Flexible Workers or Full-Time Employees? On Staffing Systems with a Blended Workforce

Jing Dong
Northwestern University, 2145 Sheridan Road, M239, Evanston, IL 60208 jing.dong@northwestern.edu

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

The rise of the blended workforce, which is identified as one of the top workplace trends in 2017, is prompting firms to re-evaluate their staffing strategies. A blended workforce melds, as a deliberate business strategy, contingent workers, e.g., independent contractors or freelancers, with permanent employees. In this paper, we study optimal staffing decisions in service systems with a blended workforce, in the context of a queueing-theoretic framework. Because part of the workforce is flexible, the number of servers in our queueing model is random. Since the staffing problem with a random number of servers is analytically intractable, we formulate two problem relaxations, demonstrate their accuracies in large systems by relying on an asymptotic, many-server, mode of analysis, and make staffing recommendations for systems with a blended workforce. We demonstrate that staffing decisions in such systems are not straightforward. Indeed, we characterize how these decisions depend on three main factors: (i) the supply-side uncertainty of the flexible agent pool, (ii) operating costs in the system, and (iii) fluctuations in incoming customer demand, and show that it may or may not be cost-effective to staff a blended workforce, depending on the interplay between those factors.

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

1. Introduction

Major multinational companies (McKinsey&Company 2015, Deloitte 2016, Accenture 2016) and leading business periodicals (The Economist 2013, Harvard Business Review 2016, Forbes 2016) identify the rise of the *blended workforce* as one of the top global workplace trends which are reshaping the modern business landscape. A blended workforce melds, as a deliberate business strategy, a layer of contingent workers, e.g., independent contractors or freelancers, with a core of permanent, full-time, employees. In order to effectively manage a blended workforce, companies must begin by "reevaluating their staffing models" (Forbes 2015). Indeed, a major problem facing those companies is how to decide on the "right number of right people at the right time"[1], by appropriately weighing the pertinent tradeoffs. This is the problem that we address in this paper.

The rise of the blended workforce. According to an Intuit (2016) report, the number of contingent workers will constitute more than 40% of the American workforce by 2020. While the proportions

[1] https://www.beeline.com/blog/10-tips-for-the-right-blend-of-flexible-workers-in-your-organization/

of such workers are on the rise in all sectors of the economy, this is especially the case for the service sector (UKCES 2016), which we focus on in this paper. Indeed, multiple Fortune 500 companies in the service sector currently rely on a blended workforce model, e.g., Walmart (*walmart.com*), Time Warner (*timewarnercable.com*), and Netflix (*netflix.com*), to name a few. Similarly, several small to medium-size companies rely on such a model as well, e.g., R & M (*renmmatrix.nl*) is a market research bureau which has 40 full-time employees in addition to 300 freelance call-center agents who enjoy a great degree of flexibility in setting their own schedules (Feinberg et al. 2005).

Managerial challenges with a blended workforce. Several tradeoffs need to be weighted when managing a blended workforce [2]. On one hand, fixed, full-time, workers are typically reliable and committed to the firm, and they usually have a number of required working hours. In other words, they are relatively easy to control. However, this control comes at the expense of a fixed and relatively steep labor cost. Moreover, a pool of regular full-time workers cannot be easily scaled to meet dynamic business needs, i.e., periods of high or low customer demand.

On the other hand, flexible, contingent, workers are recruited on a part-time basis. To be concrete, we are thinking here of contingent workers who are recruited to complete relatively low-skill tasks, such as those advertised on Wonolo (*wonolo.com*) or Zaarly (*zaarly.com*) which offer business-to-business services by matching qualified workers to client companies. Flexible workers have a legal right to various degrees of flexibility, e.g., in setting their own work schedules, at the expense of being deprived of benefits which are usually granted to regular employees, e.g., unemployment insurance, overtime compensation, etc. Contingent workers are also typically less reliable than regular employees: They tend to have high turnover rates and uncertain availabilities, mostly due to the flexibility that is inherent in their work contracts. However, a pool of flexible workers can be easily scaled to meet seasonal demand fluctuations.

A service-system manager who relies on a blended workforce must therefore decide, as a long-term business strategy in an initial planning stage, on the numbers of flexible and fixed agents to staff so as to effectively balance operating costs, varying customer demand patterns, and supply-side uncertainty, while not compromising on the quality of service offered to customers.

Modelling framework and overview. In this paper, we study the problem of staffing a service system with a blended workforce in the context of a stylized queueing model. We assume that there are k working periods, customer demand rates are time-varying [3], and the agent pool comprises both fixed and flexible agents. (Hereafter, we use "agent" and "server" interchangeably.) A fixed server is available in each period, and is compensated c_0 per unit time. To capture the supply-side

[2] https://www.xero.com/uk/small-business-guides/business-management/independent-contractor-or-employee

[3] For time-varying demand, our analysis is based on a two-period setting with high or low demand.

uncertainty associated with flexible agents, we assume that a flexible server may or may not be available for work in any given period. If a flexible server is available, she earns c_1 per unit time. That is, c_0 and c_1 are the staffing costs in the system. Because part of the agent pool is flexible, the total number of available servers in our queueing model is random.

In an initial planning stage, the system manager must decide on the number of fixed servers, m, and the *expected* number of flexible servers, n. Indeed, since flexible servers may not be available, the manager cannot enforce a *realized* number of flexible servers, and must plan on an expected number instead (the distribution of the random number of flexible agents who actually show up, which depends on that expected value, is assumed to be known to the manager). In selecting the respective pool sizes of fixed and flexible servers, the decision maker is effectively controlling the supply-side uncertainty in the system, i.e., the *distribution* of the number of available servers. As in Gurvich et al. (2017), we also allow for the system manager to impose a cap on the expected number of flexible agents that she allows in a period. That is, we allow the system manager to impose on a flexible server not to show up in a given period, even when the server is willing to do so. The imposition of a cap is an additional control lever available to the system manager, which allows her to dynamically scale the pool size of flexible agents to meet fluctuations (time variations) in customer demand. Customers are assumed to be both impatient and delay sensitive, as is usually the case in service systems (Garnett et al. 2002).

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, many-server, mode of analysis. In particular, our modelling approach is close to the one in Bassamboo et al. (2010), who consider a single-period capacity-sizing problem with random arrival rates instead.

Here is a brief summary of our modelling approach. 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. The first formulation assumes that uncertainty effects dominate stochastic fluctuations. In this regime, we derive the optimal staffing levels by solving a *stochastic-fluid* optimization problem which ignores stochastic variability; this problem is of a multi-period newsvendor type in our setting. The second formulation assumes that both uncertainty effects and stochastic fluctuations are negligible. In this regime, we derive the optimal staffing levels by solving a *fluid* optimization problem instead. Our objective is to understand the properties of the alternative solutions which are obtained under those two approximations, to rigorously justify their

accuracy by quantifying their corresponding errors (asymptotically in large systems) and, most importantly, to characterize the relevant tradeoffs so as to draw insights into the effective staffing of service systems with a blended workforce.

A key technical challenge in our analysis is that the ensuing randomness in the number of servers, i.e., the uncertain parameter in our queueing system, depends itself on the staffing levels, i.e., our decision variables. For example, the resulting multi-period newsvendor problem in the stochastic-fluid relaxation of our problem can be equivalently formulated as one where demand depends on the stocking quantity and, to the best of our knowledge, there do not exist simple closed-form solutions to this type of problems [4]. More generally, this also implies that the asymptotic accuracy of our respective relaxations may depend on the specific solutions to those problems. To circumvent this difficulty, we derive approximate, asymptotically optimal (in a sense to be made more precise later), solutions to the stochastic-fluid problems as well, and quantify their corresponding accuracies. Essentially, those approximate solutions are based on determining appropriate additional capacity refinements (safety hedges) to fluid-approximation counterparts. Those refinements depend, in turn, on the variability, if any, in the staffed pool of servers.

Specific contributions of our paper. The main contribution of this paper is two-fold: (1) We formulate several recommendations on staffing systems with a blended workforce and draw corresponding managerial insights; and (2) we derive technical results pertaining to the analysis of queueing systems with randomness in capacity.

1. Here is a summary of our staffing recommendations and main related insights:
    - When customer demand rates do not vary, we find that a manager should *not* use a blended workforce; instead, she should, in general, rely solely on the cheaper alternative, fixed or flexible. This result suggests that calls to "all human resources leaders to take action now and plan for a blended workforce" (Forbes 2017) may not always be appropriate, particularly when managers do not face strong fluctuations in customer demand.
    - If the flexible resource entails high variability, specifically when the coefficient of variation is bounded away from 0 in large systems, then it may be cost-effective to staff a *more expensive* fixed resource instead, i.e., there is a price to be paid for that variability. This is insightful because one of the main reasons why independent contractors are preferred over regular employees, in practice, is that they are the cheaper alternative (Brumm 2017); we show that even if independent contractors were indeed cheaper, then it may still be, overall, more cost effective to staff a workforce consisting *solely* of employees.

[4] This challenge does not arise when considering random arrival rates as in Bassamboo et al. (2010): There, the stochastic-fluid optimal solution has a remarkably simple critical fractile form.

- Variation in demand patterns makes using the flexible resource *more attractive* over the fixed resource, because it can be easily scaled to dynamically meet variations in customer demand: A manager would always use a flexible resource (alone or through blending) so long as it is not "much more" expensive than the fixed resource. In particular, our result provides support to current business practices where independent contractors are staffed to meet fluctuations in customer demand (New York Times 2012).
- When customer demand rates do vary, it may be cost effective for the manager to use a blended workforce. However, this is only the case if the fixed resource is *cheaper* and the *disparity in pay* between the fixed and flexible resources is not too great. If the fixed resource is more expensive, and the flexible resource is not too variable, then the manager should rely solely on the flexible resource. From a practical standpoint, this result provides insight into the appropriateness of alternative business models which are adopted in different on-demand service platforms. To illustrate, we consider the example of ride-sharing services, which are typically confronted with a demand that varies over time. Ride-sharing services, such as Uber (*uber.com*) or Lyft (*lyft.com*) rely solely on independent contractors to field ride requests from their customers. Assuming that the supply of independent contractors is not too variable, i.e., the coefficient of variation is close to 0 in large systems, we find that this model of using only independent contractors is appropriate if, e.g., independent contractors are cheaper, but that incorporating permanent drivers into the workforce (blending) is more appropriate if e.g., permanent drivers are cheaper (but not by much). If permanent drivers are much cheaper, then the manager should rely strictly on permanent drivers instead.

