bar{m}_\lambda$,

  — *For $i \le k_0$, $\tilde{n}^i_\lambda = 0$,*

  — *For $i > k_0$, $\tilde{n}^i_\lambda$ is the minimizer of the following single-period problem:*

$$\min_{n_\lambda \ge 0} c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda - \sigma_{n_\lambda}\epsilon)^+], \tag{23}$$

  as given by cases I, II, and III of Theorem 3, depending on the value of q.

- ***Case 2. Extremely uncertainty-dominated regime.*** *If $\sigma_n = an$, for $0 < a < 1$, then for c^h_{fix} in (21) and $g(\cdot)$ in (22), let:*

$$k_1 = \begin{cases} 0, & \text{if } c_{flex} < g(c_{fix}) \\ \max\{1 \le h \le k : c_{flex} \ge g(c^h_{fix})\}, & \text{otherwise.} \end{cases}$$

  The optimal solution is:

  — $\tilde{m}_\lambda = \lambda_{k_1}/\mu$,

  — *For $i \le k_1$, $\tilde{n}^i_\lambda = 0$,*

  — *For $i > k_1$, $\tilde{n}^i_\lambda$ is the minimizer of the following single-period problem:*

$$\min_{n_\lambda \ge 0} c_{flex} n_\lambda + \beta \mathbb{E}[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda - an_\lambda\epsilon)^+], \tag{24}$$

  as given by case IV in Theorem 3.

Recall from Lemma 2 that the optimal solution to the fluid problem is to rely strictly on the fixed resource in lower-demand periods, up to some period index k_0, and to blend resources in higher-demand periods. Theorem 6 shows that the optimal staffing policy for the stochastic-fluid problem has a similar structure when $\sigma_n = an^q$ for $q < 1$. Indeed, in lower-demand periods, up to period k_0, the manager should also rely strictly on the fixed resource. Moreover, the optimal staffing level for the fixed resource in the stochastic-fluid problem remains the same as for the fluid problem, i.e., $\tilde{m}_\lambda = \bar{m}_\lambda$ and $\tilde{n}^i_\lambda = 0$ for $i \le k_0$. In higher-demand periods, i.e., periods whose index exceeds k_0, the manager should blend her workforce. The staffing levels for the flexible resource in the stochastic-fluid problem are slightly different than those given by the fluid solution in Lemma 2. In particular, for each period $i > k_0$, we must solve (23), as in Theorem 3. The solution to (23) essentially adds an uncertainty hedging of order σ_λ to the fluid solution.

When $\sigma_n = an$, for $0 < a < 1$, it remains optimal to rely on the fixed resource in low-demand periods and to blend in high demand periods. However, the fluid-based prescription of Lemma 2

could lead to substantial errors in this case. In particular, as $g(c^h_{fix}) < c^h_{fix}$, there is essentially a "risk premium" that is incurred on the cost of the flexible resource. This implies that the period index k_1 may be larger than k_0. For example, with all other parameters held equal, it may be optimal to rely on the flexible resource in Case 1 but not in Case 2.

Theorem 6 shows that the key in determining the optimal staffing policy lies in the order of magnitude of the variability in the flexible resource. As $\lambda_i - \tilde{m}_\lambda = \Theta(\lambda)$ for $i > k_0$, the presence of fixed servers in blended periods does not, in loose terms, impact the scale of supply-side variability.

4.4. Supporting Numerical Study

In this section, we describe results of numerical experiments which illustrate the benefit and cost of staffing a blended workforce. We focus on two alternative perspectives: The firm's in section 4.4.1, and the customer's in section 4.4.2.

We consider three different scenarios: i) fixed servers only, ii) flexible servers only, iii) both fixed and flexible servers are allowed in the agent pool. The main objective is to compare the optimal staffing policies, the costs incurred by the service provider, and the quality of service experienced by customers, in those three scenarios. Notice that even though we partially incorporate the quality of service through the waiting cost and the abandonment cost in the objective function, we would like to study performance measures beyond these two. In particular, we consider the probability of delay in steady state. It is customary to consider the probability of delay as a measure of the quality of service, e.g., see Halfin and Whitt (1981) and Garnett et al. (2002). Indeed, in the asymptotic regimes that arise at optimum for our staffing problem, the queue length is generally small (of a smaller order of magnitude than the average number of servers), and the waiting time is negligible when the system is large, so that considering the probability of delay is of interest.

We consider a nontrivial 2-period case where it is optimal to use a blended workforce in the high demand period when blending is allowed. This essentially requires that $\frac{T_h}{T_l+T_h} c_{flex} < c_{fix} < c_{flex}$ where T_l is the length of the low-demand period and T_h is the length of the high demand period. We set the low demand rate $\lambda_l = 25$ and the high demand rate $\lambda_h = 50$. We let the lengths of the respective periods be $T_l = 2$ and $T_h = 1$. The staffing-cost parameters are $c_{fix} = 2/9$ and $c_{flex} = 1/3$, and we set the customer-cost parameters $h = r = 1$. The service rate $\mu = 1$ and the abandonment rate $\theta = 0.5$. We vary the value of q, where $\sigma_n = n^q$ in (6): We consider values of q between 0.2 and 0.99, with increments of 0.01, which correspond to various magnitudes of supply-side variability.

For the optimal staffing levels, we numerically solve the original staffing problem in (7). Our example illustrates a case where the index k_0 in Lemma 2 is different from k_1 in Theorem 6. Indeed, for $q < 1$, the optimal solution in the stochastic-fluid problem is to blend the resources in the high-demand period i.e, we have that $k_0 = 1$. On the other hand, for $q = 1$, we have that

$k_1 = 2$, so that it is optimal not to use a blended workforce because of the "risk premium" induced when $q = 1$. In Figure 12, we plot the optimal staffing levels under each alternative in the high and low-demand periods. We include two distinct curves for the case of a blended workforce: The level of the fixed resource, and the level of the flexible resource. In contrast, when the manager is restricted to staffing strictly one of the two resources, we include a single curve in the plot instead. We note that while there is a phase transition when solving the stochastic-fluid problem, i.e., $q < 1$ versus $q = 1$, solving (7) directly illustrates a more gradual transition as q increases.

4.4.1. Firm Perspective: Cost Reduction. In Figure 13, we plot the optimal cost curves for the firm, as a function of q, under the three workforce models. While it is intuitively clear that allowing for blending in the workforce would reduce costs, it is interesting to understand what is the cost reduction entailed compared to the two single-resource benchmarks.

The relative improvement of a blended workforce over using a fixed resource only decreases as q increases, for large values of q. This is because when supply-side variability is very large, the number of flexible servers staffed decreases as q increases in this example; see the lowermost curve in the lower subplot of Figure 12, which corresponds to the high-demand period. Eventually, for $q = 1$, the manager relies strictly on fixed servers, even when allowed to rely on flexible servers too. Thus, the objective cost under the blended workforce model becomes increasingly close to that of the model with fixed servers only, as is illustrated by the two lowermost curves in Figure 13. In contrast, the improvement in cost between the blended workforce model and the model with a flexible resource only increases as a function of q. This is because when only flexible servers are allowed, the manager must staff increasingly more flexible servers to hedge against supply-side uncertainty as q increases in this example (see Figure 12). Thus, the objective cost increases steeply in this case; see the top curve in Figure 13. On the other hand, when fixed servers are allowed, the manager may rely on this alternative resource instead, which reduces her cost.

4.4.2. Customer Perspective: Quality of Service. We now illustrate the impact of blending the workforce on the quality of service offered in the system. For our choice of system parameters, the solution in scenario (i) is most cost-effective from the point of view of the manager; however, it is unclear whether customers will experience a higher quality of service under that scenario. Figure 14 illustrates that the impact of blending on customers depends on the period. In the low-demand period, the smallest delay probability corresponds to staffing solely from the fixed resource. This is because the manager staffs a high-enough level to match demand in the high period (lower subplot in Figure 12), which leaves the system overstaffed in the low-demand period. Indeed, the delay probability is almost 0 in this case (lowermost curve in the upper subplot of Figure 14). However, customers in the high-demand period are worse off when the manager staffs

strictly from the fixed resource. This is because when flexible servers are allowed in the pool, the manager hedges against uncertainty by staffing a larger pool, which benefits customers who, as a result, experience a smaller probability of delay. We see this by comparing the uppermost curve, which corresponds to the fixed workforce, to the middle curve, which corresponds to the blended workforce, in the lower subplot of Figure 14.

Figure 14 also shows that blending may either help or hurt customers, compared with staffing from the flexible resource only. In the low-demand period, the uncertainty hedge is not large enough, so that customers are benefitted from blending. We see this by comparing the uppermost probability of delay curve, which corresponds to having only flexible servers, to the middle probability of delay curve, which corresponds to the blended workforce model (in the low demand period, only fixed servers are used). In the high-demand period, the uncertainty hedge is large, and customers benefit from this. This can be seen by comparing the lowermost curve in the lower subplot of Figure 14, which corresponds to the case with flexible servers only, to the middle curve which corresponds to a blended workforce.

5. General Abandonment

Our results so far are all under the assumption of exponentially-distributed patience times. Since there is statistical evidence indicating that patience times may not be exponentially-distributed (Brown et al. 2005), it is important to go beyond that assumption. We do so in this section by describing results from a numerical study quantifying the optimality gaps for fluid-based and stochastic-fluid based approximation with non-exponential abandonment. We shall focus on flexible servers only in the agent pool.

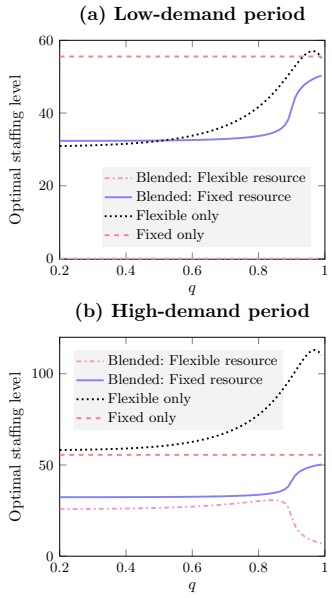

Figure 12 Staffing decisions.

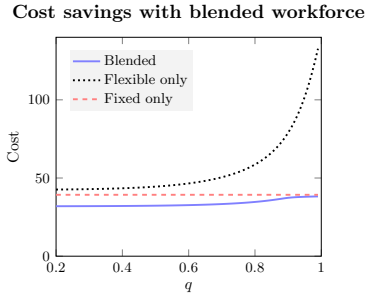

Figure 13 Optimal costs.

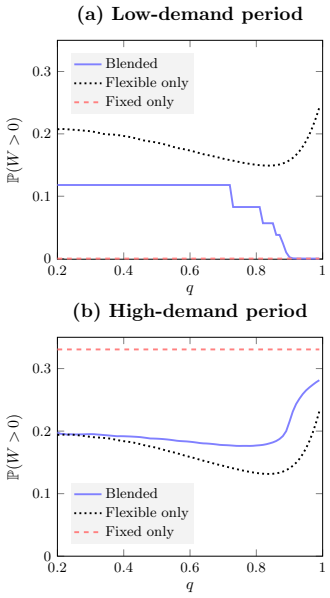

Figure 14 Delay probabilities.

Abandonment distributions and hazard rates. Using hazard rates to describe customer patience is known since the work of Palm (1953). Data collected in service systems suggest that customer patience times may have a decreasing hazard rate. For example, Zeltyn and Mandelbaum (2005) find that "customers who have already waited for a significant time, tend to remain increasingly patient" (p.14). Additionally, Bolandifar et al. (2019) also conclude that patience times have a decreasing hazard rate. In what follows, we go beyond exponential patience times by considering patience-time distributions with both increasing and decreasing hazard rates.

Staffing problem and relaxations. We begin by formulating the firm's optimization problem when times to abandon have a general distribution. As in (7), the firm's original problem is given by:

$$\min_{m_\lambda, \mathbf{n}_\lambda} \Pi_\lambda(m_\lambda, \mathbf{n}_\lambda) \equiv \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + h \cdot \mathbb{E}[Q^i(m_\lambda, n_\lambda^i)] + r \cdot \xi(m_\lambda, n_\lambda^i) \right).$$

Because of the difficulty in solving this staffing problem, we now describe both the stochastic-fluid and fluid relaxations of the problem. To do so, we let G denote the cdf of the abandonment-time distribution, $\bar{G}$ its tail cdf, and g its pdf. We assume that g is strictly positive so that $\bar{G}$ is invertible. The fluid formulation of the problem is given by:

$$\begin{aligned} \min_{m_\lambda, \mathbf{n}_\lambda} \quad & \bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \\ & \equiv \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + r(\lambda_i - m_\lambda \mu - n_\lambda^i \mu)^+ + h \left(\int_0^{w^i(m_\lambda, n_\lambda^i)} \lambda_i \bar{G}(u) du \right) \right), \end{aligned}$$

where $w^i(m_\lambda, n_\lambda^i) = \bar{G}^{-1}((m_\lambda \mu + n_\lambda^i \mu)/\lambda_i)$ denotes the fluid approximation of the waiting time in the i^{th} period. Accordingly, the stochastic-fluid formulation of the problem is given by:

$$\begin{aligned} \min_{m_\lambda, \mathbf{n}_\lambda} \quad & \tilde{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \\ & \equiv \sum_{i=1}^{k} T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + r\mathbb{E}[(\lambda_i - N(m_\lambda, n_\lambda^i)\mu)^+] + h\mathbb{E}\left[\int_0^{W^i(m_\lambda, n_\lambda^i)} \lambda_i \bar{G}(u) du \right] \right), \end{aligned}$$

where $W^i(m_\lambda, n_\lambda^i) = \bar{G}^{-1}(N(m_\lambda, n_\lambda^i)\mu/\lambda_i)$ denotes the stochastic-fluid approximation of the waiting time in the i^{th} period.

Numerical results. In Figures 15-18, we consider a problem with a single period and flexible servers only. Our objective here is to quantify the accuracies of the alternative staffing-problem relaxations. For the patience-time distribution, we consider Pareto (mean 1, shape 2) and Weibull (mean 1, shape 2). We choose these two distributions because they exhibit, for those selected parameter values, different properties for their hazard-rate functions: While the Pareto distribution has a decreasing hazard rate, the Weibull distribution has an increasing harzard rate. We consider the following cost parameters: $c_{flex} = 1$, $h = 1$, $r = 0.45$, and $\mu = 1$. For each distribution, we

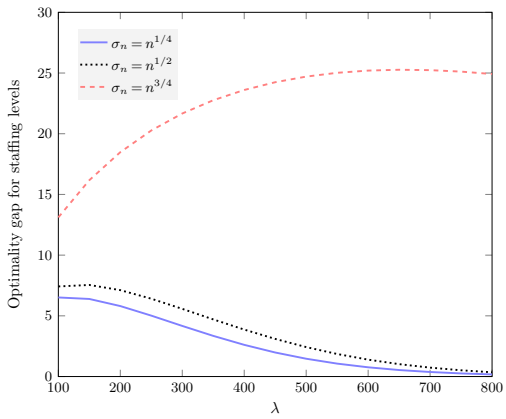

Figure 15 Unscaled errors for optimal fluid staffing levels, $|\bar{n}_\lambda - n^*_\lambda|$, as a function of λ, with Pareto abandonment.

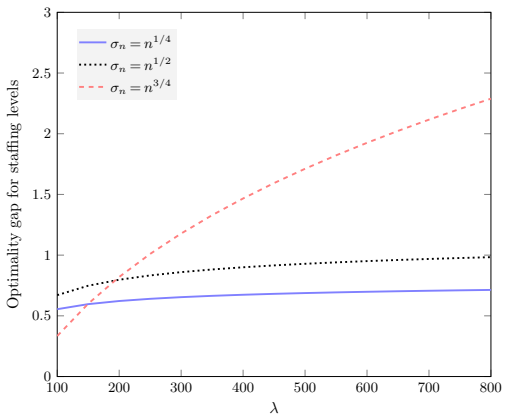

Figure 16 Scaled errors for optimal fluid staffing levels, $|\bar{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ, with Weibull abandonment.

compare the fluid, $\bar{n}_\lambda$, stochastic-fluid, $\tilde{n}_\lambda$, and original, n^*_λ, optimal solutions, by plotting their respective differences for varying arrival rates.

We first discuss our numerical results with Pareto abandonment. In this case, the overloaded regime is asymptotically optimal at fluid scale, i.e., the optimal prescription is not to match mean demand and mean supply. Thus, we expect that fluid prescriptions should be extremely accurate, i.e., with absolute errors on the order of magnitude of $\mathcal{O}(1)$ as in Bassamboo and Randhawa (2010). In other words, we expect that stochastic-fluid prescriptions would not lead to a substantial improvement over their fluid counterparts; this is confirmed by Figures 15 and 17, where we plot **unscaled** absolute differences $|\bar{n}_\lambda - n^*_\lambda|$ and $|\tilde{n}_\lambda - n^*_\lambda|$.

With Weibull abandonment, the fluid solution prescribes a critically-loaded regime, i.e., to match the mean demand and the mean supply. Because this is the same asymptotic regime prescribed with exponential abandonment, we expect the optimality gaps of our respective solutions to be close to those with exponential abandonment. Figures 16 and 18 confirm that this is indeed the case. In particular, the stochastic-fluid formulation is remarkably accurate, yielding an order of magnitude improvement over the fluid prescription (in most cases, $\tilde{n}_\lambda$ and n^*_λ are indistinguishable).

6. Concluding Remarks

In this paper, we studied the problem of staffing a service system where the manager must decide on cost-minimizing levels of fixed and/ or flexible agents. Our analysis suggests that it may be cost-effective to staff either strictly one of the two resources, or to use a blended workforce, depending on the interaction between three competing factors: (i) operational costs; (ii) the supply-side flexibility to meet the time-variation in customers' demand; and (iii) the supply-side uncertainty which is associated with staffing flexible agents. In broad terms, we showed that the optimal staffing levels

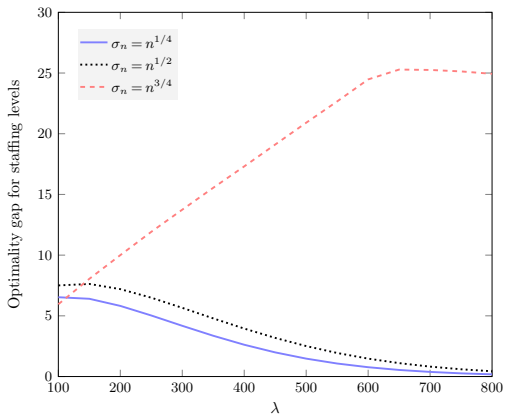

Figure 17 Unscaled errors for optimal stochastic-fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|$, as a function of λ, with Pareto abandonment.

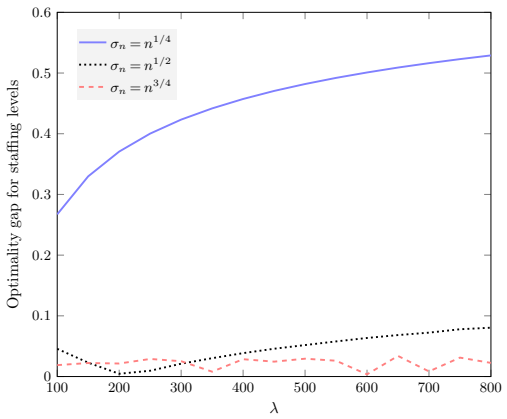

Figure 18 Scaled errors for optimal stochastic-fluid staffing levels, $|\tilde{n}_\lambda - n^*_\lambda|/\sqrt{\lambda}$, as a function of λ, with Weibull abandonment.

involve both a base capacity, which is used to match mean demand, and an additional safety capacity which hedges against supply-side variability.