2. Here is a summary of our technical results:
   - We quantify the asymptotic order of magnitude of errors for the fluid and stochastic-fluid approximations in a multi-period setting, both with and without time-varying demand.
   - We derive exact (in the form of an implicit relation) and approximate solutions of the stochastic-fluid formulation, which is difficult to solve in our setting, and quantify the accuracy of those approximate solutions.
   - We show that with a random number of servers, as with random arrival rates (Bassamboo et al. 2010), one can distinguish between an *uncertainty-dominated* regime and a *variability-dominated* regime, depending on the magnitude of uncertainty in capacity. In particular, the stochastic-fluid optimal solution involves a base capacity and an uncertainty hedge which is "extremely" accurate in the *uncertainty-dominated* regime, but there is no concrete benefit from that uncertainty hedge over the regular square-root staffing hedge (Halfin and Whitt 1981, Garnett et al. 2002) in the *variability-dominated* regime.

- We supplement our analytical results with detailed numerical studies throughout. As robustness checks, we consider both a random arrival rate (along with a random capacity) and a general patience distribution. With a random arrival rate, we find that the additional safety capacity that should be used depends on which of the two types of uncertainty, in arrivals or in capacity, dominates asymptotically (i.e., is of a higher order of magnitude). With a general patience distribution, we find that investigating the asymptotic accuracy of the respective relaxations, i.e., fluid and stochastic-fluid, boils down to determining the asymptotically optimal regime which is prescribed *at fluid scale* (depending on properties of the failure-rate function of the patience distribution).

The rest of this paper is organized as follows. In §2, we review the relevant literature. In §3, to build intuition, we formulate initial insights into the impact of randomness in capacity by considering a special case. In §4, we formulate our capacity-sizing problem, as well as its stochastic-fluid and fluid relaxations; we also quantify the corresponding optimality gaps. In §5, we derive the exact and asymptotically optimal solutions of the stochastic-fluid relaxation with stationary demand, and draw insights into staffing systems with a blended workforce. In §6, we extend our results to the two-period case with time-varying demand rates. In §8, we consider general abandonment, and in §9, we draw conclusions. We relegate all proofs to the e-companion.

2. Related Literature

Our modelling approach is close to the stream of literature initiated by Harrison and Zeevi (2005) and used in Bassamboo et al. (2010) who address the question of capacity planning under parameter uncertainty. Our paper is related to the extensive literature analyzing asymptotics of many-server queueing systems with impatient customers (Garnett et al. 2002, Zeltyn and Mandelbaum 2005, Whitt 2004, 2006a, Bassamboo and Randhawa 2010, Bassamboo et al. 2010), and to the large literature on optimal staffing decisions in service systems (Maglaras and Zeevi 2003, Borst et al. 2004, Harrison and Zeevi 2005, Bassamboo et al. 2005, 2010); for other references, see Gans et al. (2003) and Akşin et al. (2007). However, none of those papers considers a random number of servers. Whitt (2006b) considers many-server queues with an uncertain arrival rate, an uncertain number of servers, and a single period. Our focus here is on staffing multiple periods instead, and we go beyond the fluid approximation in that paper. Atar (2008) derives a diffusion limit for the number of customers with a random number of servers and random service rates. However, the staffing question is not addressed there.

There is a body of research within the queueing games literature which considers strategic servers that may select their service rates (Cachon and Harker 2002, Cachon and Zhang 2007). However, such papers do not consider staffing decisions, and the maximum number of servers considered is

two. Recent exceptions are Gopalakrishnan et al. (2016) and Zhan and Ward (2017). Our work is related to papers on nurse staffing with absenteeism, such as Green et al. (2013) and Wang and Gupta (2014), however our asymptotic mode of analysis is different, as well as our consideration of a blended workforce.

This paper is most closely related to recent papers on queues with a self-scheduling capacity. Gurvich et al. (2017) 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 (2017) 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 systems with self-scheduling agents. However, both Gurvich et al. (2017) and Ibrahim (2017) rely solely on fluid approximations to the system, and do not consider a blended workforce.

Taylor (2017) examines how two defining features of an on-demand service platform, delay sensitivity and agent independence, impact the platform's optimal per-service price and wage. Ozkan and Ward (2017) study optimal matching decisions in a ride-sharing platform and demonstrate the need to go beyond the prevailing closed-driver matching policy. Braverman et al. (2017) model a ride-sharing system as a closed queueing network and rely on a fluid model to derive an optimal routing policy. Cachon et al. (2017) study optimal contracts in a platform with self-scheduling capacity; the agent pool size is large and assumed to be given a priori, i.e., the staffing question is not addressed. Riquelme et al. (2017) model a ride-sharing service and determine optimal platform pricing. They find that threshold-based dynamic pricing does not outperform a static pricing policy, but that dynamic pricing is more robust to fluctuations in system parameters.

3. A Random Number of Servers

In this section, we specify our queueing framework. To build intuition, we also formulate initial insights into the impact of randomness in capacity: In a special case, we demonstrate (Lemma 1) that randomness in capacity leads to a deterioration in the system's performance. We then describe simulation results which illustrate that similar insights hold more generally as well.

3.1. Queueing Model

We consider a single-class $M/M/N+M$ queueing model [5]. Customers arrive to the system according to a Poisson process with rate λ, and service times are independent and identically distributed (i.i.d.) exponential random variables with rate μ. Customers are impatient, and their patience times

[5] For now, we do not distinguish between fixed and flexible capacity; we will do so later when investigating optimal staffing decisions in the system.

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 number of servers, N, is a nonnegative integer random variable. 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). Thus, a proper steady-state exists, and we focus on steady-state performances throughout.

3.2. Impact of Randomness in Capacity

Let $Q(N)$ denote the steady-state number of customers in a system with N servers. For simplicity, we assume here that the abandonment rate, θ, is equal to the service rate, μ. Let $\sigma_1 > 0$, and $\sigma_2 > 0$, and consider two queueing systems with N_1 and N_2 servers, respectively, given by:

$$N_1 = n + \sigma_1 \epsilon \quad \text{and} \quad N_2 = n + \sigma_2 \epsilon,$$

for a proper random variable $\epsilon \geq \max\{-n/\sigma_1, -n/\sigma_2\}$ with $\mathbb{E}[\epsilon] = 0$. That is, $\mathbb{E}[N_1] = \mathbb{E}[N_2] = n$, $\text{Var}[N_1] = \sigma_1^2 \text{Var}[\epsilon]$, and $\text{Var}[N_2] = \sigma_2^2 \text{Var}[\epsilon]$. All remaining parameters are assumed to be identical across the two systems. Then, the following lemma holds.

LEMMA 1. *If $\sigma_1 \leq \sigma_2$, then $\mathbb{E}[Q(N_1)] \leq \mathbb{E}[Q(N_2)]$.*

Lemma 1 shows that, with all else held constant (including the expected number of servers), increased variability in the number of servers leads to *worse* system performance. In other words, a decision maker who ignores that variability, and approximates her available capacity by its expected value instead, would have an overoptimistic view of performance in her system.

3.3. Numerical Study

We now describe supporting results from a short simulation study (Table 1). Our objective here is two-fold: (i) to quantify the deterioration in system performance which results from an increased variability in capacity, and (ii) to show that the preliminary results of Lemma 1 hold under more general distributional assumptions as well.

We let the number of servers N_λ, in a system with arrival rate λ, have a truncated normal distribution; specifically, we let $N_\lambda = \lceil Z_\lambda^+ \rceil$ where $Z_\lambda \sim \text{Nor}(\lambda/\mu, \sigma_\lambda^2)$. We note that $\mathbb{E}[Z_\lambda] = \lambda/\mu$ represents a "base capacity" to match mean demand. We consider alternative values of the arrival rate, $\lambda = 50, 100, 500$, and 1000, and hold the service rate $\mu = 1$ fixed. We also consider alternative functional forms for the variance, $\sigma_\lambda = \sqrt{\lambda}$, $0.5\lambda^{3/4}$, and 0.25λ. As a benchmark, we consider $\sigma_\lambda = 0$, which represents the classical case where the number of servers is deterministic.

For patience times, we consider both the exponential ($\theta = 2.0$ and $\theta = 0.5$) and H_2 (hyperexponential with squared coefficient of variation equal to 4, balanced means, and mean equal to

2) distributions. For service times, we consider both the exponential ($\mu = 1$) and LN (lognormal with mean and variance both equal to 1) distributions. We consider H_2 for abandonment and LN for service times because there is empirical evidence suggesting good fits to those distributions in practice (Roubos and Jouini 2013, Brown et al. 2005). Our simulations estimates are based on 100 independent replications of length one million arrivals each, and we discard an initial transient period of length 20,000 arrivals from each replication.

$M/M/N + M$ with $\theta = 2.0$

	$\sigma_\lambda = 0$		$\sigma_\lambda = \sqrt{\lambda}$		$\sigma_\lambda = 0.5\lambda^{3/4}$		$\sigma_\lambda = 0.25\lambda$	
λ	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$
50	1.67	8.27	2.16	12.3	2.45	15.2	2.87	20.1
100	2.29	15.7	3.02	24.1	3.73	34.7	5.11	60.8
500	5.24	78.5	6.64	117	10.8	276	22.9	1137
1000	7.37	155	9.30	231	17.4	704	45.1	4367

$M/M/N + H_2$

	$\sigma_\lambda = 0$		$\sigma_\lambda = \sqrt{\lambda}$		$\sigma_\lambda = 0.5\lambda^{3/4}$		$\sigma_\lambda = 0.25\lambda$	
λ	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$
50	3.32	23.1	4.88	46.4	5.67	63.6	6.88	95.6
100	4.65	45.0	6.74	90.4	8.74	154	12.5	325
500	10.5	222	14.6	432	26.0	14.0×10^2	57.5	70.4×10^2
1000	14.9	442	20.3	840	42.5	37.6×10^2	114	278×10^2

$M/LN(1,1)/N + M$ with $\theta = 0.5$

	$\sigma_\lambda = 0$		$\sigma_\lambda = \sqrt{\lambda}$		$\sigma_\lambda = 0.5\lambda^{3/4}$		$\sigma_\lambda = 0.25\lambda$	
λ	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$	$\mathbb{E}[Q(N_\lambda)]$	$\text{Var}[Q(N_\lambda)]$
50	4.60	36.6	7.20	90.5	8.50	132	10.5	208
100	6.52	71.8	9.94	177	13.3	335	19.3	743
500	14.1	334	21.7	853	40.5	32.9×10^2	90.0	169×10^2
1000	20.0	671	30.1	168×10^2	66.3	89.8×10^2	178	671×10^2

Table 1 **Expected value and variance of the steady-state queue length in the $M/GI/N + GI$ queueing system with $N = Z^+$ where $Z \sim \text{Nor}(\lambda/\mu, \sigma_\lambda^2)$.**

Inspecting the rows of Table 1 shows that as σ_λ increases for a fixed λ, both $\mathbb{E}[Q(N_\lambda)]$ and $\text{Var}[Q(N_\lambda)]$ increase as well. This substantiates the preliminary results of Lemma 1. Inspecting the columns of Table 1 (for each model) allows for a more detailed understanding of the system's performance. As expected, when the number of servers is constant, we observe that as λ increases by a factor $l > 0$, $\mathbb{E}[Q(N_\lambda)]$ increases by a factor of $\sqrt{l}$ and $\text{Var}[Q(N_\lambda)]$ increases by that same factor l (Bassamboo et al. 2010). However, for $\sigma_\lambda > 0$, as λ increases by a factor $l > 0$, $\mathbb{E}[Q(N_\lambda)]$

increases by a factor of σ_l instead, and $\mathrm{Var}[Q(N_\lambda)]$ continues to increase by that same factor l. In other words, Table 1 suggests that, as λ increases, $\mathbb{E}[Q(N_\lambda)]$ is on the order of magnitude of σ_λ (when $\sigma_\lambda \geq \sqrt{\lambda}$), and $\mathrm{Var}[Q(N_\lambda)]$ is on the order of magnitude of the maximum between λ and σ_λ^2. In §4.3, we will establish such results (for the expectation) formally by considering an asymptotic many-server queueing framework where we let λ increase without bound.