In this work, we focused solely on the long-run staffing decision in the system. Our focus on that long-run strategic planning decision was motivated by: (1) the longer time scale which is associated with the staffing decision in practice, e.g., to allow for the training of agents, and (2) the fact that even though real-time pricing is used by some on-demand service platforms, such as ride-sharing services, most such platforms have to commit to the prices that they offer to their agents well in advance (Taylor 2018). Nevertheless, it would be an interesting future research to investigate the dynamic compensation decision in the setting with a blended workforce.

In addition, in this work, we consider deterministic arrival rates. The goal is to highlight the impact of supply-side uncertainty. We can also consider demand-side uncertainty, i.e., random arrival rate, in addition to the supply-side uncertainty. In particular, following the setting in this paper, assume the staffing decisions are made before the random arrival rates and the random numbers of servers are realized. Then applying similar lines of analysis, we can show that the stochastic-fluid problem still leads to a remarkably accurate staffing prescription and the uncertainty hedge in the optimal stochastic-fluid solution will be on the order of the maximum between the two magnitudes of uncertainty (in supply and in demand). We also note that in an alternative setting, the uncertainty in the arrival rate may be realized before the staffing decision is finalized. In particular, additional flexible capacity may be called upon in the last minute to hedge against unforseen changes in the arrival rates, i.e., we can update our original staffing decision for the flexible servers after seeing the realized arrival rate. Then, flexible servers will bring an extra layer of benefit to the manager. We believe this setting is also of practical relevance and would be an interesting future research direction.

Lastly, in this work, we focus on exponential patience time distribution. We also demonstrate, through numerical experiments, that if the hazard rate of patience time distribution is increasing, the fluid prescriptions is extremely accurate, while if the hazard rate of patience time distribution is decreasing, the refined stochastic-fluid prescriptions yields orders of magnitude improvement over the fluid prescription. Thus, understanding the application-specific patience time distributions is also very important for making the right staffing decisions.

Acknowledgments

The authors are grateful to the department editors and the anonymous associated editor and reviewers for their constructive suggestions.

Appendix A: Proofs of the optimality gaps

In this section, we prove the optimality gaps results (Theorem 1, 2, 4 and 5). As we shall explain, Theorem 1 is actually a special case of Theorem 4, and Theorem 2 is a special case of Theorem 5.

We first state and prove some auxiliary lemmas (Lemma 3-5), which will be useful for the proof of our theorems. These lemmas are stated for the single-period case and allow for general values of $m_\lambda \geq 0$ and $n_\lambda \geq 0$. The proofs follow similar lines of the arguments as Bassamboo et al. (2010).

A.1. Additional lemmas for a single period

LEMMA 3. *When $\mu = \theta$,*

$$\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \leq \mathbb{E}[(X(m_\lambda, n_\lambda) - N(m_\lambda, n_\lambda))^+] \leq \mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] + \mathcal{O}(\sqrt{\lambda}),$$

where $N(m_\lambda, n_\lambda) = m_\lambda + n_\lambda + \sigma_{n_\lambda}\epsilon$. Moreover, if $n_\lambda = \Theta(\lambda)$ and $\sigma_\lambda = \Theta(\lambda^q)$ for $q > 1/2$ then:

$$\mathbb{E}[(X(m_\lambda, n_\lambda) - N(m_\lambda, n_\lambda))^+] \leq \mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] + \mathcal{O}\left(\lambda/\sigma_\lambda\right).$$

PROOF. When $\mu = \theta$, $X(m_\lambda, n_\lambda) \sim$Poisson$(\lambda/\mu)$. By Lemma 3 of Bassamboo et al. (2010):

$$\begin{aligned}
&\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+ \leq \mathbb{E}[(X(m_\lambda, n_\lambda) - N(m_\lambda, n_\lambda))^+ | N(m_\lambda, n_\lambda)] \\
&\leq \left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+ + \sqrt{\frac{4\pi\lambda}{\mu}} \exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^2\right) + \frac{1}{\log 2}
\end{aligned}$$

Then as $\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^2\right) \leq 1$, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] \leq\ & \mathbb{E}[X(m_\lambda, n_\lambda) - N(m_\lambda, n_\lambda))^+] \\
& \leq \mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}$$

For the second part of the lemma, we let $f_N^\lambda(s)$ denote the pdf of $N(m_\lambda, n_\lambda)$ and define, for $y \geq 0$:

$$M_\lambda(y) \equiv \sup_{y - \sqrt{\lambda}\log\lambda < s < y + \sqrt{\lambda}\log\lambda} \lambda f_N^\lambda(s).$$

We can then write:

$$
\begin{aligned}
&E\left[\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-N(m_\lambda,n_\lambda)\right)^2\right)\right]\\
&=\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\quad+\int_{\lambda/\mu+\sqrt{\lambda}\log\lambda}^{\infty}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\quad+\int_{-\infty}^{\lambda/\mu-\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds.
\end{aligned}
$$

Letting F_N^λ denote the cdf of $N(m_\lambda,n_\lambda)$, we see that: $F_N^\lambda(s)=\mathbb{P}\left(\epsilon\le\frac{s-m_\lambda-n_\lambda}{\sigma_{n_\lambda}}\right)=F_\epsilon\left(\frac{s-m_\lambda-n_\lambda}{\sigma_{n_\lambda}}\right)$. Thus, $f_N^\lambda(s)=\frac{1}{\sigma_{n_\lambda}}f_\epsilon\left(\frac{s-m_\lambda-n_\lambda}{\sigma_{n_\lambda}}\right)$. That is, when $n_\lambda=\Theta(\lambda)$, it must be that $M_\lambda(y)=\mathcal{O}(\lambda/\sigma_\lambda)$. Now,

$$
\begin{aligned}
&\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\le M_\lambda(\lambda/\mu)\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi}{\mu}}\frac{\sqrt{\lambda}}{\lambda}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)ds\\
&\le M_\lambda(\lambda/\mu)\int_{\lambda/\mu-\sqrt{\lambda}\log\lambda}^{\lambda/\mu+\sqrt{\lambda}\log\lambda}\frac{K_1}{\sqrt{\lambda}}\exp\left(-\frac{K_2}{\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)ds \text{ for some } K_1,K_2>0,\\
&=\mathcal{O}(\lambda/\sigma_\lambda).
\end{aligned}
$$

In addition,

$$
\begin{aligned}
&\int_{\lambda/\mu+\sqrt{\lambda}\log\lambda}^{\infty}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds\\
&\le\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\lambda(\log\lambda)^2\right)\\
&=o(1).
\end{aligned}
$$

Similarly,

$$
\int_{-\infty}^{\lambda/\mu-\sqrt{\lambda}\log\lambda}\sqrt{\frac{4\pi\lambda}{\mu}}\exp\left(-\frac{\mu}{4\lambda}\left(\frac{\lambda}{\mu}-s\right)^2\right)f_N^\lambda(s)ds=o(1).
$$

Thus, if $n_\lambda=\Theta(\lambda)$ then:

$$
\mathbb{E}\left[(X(m_\lambda,n_\lambda)-N(m_\lambda,n_\lambda))^+\right]\le\mathbb{E}\left[(\lambda/\mu-N(m_\lambda,n_\lambda))^+\right]+\mathcal{O}(\lambda/\sigma_\lambda).
$$

■

Lemma 4.

$$
\tilde{\Pi}_\lambda(m_\lambda,n_\lambda)\le\Pi_\lambda(m_\lambda,n_\lambda)\le\tilde{\Pi}_\lambda(m_\lambda,n_\lambda)+\mathcal{O}(\sqrt{\lambda})
$$

Moreover, when $n_\lambda=\Theta(\lambda)$ *and* $\sigma_\lambda=\Theta(\lambda^q)$ *for* $q>1/2$*:*

$$
\tilde{\Pi}_\lambda(m_\lambda,n_\lambda)\le\Pi_\lambda(m_\lambda,n_\lambda)\le\tilde{\Pi}_\lambda(m_\lambda,n_\lambda)+O\left(\lambda/\sigma_\lambda\right).
$$

PROOF. We prove the first statement in detail. The second statement, where $n_\lambda = \Theta(\lambda)$ and $q > 1/2$, follows along the same line of arguments.

When $\mu = \theta$, the result follows directly from Lemma 3.

When $\mu > \theta$**,** we first consider an auxiliary "upper bound" system with abandonment rate μ. On each sample path, we assume that the two systems have the same (randomly drawn) number of servers. Let $A_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \mu, \theta) - N(m_\lambda, n_\lambda))^+]$ and $A^I_\lambda(N(m_\lambda, n_\lambda)) \equiv \mu\mathbb{E}[(X(m_\lambda, n_\lambda; \mu, \mu) - N(m_\lambda, n_\lambda))^+]$ where $X(m_\lambda, n_\lambda; x, y)$ is the steady-state number-in-system with service rate x and abandonment rate y. As $A_\lambda(N(m_\lambda, n_\lambda)) \le A^I_\lambda(N(m_\lambda, n_\lambda))$:

$$\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda)) \\
&\le c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A^I_\lambda(N(m_\lambda, n_\lambda)) \\
&\le c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] + \mathcal{O}(\sqrt{\lambda}) \quad \text{by Lemma 3} \\
&= \tilde{\Pi}_\lambda(m_\lambda, n_\lambda) + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}$$

We then consider an auxiliary "lower bound" system with service rate θ. On each sample path, we assume that the two systems have the same (randomly drawn) number of servers. Let $A^{II}_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \theta, \theta) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \ge A^{II}_\lambda(N(m_\lambda, n_\lambda)\mu/\theta)$ (Bassamboo et al. 2010):

$$\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda\left(N(m_\lambda, n_\lambda)\right) \\
&\ge c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A^{II}_\lambda\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right) \\
&\ge c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)^+\right] \quad \text{by Lemma 3} \\
&= c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}\left[(\lambda/\mu - N(m_\lambda, n_\lambda))^+\right] \\
&= \tilde{\Pi}_\lambda(m_\lambda, n_\lambda).
\end{aligned}$$

When $\mu < \theta$**,** the proof is similar to the case of $\mu > \theta$. We first consider an auxiliary "upper bound" system with service rate θ. Let $A^{II}_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \theta, \theta) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \le A^{II}_\lambda\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)$:

$$\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda)) \\
&\le c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A^{II}_\lambda\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right) \\
&\le c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} m_\lambda - \frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)^+\right] + \mathcal{O}(\sqrt{\lambda}) \text{ by Lemma 3} \\
&= c_{fix} m_\lambda + c_{flex} n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \\
&= \tilde{\Pi}_\lambda(m_\lambda, n_\lambda) + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}$$

We then consider an auxiliary "lower upper" bound system with abandonment rate μ. Let $A_\lambda(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(N(m_\lambda, n_\lambda); \mu, \theta) - N(m_\lambda, n_\lambda))^+]$ and $A^I_\lambda(N(m_\lambda, n_\lambda)) \equiv \mu\mathbb{E}[(X(N(m_\lambda, n_\lambda); \mu, \mu) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \ge A^I_\lambda(N(m_\lambda, n_\lambda))$:

$$\Pi_\lambda(m_\lambda, n_\lambda) = c_{fix} m_\lambda + c_{flex} n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda))$$

$$
\begin{aligned}
&\geq c_{fix}m_\lambda + c_{flex}n_\lambda + (h/\theta + r)A^I_\lambda(N(m_\lambda, n_\lambda)) \\
&\geq c_{fix}m_\lambda + c_{flex}n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \quad \text{by Lemma 3} \\
&= \tilde{\Pi}_\lambda(m_\lambda, n_\lambda).
\end{aligned}
$$

■

LEMMA 5.

$$
\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+ \leq \mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] \leq \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+ + \mathcal{O}(\sigma_{n_\lambda}).
$$

PROOF. We notice that by Jensen's inequality,

$$
\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] \geq \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+.
$$

For the upper bound, as $-1 < \epsilon < 1$,

$$
\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] = \mathbb{E}\left[\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda - \sigma_{n_\lambda}\epsilon\right)^+\right] =
$$

$$
\begin{cases}
0 & \text{for } \frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}} < -1, \\
\sigma_{n_\lambda}\int_{-1}^{\frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}}} F_\epsilon(x)dx & \text{for } -1 \leq \frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}} \leq 1, \\
\lambda/\mu - m_\lambda - n_\lambda & \text{for } \frac{\lambda/\mu - m_\lambda - n_\lambda}{\sigma_{n_\lambda}} > 1.
\end{cases}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right] &= \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right) \cdot \mathbf{1}\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda > \sigma_{n_\lambda}\right) + \mathcal{O}(\sigma_{n_\lambda}) \\
&\leq \left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)^+ + \mathcal{O}(\sigma_{n_\lambda}).
\end{aligned}
$$

■

A.2. Proofs of Theorems 4 and 5

We denote

$$
\Pi^i_\lambda(m, n^i) := c_{fix}m + c_{flex}n + \beta E[Q^i_\lambda(m, n^i)].
$$

We also write $\bar{\Pi}^i_\lambda(m, n^i)$ and $\tilde{\Pi}^i_\lambda(m, n^i)$ as the corresponding fluid and stochastic-fluid approximations, respectively, for period i. We start with the solution to the stochastic-fluid relaxation, $(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda)$. From Lemma 4, we have:

$$
\begin{aligned}
\Pi_\lambda(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda) &= \sum_{i=1}^k T_i \Pi^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) \\
&\leq \sum_{i=1}^k T_i \left\{\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) + \mathcal{O}(\sqrt{\lambda})\right\} \\
&\leq \sum_{i=1}^k T_i \tilde{\Pi}^i_\lambda(m^*_\lambda, n^{*,i}_\lambda) + \mathcal{O}(\sqrt{\lambda}) \\
&\leq \sum_{i=1}^k T_i \Pi^i_\lambda(m^*_\lambda, n^{*,i}_\lambda) + \mathcal{O}(\sqrt{\lambda}) = \Pi_\lambda(m^*_\lambda, \mathbf{n}^*_\lambda) + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}
$$

Moreover, if $\tilde{n}_\lambda^{(m)} = \Theta(\lambda)$ and $\sigma_\lambda = \Theta(\lambda^q)$ for $q > 1/2$, from Lemma 4, we have:

$$\begin{aligned}
\Pi_\lambda(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda) &= \sum_{i=1}^{k} T_i \Pi_\lambda^i(\tilde{m}_\lambda, \tilde{n}_\lambda^i) \\
&\le \sum_{i=1}^{k} T_i \left(\tilde{\Pi}_\lambda^i(\tilde{m}_\lambda, \tilde{n}_\lambda^i) + \mathcal{O}(\lambda/\sigma_\lambda) \right) \\
&\le \tilde{\Pi}_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda) \\
&\le \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}(\lambda/\sigma_\lambda).
\end{aligned}$$

We next analyze the solution to the fluid relaxation, $(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda)$. From Lemmas 4 & 5, we have:

$$\begin{aligned}
\Pi_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) &= \sum_{i=1}^{k} T_i \Pi_\lambda^i(\bar{m}_\lambda, \bar{n}_\lambda^i) \\
&\le \sum_{i=1}^{k} T_i \left\{ \tilde{\Pi}_\lambda^i(\bar{m}_\lambda, \bar{n}_\lambda^i) + \mathcal{O}(\sqrt{\lambda}) \right\} \\
&\le \sum_{i=1}^{k} T_i \left\{ \bar{\Pi}_\lambda^i(\bar{m}_\lambda, \bar{n}_\lambda^i) + \mathcal{O}(\sqrt{\lambda}) + \mathcal{O}(\sigma_{\bar{n}_\lambda^i}) \right\} \\
&\le \sum_{i=1}^{k} T_i \bar{\Pi}_\lambda^i(m_\lambda^*, n_\lambda^{*,i}) + \mathcal{O}(\sqrt{\lambda}) + \mathcal{O}(\sigma_{\bar{n}_\lambda^{(M)}}) \\
&\le \tilde{\Pi}_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}(\sqrt{\lambda}) + \mathcal{O}(\sigma_{\bar{\mathbf{n}}_\lambda^{(\mathrm{M})}}) \\
&\le \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}(\sqrt{\lambda}) + \mathcal{O}(\sigma_{\bar{\mathbf{n}}_\lambda^{(\mathrm{M})}}).
\end{aligned}$$

In particular, if $\bar{n}_\lambda^{(M)} = \Theta(\lambda)$ then

$$\Pi_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) \le \Pi_\lambda(m_\lambda^*, \mathbf{n}_\lambda^*) + \mathcal{O}(\max\{\sqrt{\lambda}, \sigma_\lambda\}).$$

This concludes the proofs for Theorems 4 and 5.

Theorem 1 and Theorem 2 are essentially special cases of Theorem 4 and 5, respectively, so we do not include separate proofs for them. In particular, when we set $c_{fix} = \infty$, we obtain $\bar{m}_\lambda = 0$, i.e., we shall rely on the flexible servers only. This is equivalent to the cases considered in Theorem 1 and Theorem 2. Lastly, note that from the analysis in §3 with flexible servers only, for each period i, we have $\tilde{n}_\lambda^i = \Theta(\lambda)$.

Appendix B: Proofs for the optimal staffing rule with flexible servers only

In this section, we prove the results for the optimal staffing policy with flexible servers only (Lemma 1 and Theorem 3).

B.1. Proof of Lemma 1

The proof builds on extensive applications of Taylor expansion and the mean value theorem. We first notice that from $\tilde{\Pi}'_\lambda(\tilde{n}_\lambda) = 0$, we have

$$\tilde{n}_\lambda = \frac{\lambda}{\mu} - F_\epsilon^{-1}\left(\frac{c_{flex}}{\beta} - \sigma'_{\tilde{n}_\lambda} \int_{-1}^{\frac{\lambda/\mu - \tilde{n}_\lambda}{\sigma_{\tilde{n}_\lambda}}} x f_\epsilon(x) dx \right) \sigma_{\tilde{n}_\lambda}.$$

We define $\gamma_\lambda \equiv F_\epsilon^{-1}\left(\frac{c_{flex}}{\beta} - \sigma'_{\tilde{n}_\lambda}\int_{-1}^{(\lambda/\mu-\tilde{n}_\lambda)/\sigma_{\tilde{n}_\lambda}} x f_\epsilon(x)dx\right)$. As $\sigma'_{\tilde{n}_\lambda}\int_{-1}^{(\lambda/\mu-\tilde{n}_\lambda)/\sigma_{\tilde{n}_\lambda}} x f_\epsilon(x)dx \le 0$, and it converges to zero as $\lambda\to\infty$, we have $\gamma_\lambda \ge \gamma$ and $\gamma_\lambda \to \gamma$ as $\lambda\to\infty$.

$$\begin{aligned}\hat{n}_\lambda - \tilde{n}_\lambda &= -\gamma\sigma_{\lambda/\mu} + \gamma_\lambda\sigma_{\tilde{n}_\lambda}\\ &= -\gamma(\sigma_{\lambda/\mu} - \sigma_{\tilde{n}_\lambda}) + (\gamma_\lambda - \gamma)\sigma_{\tilde{n}_\lambda}. \end{aligned} \tag{25}$$

Based on (25), we notice that, as

$$\sigma_{\lambda/\mu} - \sigma_{\tilde{n}_\lambda} = \mathcal{O}(\sigma'_{\lambda/\mu}\sigma_{\lambda/\mu}) = O(\sigma_\lambda^2/\lambda),$$

and

$$\begin{aligned}\gamma_\lambda - \gamma &= \frac{1}{f_\epsilon(\gamma_0)}\sigma'_{\tilde{n}_\lambda}\int_{-1}^{\frac{\lambda/\mu-\tilde{n}_\lambda}{\sigma_{\tilde{n}_\lambda}}} x f_\epsilon(x)dx \quad \text{for some } \gamma_0\in[\gamma_\lambda,\gamma],\\ &= \mathcal{O}(\sigma'_{\lambda/\mu}) = \mathcal{O}(\sigma_\lambda/\lambda),\end{aligned}$$

we have

$$|\hat{n}_\lambda - \tilde{n}_\lambda| = \mathcal{O}(\sigma_\lambda^2/\lambda).$$