Table 1 illustrates that the order of magnitude of "noise" in a system with a random number of servers may be higher than the usual square-root order of stochastic fluctuations. Thus, when investigating appropriate staffing decisions with a blended workforce, there is a need to derive an appropriate hedge against such variability, which may be of a different order than the standard square-root safety capacity (Garnett et al. 2002). We describe our staffing problem in the following section; the solution to this problem allows for the specification of an appropriate hedge.

4. Capacity Sizing with a Blended Workforce

In this section, we formulate the capacity-sizing problem faced by the system manager. 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 extend our modelling framework in §7; there, we consider a doubly-stochastic Poisson arrival process with a random arrival rate.) We fix $\lambda > 0$ and let $\lambda_i \equiv \lambda\xi_i$, where $\xi_i \geq 0$ for each i. In formulating the capacity-sizing problem, we index all relevant quantities by λ, to indicate dependence on the arrival rates. In our asymptotic analysis, we let λ grow without bound while keeping each ξ_i constant.

The system manager must determine the staffing levels for the fixed and flexible agent pools so as to strike a balance between customer-related and staffing costs in the system. Specifically, consistently with Bassamboo and Randhawa (2010) and Bassamboo et al. (2010), we consider two quality-of-service 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. The assumption of time-homogeneous customer costs is consistent with having a single class of customers across periods. The (per unit of time) staffing costs are given by c_0 for a fixed server, and c_1 for a flexible server. In practice, it is usually the case that $c_1 < c_0$, but we do not impose this assumption here. Instead, we study how the solution to our capacity-sizing problem depends on the individual staffing costs, c_0 and c_1.

We let m_λ denote the number of fixed servers, and n_λ denote the *expected* number of flexible servers. Since our choices of m_λ and n_λ affect the distribution of the random number of servers,

$N(m_\lambda, n_\lambda)$, we need to make further specifications. In particular, we assume that the total number of servers can be expressed as:

$$N(m_\lambda, n_\lambda) = m_\lambda + n_\lambda + \sigma_{n_\lambda}\epsilon, \tag{1}$$

for some random variable $-1 \leq \epsilon \leq 1$ with $\mathbb{E}[\epsilon] = 0$. We assume that ϵ 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. We assume that $\sigma_{n_\lambda} \geq 0$ is some function of n_λ. Later, we will demonstrate that the optimal staffing level in our problem depends on how σ_{n_λ} compares to the magnitude of stochastic variability in the system. The expression in (1) implies that $\mathbb{E}[N(m_\lambda, n_\lambda)] = m_\lambda + n_\lambda$ and $\text{Var}[N(m_\lambda, n_\lambda)] = \sigma^2_{n_\lambda}\text{Var}[\epsilon]$. We also make the following assumptions.

ASSUMPTION 1. *For σ_n in (1), we assume that $\sigma'_n > 0$ and $\sigma''_n < 0$; also, $c_1, c_0 < (h/\theta + r)\mu$.*

Assumption 1 guarantees that σ_n is strictly increasing and concave (which will be used later to ensure uniqueness of solutions). The assumption on costs ensures that the fixed and flexible resources are cheap enough to avoid pathological cases where the system manager would not staff any of the two resources. In (1), we ignore the integrality assumptions on m_λ, n_λ, and $N(m_\lambda, n_\lambda)$: 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 the all-Markovian Erlang-A queueing model can be extended to real values of the number of servers (Mandelbaum and Zeltyn 2007). The staffing problem faced by the manager is defined for both integer and non-integer values of m_λ and n_λ.

As in Gurvich et al. (2017), we also allow for the system manager to impose a cap on the expected number of flexible agents in a period. That is, we allow the system manager to impose on a flexible server not to show up in a given period, even when the server is willing to do so. In particular, we let α^i_λ denote the cap in period i, and $\alpha^i_\lambda n_\lambda$ be the expected number of flexible servers that will be allowed in period i. We let $Q^i(m_\lambda, \alpha^i_\lambda n_\lambda)$ and $\Xi^i(m_\lambda, \alpha^i_\lambda n_\lambda)$ denote the steady-state queue-length and steady-state rate of customer abandonment in period i. We let $X^i(m_\lambda, \alpha^i_\lambda n_\lambda)$ denote the steady-state number of customers in the system, in period i, so that:

$$Q^i(m_\lambda, \alpha^i_\lambda n_\lambda) = (X^i(m_\lambda, \alpha^i_\lambda n_\lambda) - N(m_\lambda, \alpha^i_\lambda n_\lambda))^+,$$

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

$$\Xi^i(m_\lambda, \alpha^i_\lambda n_\lambda) = \theta \cdot \mathbb{E}[Q^i(m_\lambda, \alpha^i_\lambda n_\lambda)],$$

where θ is the rate of the patience-time distribution (Mandelbaum and Zeltyn 2007). The system manager's long-run staffing problem is given by:

$$\min_{m_\lambda, n_\lambda, \alpha^i_\lambda} \quad \Pi_\lambda(m_\lambda, n_\lambda, \alpha^i_\lambda) \tag{2}$$

$$
\begin{aligned}
&\equiv \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha_\lambda^i n_\lambda + h \cdot \mathbb{E}[Q^i(m_\lambda, \alpha_\lambda^i n_\lambda)] + r \cdot \Xi(m_\lambda, \alpha_\lambda^i n_\lambda) \right), \\
&= \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha_\lambda^i n_\lambda + (h + r\theta) \mathbb{E}[Q^i(m_\lambda, \alpha_\lambda^i n_\lambda)] \right), \\
&= \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha_\lambda^i n_\lambda + (h + r\theta) \mathbb{E}[(X^i(m_\lambda, \alpha_\lambda^i n_\lambda) - N(m_\lambda, \alpha_\lambda^i n_\lambda))^+] \right), \\
&= \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha_\lambda^i n_\lambda + (h + r\theta) \mathbb{E}[(X^i(m_\lambda, \alpha_\lambda^i n_\lambda) - m_\lambda - \alpha_\lambda^i n_\lambda - \sigma_{\alpha_\lambda^i n_\lambda} \cdot \epsilon)^+] \right).
\end{aligned}
$$

We note that the problem formulation in (2) assumes that m_λ is fixed across all periods. In some settings, it may be possible to decide on a different level for the fixed resource in each period. In this case, the manager can decide on both her fixed and flexible staffing levels for each period separately. In other words, the optimal staffing policy reduces to the one with a single period instead or, equivalently, with stationary demand, which we treat in §5.

The problem formulation in (2) is prohibitively difficult to solve in closed form, because our choices of m_λ and n_λ affect the distribution of the number of servers, which, in turn, affects the distributions of both the queue-length, Q, and the rate of abandonment, Ξ. Thus, we turn next to formulating stochastic-fluid (ignoring stochastic variability) and fluid (ignoring both stochastic variability and parameter uncertainty) relaxations of that problem.

4.1. Stochastic-Fluid Problem

For the stochastic-fluid relaxation of our problem, we ignore stochastic fluctuations in the system. In particular, customers arrive to period i at the rate of λ_i per unit of time. The processing capacity is $N(m_\lambda, \alpha_\lambda^i n_\lambda)\mu$ and, by conservation of flow, the resulting stochastic-fluid abandonment rate is given by $(\lambda_i - N(m_\lambda, \alpha^i n_\lambda)\mu)^+$. Thus, the resulting stochastic-fluid approximation to (2) is:

$$
\begin{aligned}
\min_{m_\lambda, n_\lambda, \alpha_\lambda^i} \bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda^i) &\equiv \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha_\lambda^i n_\lambda + \left(\frac{h}{\theta} + r \right) \left(\lambda_i - N(m_\lambda, \alpha^i n_\lambda) \cdot \mu \right)^+ \right), \qquad (3) \\
&= \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha_\lambda^i n_\lambda + \left(\frac{h}{\theta} + r \right) \left(\lambda_i - m_\lambda \mu - \alpha_\lambda^i n_\lambda \mu - \sigma_{\alpha_\lambda^i n_\lambda} \mu \epsilon \right)^+ \right).
\end{aligned}
$$

The stochastic-fluid formulation in (3) may be viewed as a multi-period newsvendor-type problem, albeit one where the "demand" in period i, $\lambda_i - \sigma_{\alpha_\lambda^i n_\lambda} \mu\epsilon$, is dependent on the overprocessing capacity, $m_\lambda \mu + \alpha_\lambda^i n_\lambda \mu$, where m_λ and n_λ are the decision variables. There have been relatively few papers in the literature which study newsvendor-type problems where the distribution of the demand depends on the stocking quantity and these, to the best of our knowledge, are restricted to a single-period setting only; e.g., see Balakrishnan et al. (2008). Because it is difficult to derive a simple, closed-form, analytical solution to (3), we resort to deriving an asymptotically optimal solution instead, in a sense that will be made precise in §5.

4.2. Fluid Problem

We are now ready to formulate the fluid relaxation of our problem. For this, we ignore both uncertainty effects and stochastic fluctuations in the system. In particular, the fluid abandonment rate in our problem is given by $(\lambda_i - m_\lambda\mu - \alpha^i_\lambda n_\lambda\mu)^+$, which is obtained by substituting the random number of servers, $N(m_\lambda, \alpha^i_\lambda n_\lambda)$, by its expected value, $m_\lambda + \alpha^i_\lambda n_\lambda$. This leads to the following:

$$\min_{m_\lambda, n_\lambda, \alpha^i_\lambda} \tilde{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha^i_\lambda) \equiv \sum_{i=1}^{k} T_i \left(c_0 m_\lambda + c_1 \alpha^i_\lambda n_\lambda + \left(\frac{h}{\theta} + r\right) \left(\lambda_i - m_\lambda\mu - \alpha^i_\lambda n_\lambda\mu\right)^+ \right). \tag{4}$$

Comparing the optimal staffing policies given by (3) and (4) allows for a deeper understanding into the impact of randomness in capacity on the system's optimization. Given its simple form, the fluid approximation in (4) may be preferred, provided that it does not entail a significantly less accurate solution than (3). Next, we characterize when this is indeed the case.