We next analyze the gap between $\tilde{\Pi}_\lambda(\tilde{n}_\lambda)$ and $\tilde{\Pi}_\lambda(\hat{n}_\lambda)$. Denote

$$M \equiv \sup_\lambda \sup_{n-\sigma_n<\lambda<n+\sigma_n} \sigma_{\lambda/\mu}\tilde{\Pi}''_\lambda(n) \in (0,\infty).$$

We note that

$$\begin{aligned}\tilde{\Pi}_\lambda(\hat{n}_\lambda) &= \tilde{\Pi}_\lambda(\tilde{n}_\lambda) + \tilde{\Pi}'_\lambda(\tilde{n}_\lambda)(\hat{n}_\lambda - \tilde{n}_\lambda) + \frac{1}{2}\tilde{\Pi}''_\lambda(n_0)(\hat{n}_\lambda - \tilde{n}_\lambda)^2 \quad \text{for some } n_0\in[\hat{n}_\lambda,\bar{n}_\lambda],\\ &\le \tilde{\Pi}_\lambda(\tilde{n}_\lambda) + \frac{1}{2}\frac{M}{\sigma_{\lambda/\mu}}(\hat{n}_\lambda - \tilde{n}_\lambda)^2. \end{aligned} \tag{26}$$

As $|\hat{n}_\lambda - \tilde{n}_\lambda| = \mathcal{O}(\sigma_\lambda^2/\lambda)$, (26) implies

$$\bar{\Pi}_\lambda(\hat{n}_\lambda) - \tilde{\Pi}_\lambda(\tilde{n}_\lambda) = \mathcal{O}(\sigma_\lambda^3/\lambda^2).$$

B.2. Proof of Theorem 3

Recall that we denote by $\tilde{n}_\lambda$ the optimal solution of $\tilde{\Pi}_\lambda(n)$, and n^*_λ denotes the optimal solution of $\Pi_\lambda(n)$.

Case I. $0 \le q \le 1/2$. Plugging in the fluid optimal solution, $n_\lambda = \bar{n}_\lambda = \lambda/\mu$, we have

$$\Pi_\lambda(n_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\max\{\sqrt{\lambda},\sigma_\lambda\}) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\sqrt{\lambda}),$$

where the first equality follows from Theorem 1.

Case II. $1/2 < q < 3/4$. From Lemma 1 and Theorem 2, we have for large enough systems,

$$\begin{aligned}\tilde{\Pi}_\lambda(\hat{n}_\lambda) &= \tilde{\Pi}_\lambda(\tilde{n}_\lambda) + \mathcal{O}(\sigma_\lambda^3/\lambda^2),\\ &= \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda) + \mathcal{O}(\sigma_\lambda^3/\lambda^2),\\ &= \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda).\end{aligned}$$

Case III. $3/4 < q < 1$. From Theorem 2, we have:

$$\tilde{\Pi}_\lambda(\tilde{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda).$$

We also note from Lemma 1 that if we set $n_\lambda = \hat{n}_\lambda$ in this case, then the error term $\sigma_\lambda^3/\lambda^2$ will dominate the error term λ/σ_λ, i.e., we are no longer able to achieve $\mathcal{O}(\lambda/\sigma_\lambda)$ optimality.

Case IV. $q=1$. From the analysis in Section 3 and the definition of Ω_λ, we have the optimal

$$\eta^* \in \left(\frac{1}{1+a}, \frac{1}{1-a}\right).$$

Let

$$g(\eta) = \frac{\tilde{\Pi}_\lambda(\frac{\lambda}{\mu}\eta)}{\lambda/\mu} = c_{flex}\eta + \beta\mathbb{E}[(1-\eta-a\eta\epsilon)^+].$$

Then,

$$g'(\eta) = c_{flex} + \beta a\int_{-1}^{(1-\eta)/(a\eta)} F_\epsilon(u)du - \frac{\beta}{\eta}F_\epsilon\left(\frac{1-\eta}{a\eta}\right) \quad \text{and} \quad g''(\eta) = \frac{\beta}{a\eta^3}f_\epsilon\left(\frac{1-\eta}{a\eta}\right) > 0.$$

Thus, the η^* that minimizes $g(\eta)$ is the solution of $g'(\eta)=0$. In this case, $\tilde{n}_\lambda = \frac{\lambda}{\mu}\eta^*$. From Theorem 2, we have

$$\bar{\Pi}_\lambda(\tilde{n}_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(\lambda/\sigma_\lambda) = \Pi_\lambda(n^*_\lambda) + \mathcal{O}(1).$$

Appendix C: Proof for the optimal staffing rule with a blended workforce

In this section, we prove the results for the optimal staffing policy with a blended workforce. We first prove the optimal solution for the fluid approximation (Lemma 2). We then prove Theorem 6.

C.1. Proof of Lemma 2.

If $c_{flex} < c_{fix}$, then $n^i_\lambda = \lambda_i/\mu$ for all i, and $m=0$.

Now consider $c_{flex} > c_{fix}$. Recall that $c_{flex}, c_{fix} < \beta$, and λ_i's are arranged in an increasing order.

Fix m_λ and solve for n^i_λ. The problem for period i where we drop T_i is:

$$\min_{n^i_\lambda} c_{flex}n^i_\lambda + \beta(\lambda_i - m^i_\lambda - n^i_\lambda)^+.$$

The solution is:

- If $\lambda_i/\mu < m_\lambda$ then $n^i_\lambda = 0$;
- If $\lambda_i/\mu \geq m_\lambda$ then $n^i_\lambda = \lambda_i/\mu - m_\lambda$.

Solve for m_λ. Plugging in the above solution, the problem becomes:

$$\begin{aligned}
&\min_{m_\lambda} \quad \left(\sum_{\{i:\lambda_i/\mu<m_\lambda\}} T_i c_{fix} m_\lambda + \sum_{\{i:\lambda_i/\mu\geq m_\lambda\}} T_i\left[c_{fix}m_\lambda + c_{flex}\left(\frac{\lambda_i}{\mu} - m_\lambda\right)\right]\right)\\
\equiv &\min_{m_\lambda} \quad \left(\sum_{i=1}^{k} T_i c_{fix} m_\lambda - \sum_{\{i:\lambda_i/\mu\geq m_\lambda\}} T_i c_{flex} m_\lambda + \sum_{\{i:\lambda_i/\mu\geq m_\lambda\}} T_i c_{flex}\frac{\lambda_i}{\mu}\right)\\
\equiv &\min_{m_\lambda} \quad \left(m_\lambda \cdot \left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{\{i:\lambda_i/\mu\geq m_\lambda\}} T_i c_{flex}\right) + \sum_{\{i:\lambda_i/\mu\geq m_\lambda\}} T_i c_{flex}\frac{\lambda_i}{\mu}\right).
\end{aligned}$$

It is easy to see that there exist $1 \leq k_0 \leq k$ such that:

$$\left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{i=k_0}^{k} T_i c_{flex}\right) \leq 0 \quad \text{and} \quad \left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{i=k_0+1}^{k} T_i c_{flex}\right) > 0.$$

Then the optimal solution is given by:

$$\begin{cases}
\bar{m}_\lambda = \lambda_{k_0}/\mu & \\
\bar{n}^i_\lambda = 0 & \text{for } i \leq k_0\\
\bar{n}^i_\lambda = \lambda_i/\mu - \lambda_{k_0}/\mu & \text{for } i > k_0.
\end{cases}$$

That is, we use only the fixed capacity in low-demand periods, and we blend in the higher-demand periods. In the special case when $c_{fix} \leq \frac{T_k}{\sum_{i=1}^k T_i} c_{flex}$, we set $\bar{m}_\lambda = \lambda_k/\mu$ and $\bar{n}^i_\lambda = 0$ for all i.

C.2. Proof of Theorem 6.

The proof is divided into two part. We first prove Theorem 6 for a single period, $k=1$. We then show how to decompose the multi-period problem into k, $k\geq 2$, single period problems based on the fluid optimal solution derived in Lemma 2 and properly adjusted arrival rate.

C.2.1. Single period

$c_{fix}\leq c_{flex}$. In this case, we must have that $\tilde{n}_\lambda=0$. Aiming at a contradiction, assume that $\tilde{n}_\lambda>0$. Then,

$$\begin{aligned}\tilde{\Pi}_\lambda(\tilde{m}_\lambda+\tilde{n}_\lambda,0) &= c_{fix}(\tilde{m}_\lambda+\tilde{n}_\lambda)+\beta(\lambda/\mu-\tilde{m}_\lambda-\tilde{n}_\lambda)^+\\ &< c_{fix}\tilde{m}_\lambda+c_{flex}\tilde{n}_\lambda+\beta\mathbb{E}[(\lambda/\mu-\tilde{m}_\lambda-\tilde{n}_\lambda-\sigma_{\tilde{n}_\lambda}\cdot\epsilon)^+]=\tilde{\Pi}_\lambda(\tilde{m}_\lambda,\tilde{n}_\lambda).\end{aligned}$$

The inequality follows from the fact that $c_{fix}<c_{flex}$ and Jensen's inequality. Fixing $\tilde{n}=0$, we have $\tilde{\Pi}(m,0)=\bar{\Pi}(m,0)$. Thus, a unique optimal solution exists, and is the solution to the fluid problem.

$c_{fix}>c_{flex}$ and $0\leq q\leq 1/2$. Plugging in the fluid optimal solution, i.e., $\bar{m}_\lambda=0$ and $\bar{n}_\lambda=\lambda/\mu$, we have

$$\Pi_\lambda(\bar{m}_\lambda,\bar{n}_\lambda)=\Pi_\lambda(m^*_\lambda,n^*_\lambda)+\mathcal{O}(\max\{\sqrt{\lambda},\sigma_\lambda\})=\Pi(m^*_\lambda,n^*_\lambda)+\mathcal{O}(\sqrt{\lambda}),$$

where the first equality follows from Theorem 4.