4.3. Optimality Gaps

We now study the accuracy of the staffing prescriptions in (3) and (4), in a regime where the arrival rate, λ, is large. For ease of exposition, we focus here on a single-period setting or, equivalently, the stationary-demand case, and relegate the generalization to the multi-period case to the e-companion (§EC.3). Thus, we drop dependence on the period's index, i. With a single period, we can equivalently consider the capped expected number of flexible servers, $\alpha_\lambda n_\lambda$, to be our decision variable, i.e., we can eliminate α_λ from the problem.

We begin with a statement of our main theorem. Let $(m^*_\lambda, n^*_\lambda)$ denote the optimal staffing levels to the original staffing problem in (2). Let $(\bar{m}_\lambda, \bar{n}_\lambda)$ and $(\tilde{m}_\lambda, \tilde{n}_\lambda)$ denote the optimal solutions to its stochastic-fluid relaxation in (3) and its fluid relaxation in (4), respectively. To interpret the results of Theorem 1, and subsequent theorems, we need the following 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 $|f(n)| \le M|g(n)|$ for $n \ge C$;

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

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

We are now ready to state the main theorem of this section.

Theorem 1. *For large λ,*

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

that is, if $\tilde{n}_\lambda = \Theta(\lambda)$, then $\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\max\left\{\sigma_\lambda, \sqrt{\lambda}\right\}\right)$.
Also, for large λ,

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

moreover, if $\bar{n}_\lambda = \Theta(\lambda)$ *and* $\sigma_{n_\lambda} > \sqrt{n_\lambda}$*, then:*

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

The results of Theorem 1 are in line with those in Theorem 1 of Bassamboo et al. (2010), and indeed our proof proceeds in a similar fashion. In particular, Theorem 1 characterizes two operating modes, depending on the magnitude of variability in the random number of servers, as quantified by σ_λ. When σ_λ is "large", i.e., of an order which is larger than the square-root order of stochastic fluctuations in the system, the system can be considered to be in an *uncertainty-dominated* regime. In this regime, stochastic-fluid approximations are remarkably accurate. Indeed, the optimality gap for the stochastic-fluid solution is on the order $\mathcal{O}(\lambda/\sigma_\lambda)$. In other words, 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 square-root order of stochastic fluctuations in the system, the system can be considered to be in a *variability-dominated* regime. In this case, the optimality gap for the stochastic-fluid solution is on the order of stochastic fluctuations in the system, i.e., $\mathcal{O}(\sqrt{\lambda})$. Since the fluid approximation to our problem ignores both uncertainty and variability effects, the corresponding optimality gap is dominated by either the order of stochastic fluctuations in the system, or the order of variability in the number of servers, whichever is larger. In other words, when the variability in the number of servers is small, there is no distinct (not asymptotically negligible) advantage from using stochastic-fluid approximations over fluid approximations to the system. This implies that a first-order (fluid) approximation of the system's performance is sufficient to determine the optimal staffing policy. Since the results of Theorem 1 assume specific forms for the solutions to (3), there remains to show that these are indeed the correct forms for those solutions: We show this in §5.

4.4. Supporting Numerical Study

We close this section with a numerical investigation of the performance of our stochastic-fluid and fluid solutions. In particular, we show that the asymptotic results of Theorem 1 describe small to moderate systems well, which validates their usefulness even further. Since including fixed servers does not affect our asymptotic accuracy results, we restrict attention to having only a flexible pool. We assume the following values for the cost parameters: $c = 1/3$, $p = 1$, and $h = 1$. We also assume that $\mu = 1$ and $\theta = 3$. We let ϵ in (1) have a uniform distribution over $[-1, 1]$ and vary the functional form for σ_n: $\sigma_n = \sqrt{n}$, $n^{3/4}$, and $0.25n$. We report the optimal solutions of problems (3) and (4), $\bar{n}$ and $\tilde{n}$, and the corresponding relative and absolute errors.

We can make several observations based on our results in Table 2. First, as expected, fluid approximations are consistently less accurate than their stochastic-fluid counterparts. Second, stochastic-fluid approximations are extremely accurate, particularly when the variability in the number of

servers, i.e., σ_n, is large. Third, our asymptotic results are useful in describing small systems, e.g., consisting of tens of servers. In particular, for $\lambda = 20$, the errors in both the stochastic-fluid and fluid approximations are remarkably small. Fourth, the results in Table 2 substantiate the orders of magnitude reported in Theorem 1. For example, focusing on the fluid approximation and $\sigma_n = n^{3/4}$, we see that when λ is multiplied by a factor of $l = 50$ (in going from 20 to 1000, i.e., first to last row), the corresponding fluid errors increase roughly by a factor of 21, which is approximately equal to $l^{0.75} = 50^{0.75}$. In other words, our numerical results suggest that those errors are on the order of magnitude of σ_λ, as given by Theorem 1.

$\sigma_n = \sqrt{n}$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
20	24	22	20	0.173	0.700	1.86	7.20
50	57	53	50	0.333	1.13	1.60	5.43
100	110	105	100	0.360	1.64	0.921	4.20
150	162	156	150	0.474	2.04	0.83	3.58
500	522	511	500	0.894	3.82	0.498	2.13
1000	1031	1016	1000	1.18	5.40	0.337	1.54

$\sigma_n = n^{3/4}$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
20	25	24	20	0.010	0.556	0.101	5.38
50	59	58	50	0.015	1.12	0.0646	4.80
100	115	114	100	0.010	1.92	0.0231	4.39
150	171	169	150	0.0116	2.64	0.0182	4.15
500	551	550	500	6.19×10^{-4}	6.92	3.13×10^{-4}	3.50
1000	1086	1085	1000	2.50×10^{-5}	12.1	7.0×10^{-6}	3.14

$\sigma_n = 0.25n$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
20	24	22	20	0.143	0.595	1.51	6.30
50	57	55	50	0.0915	1.00	0.419	4.60
100	113	111	100	0.0211	1.61	0.0500	3.82
150	168	166	150	0.0212	2.23	0.0340	3.56
500	556	555	500	6.83×10^{-4}	6.70	3.35×10^{-4}	3.29
1000	1110	1109	1000	1.61×10^{-3}	13.2	3.96×10^{-4}	3.25

Table 2 Asymptotic accuracy of the stochastic-fluid and fluid solutions, $\bar{n}$ and $\tilde{n}$, to problems (3) and (4).

5. Capacity Sizing with Stationary Demand

The asymptotic results of Theorem 1 quantify the optimality gaps for the fluid and stochastic-fluid prescriptions, provided that those prescriptions are consistent with their specifications in

the theorem. There remains to show when this is indeed the case. We do so in this section by considering the stationary-demand case; this is equivalent to considering a single period, so we let $k=1$ in what follows. We consider the case with time-varying demand in the following section.

5.1. Solution of the Fluid Problem

The solution to the fluid problem relaxation in (4) is straightforward, and given by the following lemma whose proof we omit.

LEMMA 2. *With a single period or, equivalently, with stationary demand, the solution to the fluid relaxation in (4) is to staff the cheaper resource only, i.e.,*

(a) if $c_0 \le c_1$, *then* $\tilde{m}_\lambda = \lambda/\mu$ *and* $\tilde{n}_\lambda = 0$;

(b) if $c_1 < c_0$, *then* $\tilde{m}_\lambda = 0$ *and* $\tilde{n}_\lambda = \lambda/\mu$.

Lemma 2 makes intuitive sense: The fluid formulation in (4) ignores all "noise" in the system. In particular, it ignores the uncertainty entailed in staffing a flexible resource. Given that demand is stationary (so that there is no need to staff a pool that efficiently scales to demand fluctuations), and assuming out the uncertainty in capacity, a manager would have no incentive to staff a resource that is more expensive. Thus, the optimal fluid-based solution in Lemma 2 is to staff solely from the cheaper resource. In practice, independent contractors are usually cheaper than employees [6], and this is one of the main reasons why they seem to be preferred by several businesses [7]. In what follows, we characterize when reasoning as such would lead to sub-optimal staffing policies.

5.2. Exact Solution for the Stochastic-Fluid Problem

For expositional ease, we define $\beta \equiv (h/\theta + r)\mu$. Here is our first theorem.

THEOREM 2. *In an optimal solution to (3), we must have*

$$\bar{n}_\lambda + \bar{m}_\lambda = \frac{\lambda}{\mu} - \sigma_{\bar{n}_\lambda} F^{-1}\left(\frac{\sigma_{\bar{n}_\lambda}}{\sigma_{\bar{n}_\lambda} + \sigma'_{\bar{n}_\lambda}(\lambda/\mu - \bar{n}_\lambda - \bar{m}_\lambda)}\left(\frac{c_1}{\beta} - \sigma'_{\bar{n}_\lambda}\int_{-1}^{\frac{\lambda/\mu - \bar{n}_\lambda - \bar{m}_\lambda}{\sigma_{\bar{n}_\lambda}}} F(x)dx\right)\right). \tag{5}$$

That is,

$$\bar{m}_\lambda + \bar{n}_\lambda = \lambda/\mu + \mathcal{O}(\sigma_{\bar{n}_\lambda}). \tag{6}$$

The expression in (5) is an implicit relation between $\bar{m}_\lambda$ and $\bar{n}_\lambda$. It is insightful to contrast this with the remarkably simple critical fractile solution (to the stochastic-fluid problem with random arrival rates) in Bassamboo et al. (2010). The additional complexity in our setting is because when the number of servers itself is random, and the choice of capacity levels impacts the distribution of the number of servers, which complicates the solution to the stochastic-fluid problem.

[6] http://smallbusiness.chron.com/costs-employee-vs-independent-contractor-1077.html

[7] http://www.eclewis.com/hiring-independent-contractors/

The expression in (6) is useful because it allows us to quantify the order of magnitude of the required safety hedge: It indicates that the optimal staffing level is in the form of a base capacity, λ/μ, which matches the mean offered load, in addition to an uncertainty hedge, on the order of $\mathcal{O}(\sigma_{\bar{n}_\lambda})$. This additional capacity hedges against the uncertainty in the underlying parameter, i.e., the number of servers, much like the usual square-root additional capacity hedges against stochastic fluctuations in the system (Garnett et al. 2002). The asymptotic form in (6) coincides with the desired asymptotic form in Theorem 1 if $\bar{n}_\lambda = \Theta(\lambda)$: In this case, the solution to the stochastic-fluid formulation will indeed be "extremely" accurate, particularly when the uncertainty in the number of servers is large. The solution in Theorem 2 is algebraically complex and cannot be generally used to derive the exact stochastic-fluid staffing level. Thus, we go further in §5.3, and derive asymptotically optimal solutions to that problem instead. We close this section by discussing a special case where it is easy to derive an exact, closed-form, solution to problem (3).

LEMMA 3. *If $\sigma_n = an$, for $a < 1$, then there exists $\omega < 0$* [8] *such that the solution to (3) is:*

(a) If $c_1 > c_0 + \omega$, then:

$$\bar{m}_\lambda = \frac{\lambda}{\mu} \quad \text{and} \quad \bar{n}_\lambda = 0;$$

(b) If $c_1 \le c_0 + \omega$, then:

$$\bar{m}_\lambda = 0 \quad \text{and} \quad \bar{n}_\lambda = \frac{\lambda}{\mu}\eta^*,$$

where η^* denotes the solution to

$$c_1 + \beta a \int_{-1}^{1/(a\eta)-1/a} F(u)du - \frac{\beta}{\eta} F\left(\frac{1}{a\eta} - \frac{1}{a}\right) = 0.$$

In particular, $\eta^* \ge (\le) 1$ if we also have that $c_1 \le (\ge) \beta F(0) - \beta a \int_{-1}^{0} F(u)du$.

Lemma 3 covers the case where σ_n is "very large". In this case, it is possible for the manager to rely solely on the more expensive, yet certain, supply, i.e., on the fixed resource only. In other words, the manager incurs an additional cost for the uncertainty in capacity, and this cost is quantified by $|\omega|$ in the lemma. This remains true so long as the disparity in pay between the fixed and flexible resources is not too large. Contrasting Lemmas 2 and 3 is insightful because it allows us to coin the impact of uncertainty in the flexible supply. Indeed, when this uncertainty is large but assumed out, as in 2, the manager staffs only from the cheaper resource. Taking this uncertainty into account, as in 3, may lead to *reversing this staffing prescription,* i.e., relying only on the more expensive resource. In other words, hiring independent contractors simply because they are cheaper, which is believed to be an effective staffing strategy in practice [9], could be problematic.

[8] $\omega \equiv c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx < 0$ if $\mathbb{E}[\epsilon] = 0$, where $\alpha_0 \equiv \frac{c_0}{\mu(h/\theta + r)} = \frac{c_0}{\beta}$.

[9] http://money.cnn.com/2015/07/16/pf/independent-contractors-employees/

In some cases, the difference in the staffing costs between the fixed and flexible resources is so large that it remains worthwhile for the system manager to staff only the flexible resource, despite the uncertainty that this entails (this is case (b) in the lemma). *Thus, as a rule of thumb: When the flexible resource is highly uncertain, and demand is stationary, the manager relies exclusively on one of the two resources, but not necessarily the cheaper of the two.*

Interestingly, Lemma 3 also shows that when staffing from the flexible pool, different possibilities emerge: The manager may either match supply and demand (as she would do when staffing from the fixed pool), or intentionally understaff or overstaff her system, depending on how small c_1 is. In other words, conventional wisdom for workforce management, whereby supply and demand are matched in expectation, no longer applies when the pool of flexible workers is highly uncertain.

5.3. Asymptotic Analysis

We let $\Pi^*_\lambda \equiv \Pi_\lambda(m^*_\lambda, n^*_\lambda)$ denote the optimal objective value to the original problem in (2); we also let $\hat{m}_\lambda$ and $\hat{n}_\lambda$ denote the *asymptotically* optimal solutions to (3), where the exact solutions to that problem are given by the implicit relation in Theorem 2. Here is our main theorem.

THEOREM 3. *The approximate solution to the stochastic-fluid problem in (3), with stationary demand, is given by:*

(a) If $c_0 \le c_1$, then $\hat{m}_\lambda = \lambda/\mu$ and $\hat{n}_\lambda = 0$ is $\mathcal{O}(\sqrt{\lambda})$ optimal, i.e.,

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

(b) If $c_0 > c_1$, then we consider two cases:

(b.I) If $\sigma_n = \Theta(n^q)$, $0 \le q \le 1/2$, then $\hat{m}_\lambda = 0$ and $\hat{n}_\lambda = \lambda/\mu$ is $\mathcal{O}(\sqrt{\lambda})$ optimal, i.e.,

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

(b.II) If $\sigma_n = \Theta(n^q)$, $1/2 < q < 1$, then $\hat{m}_\lambda = 0$ and $\hat{n}_\lambda = \lambda/\mu + \gamma^* \sigma_{\lambda/\mu}$ is $o(\sigma_\lambda)$ optimal, i.e.,

$$\lim_{\lambda\to\infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) - \Pi^*_\lambda}{\sigma_\lambda} = 0,$$

where γ^* denotes the optimal solution to

$$\min_\gamma c_1\gamma + \beta \cdot \mathbb{E}[(-\gamma - \epsilon)^+].$$

We can make several noteworthy observations based on Theorem 3. One main managerial insight which follows from our analysis is that, when demand is stationary, a manager should *not* blend fixed and flexible resources, i.e., using a blended workforce is *not* justified in this case. Intuitively, this is so because when customer demand rates do not vary substantially, there are only two effects

which come into play: (i) the staffing cost (fixed and flexible), and (ii) the supply-side uncertainty (only flexible). The interaction between those two effects dictates the optimal staffing level in the system. *As a rule of thumb: When the flexible resource is moderately uncertain, and demand is stationary, the manager should staff only from the cheaper resource, but the additional safety-capacity hedge depends on whether the fixed or flexible resources are used.* In particular, since the fixed resource does not entail any uncertainty (fixed servers are assumed to always show up), we must have that when $c_0 < c_1$, a manager should rely only on that fixed capacity.

It is interesting to inspect the solution for $c_0 > c_1$, i.e., when the fixed capacity is the more expensive. In this case, the supply-side uncertainty comes into play and the solution to the stochastic-fluid relaxation depends on its order of magnitude, as quantified by σ_λ. In particular, when σ_λ is "small", i.e., of an order that is at most equal to the order of stochastic fluctuations in the system, the manager should rely only on the flexible capacity. In this case, there is no advantage in using an uncertainty hedge, and the familiar square-root staffing hedge (Garnett et al. 2002) would yield the same optimality gap. On the other hand, when σ_λ is "moderately large", i.e., of an order that is larger than the order of stochastic fluctuations in the system but smaller than the order of magnitude of the arrival rate (Lemma 3), the manager would still rely solely on the flexible resource, but it is beneficial in this case to introduce an uncertainty hedge which is larger (in order of asymptotic magnitude) than the familiar square-root staffing hedge. The resulting solution is asymptotically optimal to the original problem in (2), and the corresponding optimality gap is $o(\sigma_\lambda)$. Our numerical results, described in the following section, indicate that this is only a loose upper bound, and that the optimality gap is actually smaller than this.

5.4. Numerical Study: On the Magnitude of the Appropriate Hedge

The results of Theorem 3 are asymptotic results; for completeness, we now provide supporting numerical results to test whether the optimality gaps of Theorem 3 are tight. We show that this is not the case, i.e., the accuracy of the approximate stochastic-fluid solutions is superior to the one specified by the theorem. In what follows, we assume that $c_1 < c_0$ and consider a variance term $\sqrt{\lambda} \leq \sigma_\lambda < \lambda$. Then, the optimal solution to (2) is to staff only flexible servers (cases b.I and b.II in Theorem 3). In Table 3, we let $c_1 = 1/3$, $p = 1$, $h = 1$, $\mu = 1$, and $\theta = 3$. We let ϵ in (1) have a uniform distribution over $[-1, 1]$, and we vary the value of λ from 50 to 1,000.

We consider three functional forms for σ_n: $\sqrt{n}$, $n^{3/4}$, and $n^{0.9}$. In each case, we calculate numerically: the optimal solution to (2), n^*, and the optimal solution to (3), $\bar{n}$. We also calculate the approximate solution, $\hat{n}$, as given by cases (b.I) and (b.II) in Theorem 3: The solution for case I is identical to the fluid-based solution, whereas the solution to case II involves an uncertainty hedge, i.e., $\hat{n} = \lambda/\mu + \gamma^* \sigma_{\lambda/\mu}$. In each case, we calculate corresponding relative and absolute errors.

Table 3 shows that, in general, the approximate solution $\hat{n}$ of Theorem 3 is quite accurate. In particular, for $\sigma_n = \sqrt{n}$ there is no distinct advantage in using an uncertainty hedge. Indeed, when λ increases from 50 to 1,000, i.e., is multiplied by a factor of 20, the optimality gaps for $\bar{n}$ and $\hat{n}$, in the fifth and sixth columns of the table, are almost identical, and both yield absolute errors on the order of magnitude of $\sqrt{\lambda}$.

For $\sigma_n = n^{3/4}$, there is still no noticeable difference in performance between the two solutions: While $\bar{n}$ yields a slightly smaller optimality gap, $\hat{n}$ remains asymptotically accurate. With a large variance in the number of servers, i.e., for $\sigma_n = n^{0.9}$, Table 3 suggests that, while the optimality gap for $\hat{n}$ grows with λ (unlike for $\bar{n}$), it does not increase by much. In particular, our numerical results suggest that the optimality gap in this case is on the order of magnitude of $\mathcal{O}(\sqrt{\lambda})$.

$\sigma_n = \sqrt{n}$

λ	n^*	$\bar{n}$	$\hat{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})\rvert}{\Pi_\lambda(n^*)}$
50	57	53	54	0.333	0.140	1.60	0.674
100	110	105	105	0.360	0.360	0.921	0.921
300	317	309	309	0.604	0.604	0.550	0.550
500	522	511	511	0.895	0.895	0.499	0.499
700	726	713	713	1.07	1.07	0.431	0.431
900	929	915	915	1.18	1.18	0.371	0.371
1000	1031	1016	1016	1.21	1.21	0.344	0.344

$\sigma_n = n^{3/4}$

λ	n^*	$\bar{n}$	$\hat{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})\rvert}{\Pi_\lambda(n^*)}$
50	59	58	59	0.0150	0.0	0.0646	0.0
100	115	114	116	0.0101	4.66×10^{-3}	0.0232	0.0107
300	334	333	336	6.75×10^{-3}	7.67×10^{-3}	5.55×10^{-3}	6.30×10^{-3}
500	551	550	553	6.17×10^{-4}	0.0137	3.12×10^{-4}	6.91×10^{-3}
700	765	764	768	3.07×10^{-3}	0.0144	1.13×10^{-3}	5.30×10^{-3}
900	979	978	982	1.49×10^{-3}	0.0154	4.29×10^{-4}	4.44×10^{-3}
1000	1086	1085	1089	0.0	0.0184	0.0	4.80×10^{-3}

$\sigma_n = n^{0.9}$

λ	n^*	$\bar{n}$	$\hat{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\hat{n})\rvert}{\Pi_\lambda(n^*)}$
50	59	58	67	8.28×10^{-3}	0.346	0.0295	1.23
100	118	117	132	3.56×10^{-3}	0.593	6.60×10^{-3}	1.10
300	353	352	385	2.05×10^{-3}	1.16	1.34×10^{-3}	0.758
500	588	587	634	8.53×10^{-4}	1.57	3.41×10^{-4}	0.626
700	822	821	882	9.29×10^{-4}	1.99	2.69×10^{-4}	0.570
900	1056	1055	1128	6.24×10^{-4}	2.29	1.42×10^{-4}	0.520
1000	1173	1172	1251	4.11×10^{-4}	2.46	0.0	0.506

Table 3 Optimality gaps for the exact ($\bar{n}$) and approximate ($\hat{n}$) solutions of the stochastic-fluid problem in (3).