$c_{fix}>c_{flex}$ and $1/2<q<1$. We first fixed n. Denote $m(n)$ as the optimal level of fixed staffing for the given level of flexible staffing, n. As

$$\frac{\partial\Pi_\lambda(m,n)}{\partial m}=c_{fix}-\beta F_\epsilon\left(\frac{\lambda/\mu-m-n}{\sigma_n}\right)\text{ and }\frac{\partial^2\Pi_\lambda(m,n)}{\partial m^2}=\frac{\beta}{\sigma_n}f_\epsilon\left(\frac{\lambda/\mu-m-n}{\sigma_n}\right)\geq 0,$$

if $\frac{\partial\Pi_\lambda(0,n)}{\partial m}>0$, i.e. $n>\lambda/\mu-F_\epsilon^{-1}(c_{fix}/\beta)\sigma_n$, $m(n)=0$; otherwise, $m(n)=\lambda/\mu-n-F_\epsilon^{-1}(c_{fix}/\beta)\sigma_n$.

We next plug $m(n)$ in $\Pi_\lambda(m,n)$. Let $n_b=\lambda/\mu-F_\epsilon^{-1}(c_{fix}/\beta)\sigma_{n_b}$. For $n<n_b$, $n<\lambda/\mu-F_\epsilon^{-1}(c_{fix}/\beta)\sigma_n$. In this region, we have

$$\frac{\partial\Pi_\lambda(m(n),n)}{\partial n}=c_{flex}-c_{fix}-\sigma_n'c_{fix}F_\epsilon^{-1}(c_{fix}/\beta)$$

As $\sigma_n'\to 0$ as $n\to\infty$ and $c_{flex}<c_{fix}$, then for λ large enough, $\frac{\partial\Pi_\lambda(m(n),n)}{\partial n}>0$ for all $n<n_b$. This suggests that $\bar{n}_\lambda\geq n_b$. We also notice that for $n\geq n_b$, $m(n)=0$. Thus, $\tilde{m}_\lambda=0$ for λ large enough. In this case, solving for the optimal $\tilde{n}_\lambda$ reduces to the flexible resource only problem we studied in Case II and III of Theorem 3.

$c_{fix}>c_{flex}$ and $q=1$. Let $m_\lambda=x\lambda/\mu$, $n_\lambda=y\lambda/\mu$ for $x,y\in\mathbb{R}^+$. Then, $\min_{m_\lambda,n_\lambda}\tilde{\Pi}_\lambda(m_\lambda,n_\lambda)$ is equivalent to optimizing:

$$\min_{x,y}V(x,y):=\frac{\tilde{\Pi}_\lambda(m_\lambda,n_\lambda)}{\lambda/\mu}=c_{fix}x+c_{flex}y+\beta\mathbb{E}\left[(1-x-y-ay\epsilon)^+\right].$$

We denote the optimal solution to $V(x,y)$ as x^*, y^*. It suffices to consider x,y such that $-ay\leq 1-x-y\leq ay$. In this case, we first notice that for fixed y, we have

$$\frac{\partial V(x,y)}{\partial x}=c_{fix}-\beta F_\epsilon\left(\frac{1-x-y}{ay}\right)\text{ and }\frac{\partial^2V(x,y)}{\partial x^2}=\beta\frac{1}{ay}f_\epsilon\left(\frac{1-x-y}{ay}\right)\geq 0$$

Let $x(y)$ denote the optimal x given y. Then if $y<\frac{1}{1+aF_\epsilon^{-1}(c_{fix}/\beta)}$, $x(y)=1-y-ayF_\epsilon^{-1}(\beta_1)$; Otherwise, $x(y)=0$.

We next plug $x(y)$ in $V(x,y)$. Let $y_b = \frac{1}{1+aF_\epsilon^{-1}(c_{fix}/\beta)}$. For $y \le y_b$, we have

$$\frac{\partial V(x(y),y)}{\partial y} = c_{flex} - c_{fix} - c_{fix} a F_\epsilon^{-1}(\beta_1) + a\beta \int_{-1}^{F_\epsilon^{-1}(\beta_1)} F_\epsilon(u)du.$$

If $c_{flex} \ge c_{fix}\left(1 + aF_\epsilon^{-1}(c_{fix}/\beta)\right) - \beta a \int_{-1}^{F_\epsilon^{-1}(c_{fix}/\beta)} F_\epsilon(x)dx$, it is optimal to set $y = 0$ and $x(y) = 1$ in this region. Otherwise, it is optimal to set $y = y_b$ and $x(y) = 0$ in this region. Note also that as $c_{fix}\left(1 + aF_\epsilon^{-1}(c_{fix}/\beta)\right) - \beta a \int_{-1}^{F_\epsilon^{-1}(c_{fix}/\beta)} F_\epsilon(x)dx < c_{fix}$ and $c_{flex} < c_{fix}$, both cases above can happen.

Now let us consider $y > y_b$. In this region, we have

$$\frac{\partial^2 V(x(y),y)}{\partial y^2} = \frac{\beta}{ay^3} f_\epsilon\left(\frac{1}{ay} - \frac{1}{a}\right) > 0.$$

This implies that $V(x(y),y) = V(0,y)$ is convex in y for $y \ge y_b$. We also notice that if

$$c_{flex} \ge c_{fix}\left(1 + aF_\epsilon^{-1}(c_{fix}/\beta)\right) - \beta a \int_{-1}^{F_\epsilon^{-1}(c_{fix}/\beta)} F_\epsilon(x)dx,$$

$\frac{\partial V(x(y_b),y_b)}{\partial y} \ge 0$. Combining our analysis in the region where $y \le y_b$, we conclude that $y^* = 0$ and $x^* = 1$. If

$$c_{flex} < c_{fix}\left(1 + aF_\epsilon^{-1}(c_{fix}/\beta)\right) - \beta a \int_{-1}^{F_\epsilon^{-1}(c_{fix}/\beta)} F_\epsilon(x)dx,$$

$\frac{\partial V(x(y_b),y_b)}{\partial y} < 0$. Combining our analysis in the region where $y \le y_b$, we conclude that $y^* > y_b$ and $x^* = 0$. Then in this case, solving for the optimal y^* reduces to the flexible resource only problem we studied in Case IV of Theorem 3 .

C.2.2. Multiple periods

Case 1. $\sigma_n = an^q$ for $0 < q < 1$. For $\mathbf{n}_\lambda = (n_\lambda^1, n_\lambda^2, ..., n_\lambda^k)$, we can write:

$$\begin{aligned}
\tilde{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) &= \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \beta E\left[\left(\frac{\lambda_i}{\mu} - N(m_\lambda, n_\lambda^i)\right)^+\right]\right), \\
&=: \sum_{i=1}^k T_i \tilde{\Pi}_\lambda^i(m_\lambda, n_\lambda^i), \\
&= \sum_{i=1}^k T_i \left(c_{fix} m_\lambda + c_{flex} n_\lambda^i + \beta 1\{n_\lambda^i > 0\} \sigma_{n_\lambda^i} \int_{-1}^{\frac{\lambda_i/\mu - m_\lambda - n_\lambda^i}{\sigma_{n_\lambda^i}}} F_\epsilon(u)du \right).
\end{aligned}$$

When plugging the fluid solution $(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda)$, we have

$$\tilde{\Pi}_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) = \bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{\mathbf{n}}_\lambda) + \beta \sum_{i=k_0+1}^k \sigma_{\bar{n}_\lambda^i} \int_{-1}^0 F_\epsilon(u)du, \tag{27}$$

where k_0 is defined in Lemma 2. The expression in (27) suggests that the optimal policy for the stochastic-fluid problem will have the same number of fixed servers as its counterpart fluid solution, but will add some flexible resource to its counterpart fluid solution for an additional "hedge" against uncertainty; this hedge should be on the order of σ_λ. We will prove that this is indeed the case next, breaking down our proof into 3 steps.

Step 1. For all $1 \le i \le k$:

(a) If $\tilde{m}_\lambda < \lambda_i/\mu$, then $\lambda_i/\mu - \sigma_{\tilde{n}_\lambda^i} \le \bar{m}_\lambda + \bar{n}_\lambda^i \le \lambda_i/\mu + \sigma_{\tilde{n}_\lambda^i}$;

(b) If $\tilde{m}_\lambda \geq \lambda_i/\mu$, then $\tilde{n}^i_\lambda = 0$.

The second part of the statement is straightforward. We prove the first part of the statement by contradiction.

i) Aiming at a contradiction, suppose that, for period i, we have $\tilde{m}_\lambda + \tilde{n}^i_\lambda > \lambda_i/\mu + \sigma_{\tilde{n}^i_\lambda}$. Choose n' such that $\tilde{m}_\lambda + n' = \lambda_i/\mu + \sigma_{n'}$. Then $n' < \tilde{n}_\lambda$ and

$$\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) = c_{fix}\tilde{m} + c_{flex}\tilde{n}^i_\lambda > c_{fix}\tilde{m} + c_{flex}n' = \tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, n'),$$

where the first and last equalities hold because $\epsilon \in (-1, 1)$. We thus get a contradiction.

ii) Now suppose that, for period i, $\tilde{m}_\lambda + \tilde{n}^i_\lambda < \lambda_i/\mu - \sigma_{\tilde{n}^i_\lambda}$. Choose n' such that $\tilde{m}_\lambda + n' = \lambda_i/\mu - \sigma_{n'}$. Notice that $n' > \tilde{n}^i_\lambda$ as σ_n is increasing in n. Then,

$$\begin{aligned}\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) &= c_{fix}\tilde{m}_\lambda + c_{flex}\tilde{n}^i_\lambda + \beta\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - \tilde{n}^i_\lambda\right) \\ &> c_{fix}\tilde{m}_\lambda + c_{flex}n' + \beta\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - n'\right) = \tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, n').\end{aligned}$$

The inequality follows from the fact that $c_{flex} < \beta$ and $n' > \bar{n}^i_\lambda$. In this case, by increasing $\tilde{n}^i_\lambda$ to n', we reduce the cost in period i without changing the cost of any other period. We thus get a contradiction. This concludes the proof of Step 1.