6. Capacity Sizing with Time-Varying Demand

Our analysis thus far has focused on the setting with stationary demand. We now consider a setting where demand varies according to predictable patterns, e.g., due to seasonality effects. In other words, we focus here on the case with time-varying, yet deterministic, demand rates. In §7, we consider random demand rates as well.

Since our aim is to formulate general insights, we consider a simple model with only two periods: A high-demand period with arrival rate λ_H, and a low-demand period with arrival rate λ_L; we assume that $\lambda_H > \lambda_L$. Consistently with our notation in §4, we fix $\lambda > 0$ and let $\lambda_i = \lambda\xi_i$, where $\xi_i \geq 0$ for $i \in \{H, L\}$. To derive asymptotic results, we let λ increase without bound. Our original capacity-sizing problem in (2), reduced to a two-period setting, can be formulated as:

$$\max_{m_\lambda, n_\lambda, \alpha_\lambda^H, \alpha_\lambda^L} \Pi_\lambda(m_\lambda, n_\lambda, \alpha_\lambda^L, \alpha_\lambda^H) = T_H \Gamma_H(m_\lambda, n_\lambda, \alpha_\lambda^H) + T_L \Gamma_L(m_\lambda, n_\lambda, \alpha_\lambda^L), \tag{7}$$

where we define:

$$\begin{aligned}
\Gamma_H(m_\lambda, n_\lambda, \alpha_\lambda^H) &\equiv c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\mathbb{E}[(X^H(m_\lambda, \alpha_\lambda^H n_\lambda) - N(m_\lambda, \alpha_\lambda^H n_\lambda))^+], \\
\Gamma_L(m_\lambda, n_\lambda, \alpha_\lambda^L) &\equiv c_0 m_\lambda + c_1 \alpha_\lambda n_\lambda + (h + r\theta)\mathbb{E}[(X^L(m_\lambda, \alpha_\lambda^L n_\lambda) - N(m_\lambda, \alpha_\lambda^L n_\lambda))^+].
\end{aligned}$$

Given (7), it is readily seen that we can normalize $\alpha^H = 1$, i.e., we can drop it from the problem formulation. We do so, hereafter, and assume that a cap α_λ^L (which we denote by α_λ, dropping dependence on the period) is applied to the flexible pool in the low-demand period.

6.1. Optimal solution

We now turn to formulating and describing the solutions of the two approximations, fluid and stochastic-fluid, of the original problem in (7).

6.1.1. Fluid Problem. For our fluid approximation, we reduce the formulation in (4) to a two-period setting, and obtain the following:

$$\begin{aligned}
\min_{m_\lambda, n_\lambda, \alpha_\lambda} \tilde{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda) = {} & T_H \left(c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r)(\lambda_H - m_\lambda \mu - n_\lambda \mu)^+\right) \\
& + T_L \left(c_0 m_\lambda + c_1 \alpha_\lambda n + (h/\theta + r)(\lambda_L - m_\lambda \mu - \alpha_\lambda n_\lambda \mu)^+\right).
\end{aligned} \tag{8}$$

We are now ready to describe the optimal solution, $(\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda)$, to the problem in (8). This solution depends on both the staffing costs and the lengths of the respective periods. The proof of the following lemma is straightforward and will therefore be omitted.

LEMMA 4. *The solution to the fluid problem in (8), with time-varying demand, is as follows:*

i) If $c_0 \leq c_1$, then there are two subcases:

(a) If $c_0 \leq \frac{T_H}{T_H+T_L} c_1$*, then:*

$$\tilde{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \tilde{n}_\lambda = 0, \quad \textit{and} \quad \tilde{\alpha}_\lambda = 0;$$

(b) If $\frac{T_H}{T_H+T_L} c_1 < c_0 \leq c_1$*, then:*

$$\tilde{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \tilde{n}_\lambda = \frac{\lambda_H}{\mu} - \frac{\lambda_L}{\mu}, \quad \textit{and} \quad \tilde{\alpha}_\lambda = 0;$$

ii) If $c_0 > c_1$*, then:*

$$\tilde{m}_\lambda = 0, \quad \tilde{n}_\lambda = \frac{\lambda_H}{\mu}, \quad \textit{and} \quad \tilde{\alpha}_\lambda = \frac{\lambda_L}{\lambda_H}.$$

Lemma 4 coins the advantage of staffing a pool of flexible agents: Such a pool can be dynamically adjusted to meet time-fluctuations in customer demand, e.g., by setting a cap α_λ on supply in the low-demand period. Indeed, assuming out the uncertainty in staffing a flexible pool (fluid approximation), it is clear that if flexible servers are cheaper, i.e., $c_1 < c_0$, then the manager should staff only flexible servers and set a cap in the low-demand period; this is case (ii) in Lemma 4.

The only case where the manager would staff fixed servers is if $c_1 \geq c_0$. 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, as in case $i(a)$. The threshold $\frac{T_H}{T_H+T_L}$ in case (i) is the length of the high-demand period, T_H, relative to the length of the entire horizon, $T_L + T_H$. In words, as the high-demand period increases (decreases) in length, it becomes more (less) cost-effective for the manager to staff a pool of fixed servers to match demand in the high period, since the relative cost of overstaffing the low period decreases (increases).

6.1.2. Stochastic-Fluid Problem. The stochastic-fluid optimization problem is given by:

$$\begin{aligned} \min_{m_\lambda, n_\lambda, \alpha_\lambda} \bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda) = {} & T_H \left(c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r)\mathbb{E}\left[(\lambda_H - N(m_\lambda, n_\lambda)\mu)^+\right]\right) \\ & + T_L \left(c_0 m_\lambda + c_1 \alpha_\lambda n_\lambda + (h/\theta + r)\mathbb{E}\left[(\lambda_L - N(m_\lambda, \alpha_\lambda n_\lambda)\mu)^+\right]\right). \end{aligned} \quad (9)$$

For the solution of (9), we begin by treating the special case where σ_n is a linear function of n. For ease of exposition, we define $\alpha_0 \equiv c_0\theta/((h + r\theta)\mu)$ and $\beta_1 \equiv \alpha_0(T_H + T_L)/T_H$. We also recall that $\beta \equiv \left(\frac{h}{\theta} + r\right)\mu$. Additionally, it will be convenient to define the constants $\zeta_0 < \zeta_1 < c_0$ given by:

$$\zeta_0 \equiv c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx,$$

and

$$\zeta_1 \equiv c_0 + c_0 a F^{-1}(\beta_1) - a\beta \int_{-1}^{F^{-1}(\beta_1)} F(x)dx.$$

The asymptotic accuracy of the following (exact) solution is given in Theorem 1.

LEMMA 5. *If $\sigma_n = an$ for $a < 1$, then $(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda)$ is given as follows.*

(a) If $c_1 \geq \zeta_1$, then:

$$\bar{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \bar{n}_\lambda = 0, \quad \text{and} \quad \bar{\alpha}_\lambda = 0.$$

(b) If $\zeta_0 \leq c_1 < \zeta_1$, then:

$$\bar{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \bar{n}_\lambda = \frac{(\lambda_H - \lambda_L)\eta_2^*}{\mu}, \quad \text{and} \quad \bar{\alpha}_\lambda = 0;$$

(c) If $c_1 < \zeta_0$, then:

$$\bar{m}_\lambda = 0, \quad \bar{n}_\lambda = \frac{\lambda_H}{\mu}\eta_2^*, \quad \text{and} \quad \bar{\alpha}_\lambda = \frac{\lambda_L}{\lambda_H},$$

where η_2^* is the solution of

$$c_1 + \beta a \int_{-1}^{(1/(a\eta))-1/a} F(x)dx - \frac{\beta}{\eta} F\left(\frac{1}{a\eta} - \frac{1}{a}\right) = 0.$$

In particular, $\eta_2^* \geq (\leq) 1$ if we also have that $c_1 \leq (\geq) \beta F(0) - \beta a \int_{-1}^{0} F(u)du$.

Consistently with Lemma 3, Lemma 5 shows that when the flexible pool is highly variable, then it may be possible to staff the *more expensive* but fixed pool; this is case (a) in the lemma. This is only true, however, when the flexible resource is expensive enough, i.e., $c_1 > \zeta_1$. Indeed, when c_1 is moderately cheaper than the fixed capacity (case (b)), then it is cost-effective for the manager to use a blended workforce: She uses the fixed pool to match demand in the low period, and the flexible pool to staff up to demand in the high period. This lends support to existing business practices where managers typically resort to hiring temporary help during periods of peak in demand (the flexible pool is capped in the low period). This is no longer the case, however, when the flexible capacity is cheap enough (case (c)), in which case the manager staffs only from the flexible resource. *As a rule of thumb: When the flexible resource is highly uncertain but cheaper, demand is time-varying, and the disparity in compensation is moderate, the manager relies on a blended workforce. She relies exclusively on the flexible resource only when it is much cheaper and, otherwise, relies exclusively on the possibly more expensive fixed resource. When she uses flexible servers, then she also caps that supply in the low-demand period.*

We now derive asymptotically optimal solutions to problem (9), paralleling our analysis in §5.3. We denote those solutions by $(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda)$, and let Π_λ^* denote the optimal objective value for the original problem in (2).

THEOREM 4. *The approximate solution to the stochastic-fluid problem in (9), with time-varying demand, is given by:*

(a) If $\sigma_n = \Theta(n^q)$*, for* $0 \le q \le 1/2$*, then:*

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

for the fluid-optimal solution $(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda)$ *as given by Lemma 4.*

(b) If $\sigma_n = \Theta(n^q)$*, for* $1/2 < q < 1$*, then:*

$$\lim_{\lambda\to\infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda) - \Pi^*_\lambda}{\sigma_\lambda} = 0,$$

where $(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda)$ *are given as follows.*

(b.I) If $c_0 \le \frac{T_H}{T_H+T_L} c_1$*, then:*

$$\hat{m}_\lambda = \frac{\lambda_H}{\mu}, \quad \hat{n}_\lambda = 0, \quad \text{and} \quad \hat{\alpha}_\lambda = 0;$$

(b.II) If $\frac{T_H}{T_H+T_L} c_1 < c_0 \le c_1$*, then:*

$$\hat{m}_\lambda = \frac{\lambda_L}{\mu}, \quad \hat{n}_\lambda = \left(\frac{\lambda_H}{\mu} - \frac{\lambda_L}{\mu}\right) + \gamma_2^* \sigma_{\lambda_H/\mu - \lambda_L/\mu}, \quad \text{and} \quad \hat{\alpha}_\lambda = 0,$$

where γ_2^* *denote the optimal solution of:*

$$\min_\gamma c_1\gamma + \beta\mathbb{E}\left[(-\gamma-\epsilon)^+\right];$$

(b.III) If $c_1 < c_0$*, then:*

$$\hat{m}_\lambda = 0, \quad \hat{n}_\lambda = \frac{\lambda_H}{\mu} + \gamma_3^* \sigma_{\lambda_H/\mu}, \quad \text{and} \quad \hat{\alpha}_\lambda = \frac{\lambda_L}{\lambda_H} + \nu_3^* \frac{\mu}{\lambda_H} \sigma_{\lambda_H/\mu},$$

where γ_3^* *and* ν_3^* *denote the optimal solutions to:*

$$\begin{aligned}\min_{\gamma,\nu} \; & c_1\left(T_H + T_L\frac{\lambda_L}{\lambda_H}\right)\gamma + c_1 T_L \nu + (h/\theta + r)\mu T_H \mathbb{E}\left[(-\gamma-\epsilon)^+\right] \\ & + \beta T_L \mathbb{E}\left[\left(-\frac{\lambda_L}{\lambda_H}\gamma - \nu - \left(\frac{\lambda_L}{\lambda_H}\right)^q \epsilon\right)^+\right].\end{aligned}$$

Theorem 4 demonstrates that if the uncertainty in the flexible pool is "low", i.e., of an order of magnitude which is smaller than the square-root order of stochastic fluctuations in the system, then the solution to the problem remains similar to the fluid-optimal solution in Lemma 4.

More generally, when supply is uncertain but not greatly so (smaller than that in Lemma 5), we find that the manager should generally rely on the flexible pool, either alone or through blending. Indeed, the only case where the manager would rely solely on the fixed pool is case (b.I), where the fixed supply is much cheaper than the flexible one. Otherwise, when the fixed servers are cheaper than flexible ones, but not by too much (case (b.II)) then the manager should blend her workforce, matching with the fixed pool up to demand in the low period, and utilizing flexible agents to match up to demand in the high period. Finally, when flexible servers are cheaper, then the manager

should depend only on those servers, as in case (b.III). In this case, she matches up to demand in the high period, and caps her supply in the low period. *As a rule of thumb: When the flexible resource is moderately uncertain, and demand is time-varying, the manager relies exclusively on the fixed resource only when it is much cheaper. Otherwise, she staffs a blended workforce when the fixed resource is cheaper but not greatly so, and staffs exclusively flexible servers when the flexible resource is cheaper. When she staffs from the flexible resource, then she also uses a cap in the low-demand period.*

7. Random Arrival Rate

In this section, we extend our modelling framework by considering a random arrival rate, Λ_λ. In so doing, we are motivated by empirical evidence suggesting that arrival counts in service systems tend to exhibit higher variance than implied by the Poisson distribution assumption. Indeed, a doubly stochastic Poisson arrival process with a random arrival rate is usually a better fit to data (Aldor-Noiman et al. 2009, Jongbloed and Koole 2001). In this section, we consider two types of parameter uncertainty: (i) uncertainty in the arrival rates, and (ii) uncertainty in the number of servers, and derive asymptotic accuracy results and the asymptotically optimal staffing policy. To illustrate the impact of randomness in arrivals, we consider the case with a time-invariant arrival rate. Our analysis can be readily extended to both time-variation and randomness in arrivals.

7.1. Problem Formulation

Paralleling (1), we assume that Λ_λ is given by:

$$\Lambda_\lambda = \lambda + \eta_\lambda \delta, \tag{10}$$

for some random variable $-1 \leq \delta \leq 1$ with $\mathbb{E}[\delta] = 0$. We assume that δ has a strictly positive pdf, g_δ on $(-1, 1)$, and cdf G_δ. We assume that $\eta_\lambda \geq 0$ is some function of λ. Paralleling (3), we can write the stochastic-fluid relaxation of the problem as:

$$\begin{aligned}
\min_{m_\lambda, n_\lambda} \Pi_\lambda(m_\lambda, n_\lambda) &= c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\mathbb{E}\left[(X(\Lambda_\lambda, m_\lambda, n_\lambda) - N(m_\lambda, n_\lambda))^+\right] \qquad (11)\\
&= c_0 m_\lambda + c_1 n_\lambda + \beta \mathbb{E}\left[\left(\frac{\Lambda_\lambda}{\mu} - N(m_\lambda, n_\lambda)\right)^+\right]\\
&= c_0 m_\lambda + c_1 n_\lambda + \beta \mathbb{E}\left[\left(\frac{\Lambda_\lambda}{\mu} + \eta_\lambda \delta - m_\lambda - n_\lambda - \sigma_{n_\lambda}\epsilon\right)^+\right].
\end{aligned}$$

where $X(\Lambda_\lambda, m_\lambda, n_\lambda)$ denotes the steady-state number of customers in the system, and we recall $\beta \equiv \mu(h/\theta + r)$. The fluid relaxation of the problem coincides with (4), since that formulation ignores both stochastic variability and parameter uncertainty in the system. The solution to that problem is given in Lemma 2, and indicates that the optimal solution is to match mean offered load with the cheaper of the two resources, fixed or flexible.

7.2. Asymptotic Accuracy

Paralleling Theorem 1, we quantify in Theorem 5 the asymptotic accuracy of the optimal solutions $(\tilde{m}_\lambda, \tilde{n}_\lambda)$ to (4) and $(\bar{m}_\lambda, \bar{n}_\lambda)$ to (11).

THEOREM 5. *We assume that* $\eta_\lambda = \Theta(\lambda^p)$ *and* $\sigma_\lambda = \Theta(\lambda^q)$*, where* $p, q > 0$*. For large* λ*,*

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

that is, if $\tilde{n}_\lambda = \Theta(\lambda)$ *then* $\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\max\left\{\sqrt{\lambda}, \sigma_\lambda, \eta_\lambda\right\}\right)$.
Also, for large λ*,*

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

Furthermore,

(a) if $p = \max\{p, q\} > 1/2$ *and* $\bar{m}_\lambda + \bar{n}_\lambda = \mathcal{O}(\lambda)$*, then*

$$\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\min\left\{\sqrt{\lambda}, \lambda/\eta_\lambda\right\}\right);$$

(b) if $q = \max\{p, q\} > 1/2$ *and* $\bar{n}_\lambda = \Theta(\lambda)$*, then*

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

Theorem 5 illustrates that the asymptotic accuracy of those optimal staffing prescriptions depends on which of the two types of parameter uncertainty, in arrivals or in capacity, dominates asymptotically, i.e., for large λ, and how the dominating uncertainty compares to the square-root order of stochastic fluctuations in the system. Indeed, we can identify two operating modes, depending on the magnitude of variability in *either* arrivals or capacity, as quantified by η_λ and σ_n, respectively. When either η_λ or σ_λ is "large", i.e., larger than $\sqrt{\lambda}$, the system can be considered to be in an uncertainty-dominated regime where stochastic-fluid approximations become increasingly accurate as that variability increases. Otherwise, the system is in a variability-dominated regime where the optimality gap for the stochastic-fluid approximation is on the order of stochastic fluctuations in the system, i.e., $\mathcal{O}(\sqrt{\lambda})$. Since the fluid approximation in (4) ignores both types of uncertainty, the corresponding optimality gap is dominated by either the order of stochastic fluctuations, or the order of variability in either arrivals or uncertainty, whichever is greatest.

7.3. Asymptotically Optimal Staffing Policy

There remains to derive an asymptotically optimal solution $(\hat{m}_\lambda, \hat{n}_\lambda)$ to (11), as we did in Theorem 3 for problem (3). We do so in Theorem 6. There, we restrict attention to the case where $\sigma_n = o(n)$ and $\eta_\lambda = o(\lambda)$, and highlight the main distinctions compared to the optimal staffing policy in §5.

THEOREM 6. *We assume that* $\eta_\lambda = \Theta(\lambda^p)$ *and* $\sigma_\lambda = \Theta(\lambda^q)$*, where* $p, q > 0$*. The approximate solution to problem (11) is given by:*

(a) If $c_0 \le c_1$*, then* $\hat{m}_\lambda = \frac{\lambda}{\mu} + \frac{1}{\mu}\bar{G}_\delta^{-1}\left(\frac{c_0/\mu}{p+h/\theta}\right)\eta_\lambda$ *and* $\hat{n}_\lambda = 0$ *is* $\mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\eta_\lambda\})$ *optimal, i.e.,*

$$\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) = \Pi_\lambda^* + \mathcal{O}(\min\{\sqrt{\lambda}, \lambda/\eta_\lambda\}).$$

(b) If $c_1 < c_0$*, then we consider three cases:*

(b.I) If $p > q$*, then* $\hat{m}_\lambda = 0$ *and* $\hat{n}_\lambda = \lambda/\mu + \gamma_1^*\eta_\lambda$ *is* $o(\eta_\lambda)$ *optimal, i.e.,*

$$\lim_{\lambda\to\infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) - \Pi_\lambda^*}{\eta_\lambda} = 0,$$

where γ_1^* *denotes the optimal solution to* $\min_\gamma c_1\gamma + \beta\mathbb{E}\left[(\delta/\mu - \gamma)^+\right]$.

(b.II) If $q > p$*, then* $\hat{m}_\lambda = 0$ *and* $\hat{n}_\lambda = \lambda/\mu + \gamma_2^*\sigma_{\lambda/\mu}$ *is* $o(\sigma_\lambda)$ *optimal, i.e.,*

$$\lim_{\lambda\to\infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) - \Pi_\lambda^*}{\sigma_\lambda} = 0,$$

where γ_2^* *denotes the optimal solution to* $\min_\gamma c_1\gamma + \beta\mathbb{E}\left[(-\gamma - \epsilon)^+\right]$.

(b. III) If $p = q$*, i.e., there exists* $a > 0$ *such that* $\eta_\lambda = a\sigma_{\lambda/\mu}$*, then:* $\hat{m} =$ *and* $\hat{n}_\lambda = \lambda/\mu + \gamma_3^*\sigma_{\lambda/\mu}$ *is* $o(\sigma_\lambda)$ *optimal, i.e.,*

$$\lim_{\lambda\to\infty} \frac{\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) - \Pi_\lambda^*}{\sigma_\lambda} = 0,$$

where γ_3^* *is the solution to* $\min_\gamma c_1\gamma + \beta\mathbb{E}\left[(a\delta/\mu - \gamma - \epsilon)^+\right]$.