Step 2. $\tilde{m}_\lambda \geq \lambda_{k_0}/\mu$**.** Recall k_0 is defined such that $\bar{m}_{k_0} = \lambda_{k_0}/\mu$. We, again, prove this statement by contradiction. Aiming at a contradiction, suppose that $\tilde{m}_\lambda < \lambda_{k_0}/\mu$. For $i \leq k_0$, let $n^{i\prime}_\lambda = 0$; and for $k_0 < i \leq k$, let $n^{i\prime}_\lambda = \max\{\tilde{n}^i_\lambda - (\bar{m}_\lambda - \tilde{m}_\lambda), 0\}$. For $i < k_0$, we have

$$\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) \geq c_{fix}\tilde{m}_\lambda.$$

For $i = k_0$, let $x = \lambda_{k_0}/\mu - \tilde{m}_\lambda - \tilde{n}^i_\lambda = \bar{m}_\lambda - \tilde{m}_\lambda - \tilde{n}^i_\lambda$. Then we have

$$\begin{aligned}&\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) - c_{fix}\tilde{m}_\lambda \\ &= c_{flex}\tilde{n}^i_\lambda + \beta\sigma_{\tilde{n}^i_\lambda}\int_{-1}^{\frac{\lambda_i/\mu - \tilde{m}_\lambda - \tilde{n}^i_\lambda}{\sigma_{\tilde{n}^i_\lambda}}} F(u)du \\ &= c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) - c_{flex}x + \beta\sigma_{\tilde{n}^i_\lambda}\int_{-1}^{\frac{x}{\sigma_{\tilde{n}^i_\lambda}}} F(u)du \\ &\geq c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) - c_{flex}\sigma_{\tilde{n}^i_\lambda}F^{-1}(c_{flex}/\beta) + \beta\sigma_{\tilde{n}^i_\lambda}\int_{-1}^{F^{-1}(c_{flex}/\beta)} F(u)du \\ &= c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) - \beta\sigma_{\tilde{n}^i_\lambda}\int_{-1}^{F^{-1}(c_{flex}/\beta)} uf(u)du \\ &> c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda).\end{aligned}$$

Thus,

$$\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) \geq c_{fix}\tilde{m}_\lambda + c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda).$$

For $i > k_0$, from Step 1, we have $\tilde{m}_\lambda + \tilde{n}^i_\lambda > \lambda_i/\mu - \sigma_{\tilde{n}^i_\lambda}$ and $\tilde{n}^i_\lambda < \lambda_i/\mu + \sigma_{\tilde{n}^i_\lambda}$. Then for λ large enough, we have: $\tilde{n}^i_\lambda - (\bar{m}_\lambda - \tilde{m}_\lambda) > \lambda_i/\mu - \lambda_{k_0}/\mu - \sigma_{\tilde{n}^i_\lambda} > 0$. Thus, $n^{i\prime}_\lambda = \tilde{n}^i_\lambda - (\tilde{m}_\lambda - \bar{m}_\lambda)$ for λ large enough, and

$$\begin{aligned}
\tilde{\Pi}^i_\lambda(\tilde{m}_\lambda, \tilde{n}^i_\lambda) &= c_{fix}\tilde{m}_\lambda + c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) + c_{flex} n^{i\prime}_\lambda + \beta\left(\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - \tilde{n}^i_\lambda - \sigma_{\tilde{n}^i_\lambda}\epsilon\right)^+\right]\right) \\
&> c_{fix}\tilde{m}_\lambda + c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) + c_{flex} n^{i\prime}_\lambda + \beta\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - \tilde{n}^i_\lambda - \sigma_{n^{i\prime}_\lambda}\epsilon\right)^+\right] \\
&= c_{fix}\tilde{m}_\lambda + c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) + c_{flex} n^{i\prime}_\lambda + \beta\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - n^{i\prime}_\lambda - \sigma_{n^{i\prime}_\lambda}\epsilon\right)^+\right]
\end{aligned}$$

where the last inequality follows from the fact that for fixed $s \geq 0$, $\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - s - \sigma_n\epsilon\right)^+\right]$ is increasing in n, and the last equality follows from the fact that $\tilde{m}_\lambda + \tilde{n}^i_\lambda = \bar{m}_\lambda + n^{i\prime}_\lambda$.

From the proof of Lemma 2, we have $\sum_{i=1}^k T_i c_{fix} \leq \sum_{i=k_0}^k T_i c_{flex}$. Combining the bound for different values of i, we have

$$\begin{aligned}
&\tilde{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{\mathbf{n}}_\lambda) \\
&= \sum_{i=1}^k T_i \left(c_{fix}\tilde{m}_\lambda + c_{flex}\tilde{n}^i_\lambda + \beta\cdot\mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - \tilde{n}^i_\lambda - \sigma_{\tilde{n}^i_\lambda}\epsilon\right)^+\right]\right) \\
&> \sum_{i=1}^k T_i c_{fix}\bar{m}_\lambda - \sum_{i=1}^k T_i c_{fix}(\bar{m}_\lambda - \tilde{m}_\lambda) + \sum_{i=k_0}^k T_i c_{flex}(\bar{m}_\lambda - \tilde{m}_\lambda) + \sum_{i=k_0+1}^k T_i c_{flex} n^{i\prime}_\lambda \\
&\quad + \beta\sum_{i=k_0+1}^k \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \tilde{m}_\lambda - n^{i\prime}_\lambda - \sigma_{n^{i\prime}_\lambda}\epsilon\right)^+\right] \\
&\geq \sum_{i=1}^k T_i c_{fix}\bar{m}_\lambda + \sum_{i=k_0+1}^k T_i c_{flex} n^{i\prime}_\lambda + \beta\sum_{i=k_0+1}^k \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - \bar{m}_\lambda - n^{i\prime}_\lambda - \sigma_{n^{i\prime}}\epsilon\right)^+\right] \\
&= \tilde{\Pi}_\lambda(\bar{m}_\lambda, \mathbf{n}'_\lambda).
\end{aligned}$$

We therefore get a contradiction.

Step 3. $\tilde{m}_\lambda = \bar{m}_\lambda$**.** To show this, we denote $\hat{m}_\lambda = m_\lambda - \bar{m}_\lambda$ and $\hat{\mathbf{n}}_\lambda = (n^{k_0+1}_\lambda, \ldots, n^k_\lambda)$. Based on Step 2, we can write

$$\begin{aligned}
&\min_{m_\lambda \geq 0, \mathbf{n}_\lambda \geq 0} \bar{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \\
&= c_{fix}\bar{m}_\lambda \sum_{i=1}^k T_i \\
&\quad + \min_{\hat{m}_\lambda \geq 0, \hat{\mathbf{n}}_\lambda \geq 0} \sum_{i=k_0+1}^k T_i \left(\left(c_{fix}\frac{\sum_{j=1}^k T_j}{\sum_{i=k_0+1}^k T_i}\right)\hat{m}_\lambda + c_{flex} n^i_\lambda + \beta\mathbb{E}\left[(\lambda_i/\mu - \bar{m}_\lambda - N(\hat{m}_\lambda, n^i_\lambda))^+\right]\right)
\end{aligned}$$

As $c_{fix}\sum_{j=1}^k T_j > c_{flex}\sum_{i=k_0+1}^k T_i$, we must have at optimum that $\hat{m}_\lambda = 0$. To see why, we notice that:

$$\begin{aligned}
&\min_{\hat{m}_\lambda \geq 0, \hat{\mathbf{n}}_\lambda \geq 0} \sum_{i=k_0+1}^k T_i \left(\left(c_{fix}\frac{\sum_{j=1}^k T_j}{\sum_{i=k_0+1}^k T_i}\right)\hat{m}_\lambda + c_{flex} n^i_\lambda + \beta\mathbb{E}\left[(\lambda_i/\mu - \bar{m}_\lambda - N(\hat{m}_\lambda, n^i_\lambda))^+\right]\right) \\
&\geq \sum_{i=k_0+1}^k T_i \min_{m^i_\lambda \geq 0, n^i_\lambda \geq 0} \left\{\tilde{c}_{fix} m^i_\lambda + c_{flex} n^i_\lambda + \beta\mathbb{E}\left[(\lambda_i/\mu - \bar{m}_\lambda - N(m^i_\lambda, n^i_\lambda))^+\right]\right\} \\
&= \sum_{i=k_0+1}^k T_i \min_{n^i_\lambda \geq 0} \left\{c_{flex} n^i_\lambda + \beta\mathbb{E}\left[(\lambda_i/\mu - \bar{m}_\lambda - n^i_\lambda - \sigma_{n^i_\lambda}\epsilon)^+\right]\right\}
\end{aligned}$$

where $\tilde{c}_{fix} = c_{fix} \sum_{j=1}^{k} T_j / \sum_{i=k_0+1}^{k} T_i > c_{flex}$.

To sum up, the above analysis suggests that $\tilde{m}_\lambda = \bar{m}_\lambda$ and $\tilde{n}_\lambda^i = 0$ for $n \le k_0$. Thus, we can fully decompose the stochastic-fluid optimization problem into $k - k_0$ single period problems with the number of fixed servers equal to $\bar{m}_\lambda$. In particular,

$$\min_{m_\lambda \ge 0, \mathbf{n}_\lambda \ge 0} \tilde{\Pi}_\lambda(m_\lambda, \mathbf{n}_\lambda) \equiv c_{fix} \bar{m}_\lambda \sum_{i=1}^{k} T_i + \sum_{i=k_0+1}^{k} T_i \min_{n_\lambda^i \ge 0} \left\{ c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda^i - \sigma_{n_\lambda^i} \epsilon)^+ \right] \right\}.$$

This includes the special case where $k_0 = 0$ and $\tilde{m}_\lambda = 0$. Note that solving

$$\min_{n_\lambda^i \ge 0} \left\{ c_{flex} n_\lambda^i + \beta \mathbb{E}\left[(\lambda_i/\mu - \bar{m}_\lambda - n_\lambda^i - \sigma_{n_\lambda^i} \epsilon)^+ \right] \right\}$$

is essentially solving a single period flexible server only problem with arrival rate adjusted to $\lambda_i/\mu - \bar{m}_\lambda$. Thus we can use results from Case I, II and III in Theorem 3.

Case 2. $\sigma_n = an$ for $a < 1$. Recall that

$$g(\eta) = \frac{\tilde{\Pi}_\lambda(\frac{\lambda}{\mu}\eta)}{\lambda/\mu} = c_{flex}\eta + \beta E[(1 - \eta - a\eta\epsilon)^+]$$

and

$$g'(\eta) = c_{flex} + \beta a \int_{-1}^{(1-\eta)/(a\eta)} F_\epsilon(u) du - \frac{\beta}{\eta} F_\epsilon\left(\frac{1-\eta}{a\eta} \right).$$

Let η^* denote the solution of $g'(\eta) = 0$. For a fixed value of m, we can solve for the corresponding optimal $n(m)$ by optimizing each period individually. Particularly, for period i, we choose $n_i(m)$ that minimizes

$$\tilde{\Pi}_\lambda^i(m, n_i(m)) = c_{fix} m + c_{flex} n_i(m) + \beta \mathbb{E}\left[\left(\frac{\lambda_i}{\mu} - N(m, n_i(m)) \right)^+ \right].$$

Treating $\lambda_i/\mu - m$ as the new arrival rate, from our analysis for the single period flexible resource only analysis (Case IV in Theorem 3), we have $n_i(m) = \eta^*(\lambda_i/\mu - m)^+$. Then if $\lambda_i/\mu > m$,

$$\begin{aligned} \tilde{\Pi}_\lambda^i(m, n_i(m)) &= c_{fix} m + c_{flex}\eta^*(\lambda_i/\mu - m) + a\eta^*(\lambda_i/\mu - m)\beta \int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u) du \\ &= \left(c_{fix} - c_{flex}\eta^* - a\eta^*\beta \int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u) du \right) m \\ &\quad + \frac{\lambda_i}{\mu} \left(c_{flex}\eta^* + a\eta^*\beta \int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u) du \right). \end{aligned}$$

Now define $\kappa(m)$ as the first period we start blending when the fixed pool is set at m. Then, if $\lambda_k/\mu \le m$, we write $\kappa(m) \equiv k+1$ (recall that we assumed, without loss of generality, that the periods are ordered in increasing λ_i values). Otherwise, set $\kappa(m) \equiv \min\{i \ge 1 : \lambda_i/\mu > m\}$. Define $\sum_{i=k+1}^{k} z_i \equiv 0$. Then our goal is to solve

$$\begin{aligned} \min_m \sum_{i=1}^{k} T_i \tilde{\Pi}_\lambda^i(m, n_i(m)) := & \left(\sum_{i=1}^{k} T_i c_{fix} - \sum_{i=\kappa(m)}^{k} T_i \left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u) du \right) \right) \cdot m \\ & + \sum_{i=\kappa(m)}^{k} T_i \frac{\lambda_i}{\mu} \left(c_{flex}\gamma^* + a\gamma^*\beta \int_{-1}^{(1-\gamma^*)/(a\gamma^*)} F_\epsilon(u) du \right). \end{aligned} \tag{28}$$

Following the same line of argument as the proof of Lemma 2, one can show that the solution of (28) is $m = \lambda_{k_1}/\mu$, where $k_1 = 0$, if $c_{fix} > c_{flex}\eta^* + a\eta^*\beta\int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u)du$; and

$$k_1 = \max\left\{1 \le h \le k : \sum_{i=1}^{k} T_i c_{fix} \le \sum_{i=h}^{k} T_i \left(c_{flex}\eta^* + a\eta^*\beta\int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u)du\right)\right\},$$

otherwise.

Note that if $k_1 = 0$, the we use flexible servers only. If $k_1 = k$, then we use fixed servers only. If $1 \le k_1 < k$, then $k_1 + 1$ is the first period where we start blending.

We next take a closer look at the condition that determines k_1. $k_1 = h$ if

$$\begin{aligned}
&\sum_{i=1}^{k} T_i c_{fix} \le \sum_{i=h}^{k} T_i \left(c_{flex}\eta^* + a\eta^*\beta\int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u)du\right)\\
\iff &\frac{\sum_{i=1}^{k} T_i}{\sum_{i=h}^{k} T_i} c_{fix} \le c_{flex}\eta^* + a\eta^*\beta\int_{-1}^{(1-\eta^*)/(a\eta^*)} F_\epsilon(u)du\\
\iff &\frac{\sum_{i=1}^{k} T_i}{\sum_{i=h}^{k} T_i} c_{fix} \le \beta F_\epsilon\left(\frac{1-\eta^*}{a\eta^*}\right) \quad \text{as } g(\gamma^*) = 0.\\
\iff &\gamma^* \le \frac{1}{1 + aF_\epsilon^{-1}\left(c_{fix}^h/\beta\right)} \text{ where } c_{fix}^h = c_{fix}\frac{\sum_{i=1}^{k} T_i}{\sum_{i=h}^{k} T_i}.
\end{aligned} \tag{29}$$

Here $\iff$ means equivalent to. Now as $g'(\eta)$ is an increasing function of η and η^* is the solution of $g'(\eta) = 0$, to check if the inequality (29) holds, we can check whether

$$g\left(1/\left(1 + aF_\epsilon^{-1}\left(c_{fix}^h/\beta\right)\right)\right) \ge 0. \tag{30}$$

Furthermore, as

$$g\left(\frac{1}{1 + aF_\epsilon^{-1}\left(c_{fix}^h/\beta\right)}\right) = c_{flex} + \beta a\int_{-1}^{F_\epsilon^{-1}\left(c_{fix}^h/\beta\right)} F_\epsilon(u)du - c_{fix}^h\left(1 + aF_\epsilon^{-1}\left(c_{fix}^h/\beta\right)\right),$$

the inequality (30) is equivalent to

$$c_{flex} \ge c_{fix}^h + c_{fix}^h aF_\epsilon^{-1}\left(c_{fix}^h/\beta\right) - \beta a\int_{-1}^{F_\epsilon^{-1}\left(c_{fix}^h/\beta\right)} F_\epsilon(u)du.$$

References

Ata, B., D. Lee, E. Sonmez. 2018. Dynamic staffing of volunteer gleaning operations. University of Chicago, working paper.

Atar, R. 2008. Central limit theorem for a many-server queue with random service rates. *The Annals of Applied Probability* **18**(4) 1548–1568.

Azriel, David, Paul D Feigin, Avishai Mandelbaum. 2019. Erlang s: A data-based model of servers in queueing networks Technion, working paper.

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

Bassamboo, A., R. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research* **58**(5) 1398–1413.

Bassamboo, A., R. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Bolandifar, E., N DeHoratius, T. L. Olsen, J. Wiler. 2019. Modeling the behavior of patients who leave the ed without being seen The Chinese University of Hong Kong , working paper.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Operations research* **52**(1) 17–34.

Brown, Lawrence, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, Linda Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association* **100**(469) 36–50.

Cachon, Gerard P, Kaitlin M Daniels, Ruben Lobel. 2017. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management* .

Chen, M Keith, Judith A Chevalier, Peter E Rossi, Emily Oehlsen. 2017. The value of flexible work: Evidence from uber drivers. Tech. rep., National Bureau of Economic Research.

Drezner, Zvi, Nicholas Farnum. 1993. A generalized binomial distribution. *Communications in Statistics-Theory and Methods* **22**(11) 3051–3063.

Forbes. 2015. 3 secrets to leading a multi-everything blended workforce. `http://www.forbes.com/sites/meghanbiro/2015/11/07/3-secrets-to-leading-a-multi-everything-blended-workforce/#7cfdd938311a`. Accessed: 2017-02-08.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 208–227.

Green, L., S. Savin, N. Savva. 2013. "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10) 2237–2256.

Gurvich, I., M. Lariviere, T. Moreno-Garcia. 2018. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Northwestern University, working paper.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* **29**(3) 567–588.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management* **7**(1) 20–36.

Heyde, CC. 2004. Asymptotics and criticality for a correlated bernoulli process. *Australian & New Zealand Journal of Statistics* **46**(1) 53–57.

Hu, Ming, Yun Zhou. 2018. Price, wage and fixed commission in on-demand matching. University of Toronto, working paper.

Ibrahim, R. 2018. Staffing a service system with a random service capacity and impatient customers. *Production and Operations Management* .

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8) 1018–1038.

Mandelbaum, A., A.S. Zeltyn. 2007. Service engineering in action: the Palm/Erlang-A queue, with appications to call centers. *Advances in Services Innovations*. Springer-Verlag.

Ozkan, E., A. Ward. 2018. Dynamic matching for real-time ridesharing. University of Southern California, working paper.

Palm, C. 1953. Methods of judging the annoyance caused by congestion. *Tele* **4** 189–208.

PWC. 2017. The sharing economy grows up. `https://www.pwc.co.uk/issues/megatrends/collisions/sharingeconomy/outlook-for-the-sharing-economy-in-the-uk-2016.html`. Accessed: 2017-12-4.

Romano, Joseph P, Michael Wolf. 2000. A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & probability letters* **47**(2) 115–124.

Taylor, T. 2018. On-demand service platforms. University of California Berkeley, working paper.

Wang, W., D. Gupta. 2014. Nurse absenteeism and staffing strategies for hospital inpatient units. *Manufacturing and Service Operations Management* **16**(3) 439–454.

Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Operations Research* **54** 37–54.

Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15** 88–102.

Whitt, Ward. 2007. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)* **54**(5) 476–484.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems: Theory and Applications* **51**(3-4) 361–402.