Consistent with our general insights in §5, the optimal staffing prescription $\bar{m}_\lambda, \bar{n}_\lambda)$ corresponds to augmenting the fluid-based optimal solution, $(\tilde{m}, \tilde{n})$ in Lemma 2 with an appropriate safety-capacity hedge. When considering both randomness in arrivals and in capacity, the magnitude of that hedge depends on which of the two types of uncertainty dominated. In particular, if $p > q$, i.e., the uncertainty in arrivals dominates, then that hedge must be on the order of variability in arrivals, i.e., η_λ. Conversely, if $q > p$, i.e., the uncertainty in capacity dominates, then that hedge must be on the order of variability in the number of servers, i.e., σ_λ. Either way, the asymptotic order of accuracy of the solution is negligible compared to the order of magnitude of variability in the underlying parameter, i.e., it is either $o(\sigma_\lambda)$ or $o(\eta_\lambda)$, whichever is greater. Otherwise, our main insights from §5, concerning the optimal staffing policy, continue to hold.

8. Generally-Distributed Abandonment

The results of the previous sections were restricted to exponentially-distributed patience times. Since there is statistical evidence indicating that patience times may not always be exponential (Brown et al. 2005), it is important to go beyond this assumption. We do so in this section by describing results from a numerical study quantifying the optimality gaps for problems (3) and (4) with a non-exponential abandonment distribution.

In Tables 4 and 5, we present our results for Pareto (mean 1, shape 2) and Weibull (mean 1, shape 2) abandonment. We choose these two distributions because they exhibit, for those selected parameter values, different properties for their failure-rate functions: While the Pareto distribution had a decreasing failure rate, the Weibull distribution has an increasing failure rate. We consider the following cost parameters: $c = 1$, $h = 1$, $p = 0.45$, and $\mu = 1$, and restrict attention to a system with only flexible capacity. In each case, we report the fluid, $\tilde{n}$, stochastic-fluid, $\bar{n}$, and original, n^*, optimal solutions, as well as the corresponding optimality gaps, both relative and absolute.

We first discuss our numerical results with Pareto abandonment. In this case, the overloaded regime is asymptotically optimal at fluid scale (Bassamboo and Randhawa 2010). When $\sigma_n = o(n)$ (first two sub-tables in Table 4), the system remains overloaded despite the uncertainty in the number of servers. Thus, we expect that fluid prescriptions should be extremely accurate, i.e., with absolute errors on the order of magnitude of $\mathcal{O}(1)$. In other words, we expect that stochastic-fluid prescriptions would not lead to a substantial improvement over their fluid counterparts; this is confirmed by Table 4. It is unclear, a priori, how the fluid and stochastic-fluid solutions would perform when σ_n is large, i.e., of an order of magnitude equal to $\mathcal{O}(n)$ (last sub-table in Table 4), because the uncertainty in the number of servers is on the same order as the offered load in this case. Table 4 shows that, while stochastic-fluid approximations are more accurate in this case, the difference in performance in not too great.

With Weibull abandonment, the fluid solution prescribes a critically-loaded regime. Thus, we expect the optimality gaps of our respective solutions to be close to those with exponential abandonment (Theorem 1). Table 5 confirms that this is indeed the case. In particular, for all values of σ_n, the stochastic-fluid formulation is remarkably accurate, yielding an order of magnitude improvement over the fluid prescription (in most cases, n^* and $\bar{n}$ are indistinguishable). Moreover, Table 5 shows that the optimality gaps obtained are consistent with those reported in Theorem 1.

9. Concluding Remarks

In this paper, we studied the problem of staffing a service system with a blended workforce. In particular, we provided answers to two related questions: (1) Should a manager rely on a blended workforce and, if so, when? and (2) How many flexible or fixed agents should a manager staff, and what are the safety capacity hedges that the manager should use?

Our analysis suggests that the answer to the first question is not straightforward, and that it may be cost-effective to staff either strictly one of the two resources, or to use a blended workforce instead, depending on the interaction between three competing factors: (i) operational costs in the system; (ii) the time-variation in customer demand; and (iii) the supply-side uncertainty which is associated with staffing flexible agents. To answer the second question, we showed that the

Pareto, $\sigma_n = \sqrt{n}$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
50	35	41	41	0.388	0.388	0.770	0.770
100	75	83	83	0.315	0.315	0.315	0.315
300	242	248	248	0.0325	0.0325	0.0109	0.0109
500	411	413	413	2.06×10^{-3}	2.06×10^{-3}	4.13×10^{-3}	4.13×10^{-4}
650	536	537	537	1.89×10^{-4}	1.89×10^{-3}	2.90×10^{-5}	2.90×10^{-5}

Pareto, $\sigma_n = n^{3/4}$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
50	33	36	41	0.152	1.025	0.300	2.01
100	70	75	83	0.196	1.375	0.194	1.36
300	226	240	248	0.411	1.15	0.137	0.383
500	389	409	413	0.526	0.789	0.105	0.158
650	512	537	537	0.583	0.583	0.0898	0.0898

Pareto, $\sigma_n = 0.25n$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
50	34	39	41	0.278	0.592	0.550	1.17
100	72	79	83	0.324	0.861	0.322	0.857
300	226	241	248	0.485	1.18	0.161	0.392
500	383	401	413	0.447	1.43	0.089	0.286
650	501	521	537	0.425	1.64	0.0655	0.252

Table 4 Performance of the stochastic-fluid and fluid optimal solutions with Pareto abandonment.

optimal staffing levels involve both a base capacity, which is used to match mean demand, and an additional safety capacity which hedges against both stochastic fluctuations in the system and the variability due to the randomness in supply. This additional safety capacity hedge may or may not be consistent with the square-root staffing hedge (Garnett et al. 2002), and generally depends on the mix of fixed and flexible resources in the staffed pool.

Part of our analysis provides support to current business practices. For example, we showed that a manager who uses only flexible agents should staff enough to match peak loads, but will resort to capping the number of active agents under low loads; e.g., this is common in virtual call centers which hire work-from-home agents [10], and is also consistent with previous results from the strategic-server literature (Gurvich et al. 2017). We also showed that staffing flexible agents is particularly desirable to meet customer demand fluctuations and that, barring such fluctuations, the cheaper resource is generally the most desirable alternative.

However, our analysis also yields results which challenge other current business trends. For example, one of our main insights is that it may *not* always be cost-effective to staff a blended

[10] https://www.glassdoor.co.uk/Reviews/Arise-Reviews-E31617.htm

Weibull, $\sigma_n = \sqrt{n}$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
50	54	54	50	0	0.8702	0	1.48
100	107	107	100	0	2.58	0	2.29
300	315	315	300	0	10.0	0	3.09
500	521	520	500	0.0197	17.4	3.70×10^{-3}	3.27
650	674	673	650	0.0673	22.7	9.80×10^{-3}	3.31

Weibull, $\sigma_n = n^{3/4}$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
50	50	50	50	0	0	0	0
100	103	104	100	4.40×10^{-3}	0.153	3.53×10^{-3}	0.123
300	320	321	300	1.19×10^{-3}	2.63	3.3×10^{-4}	0.728
500	538	539	500	2.03×10^{-3}	6.53	3.43×10^{-4}	1.10
650	702	702	650	0	9.98	0.0	1.31

Weibull, $\sigma_n = 0.25n$

λ	n^*	$\bar{n}$	$\tilde{n}$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert$	$\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\bar{n})\rvert}{\Pi_\lambda(n^*)}$	$100 \cdot \frac{\lvert\Pi_\lambda(n^*) - \Pi_\lambda(\tilde{n})\rvert}{\Pi_\lambda(n^*)}$
50	52	53	50	0.00269	0.220	4.38×10^{-4}	0.359
100	105	105	100	0.0	0.520	0.0	0.427
300	316	316	300	0.0	1.64	0.0	0.452
500	527	527	500	0.0	2.76	0.0	0.457
650	685	685	650	0.0	3.60	0.0	0.459

Table 5 Performance of the stochastic-fluid and fluid optimal solutions with Weibull abandonment.

workforce, i.e., that the modern shift in the business world towards staffing such a workforce should be handled with caution. We also showed that it may be cost-effective to, counter-intuitively, staff from a *more expensive* resource. For example, this is the case when the variability in the flexible supply is large, i.e., flexible agents are highly unreliable. Finally, we showed that even if the reliable fixed resource is cheaper, then it may still be optimal to staff flexible agents chiefly because doing so allows for the flexibility of adjusting the agent pool size dynamically to meet fluctuations in incoming customer demand.

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 their agents well in advance (Taylor 2017). Nevertheless, it remains interesting to investigate the dynamic compensation decision in a setting with a blended workforce, both separately and jointly with the staffing decision. Since the staffing decision depends on the compensation in a non-trivial way, e.g.,

in that it may be optimal to staff a more expensive and more reliable resource, such a study promises to yield important insights into the effective management of a blended workforce in practice.

References

Accenture. 2016. Blurring the line between amployee and contractor. `https://www.accenture.com/us-en/insight-future-workforce-trends`. Accessed: 2017-02-08.

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

Aldor-Noiman, Sivan, Paul D Feigin, Avishai Mandelbaum. 2009. Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics* 1403–1447.

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

Balakrishnan, Anantaram, Michael S Pangburn, Euthemia Stavrulaki. 2008. Integrating the promotional and service roles of retail inventories. *Manufacturing & Service Operations Management* **10**(2) 218–235.

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.

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

Braverman, A., J. Dai, X. Liu, L. Ying. 2017. Empty-car routing in ridesharing systems. Cornell University, working paper.

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.

Brumm, F. 2017. Making gigs work: The new economy in context. University of Illinois at Urbana Champaign, Masters thesis.

Cachon, G., K. Daniels, R. Lobel. 2017. The role of surge pricing on a service platform with self-scheduling capacity. University of Pennsylvania, working paper.

Cachon, G., P. Harker. 2002. Competition and outsourcing with scale economies. *Management Science* **48**(10) 1314–1333.

Cachon, G., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science* **53**(3) 408–420.

Deloitte. 2016. The gig economy distraction or disruption? https://dupress.deloitte.com/dup-us-en/focus/human-capital-trends/2016/gig-economy-freelance-workforce.html. Accessed: 2017-02-08.

Feinberg, R., K. de Ruyter, L. Bennington. 2005. *Cases in Call Center Management: Great Ideas (th)at Work*. Ichor Business Books.

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.

Forbes. 2016. 10 Workplace Trends You'll See In 2017. http://www.forbes.com/sites/danschawbel/2016/11/01/workplace-trends-2017/#548a23ab3457. Accessed: 2017-02-08.

Forbes. 2017. 10 hr trends for 2017. https://www.forbes.com/sites/jeannemeister/2017/01/05/the-employee-experience-is-the-future-of-work-10-hr-trends-for-2017/#40dc1dbb20a6. Accessed: 2017-03-31.

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

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

Gopalakrishnan, Ragavendran, Sherwin Doroudi, Amy R Ward, Adam Wierman. 2016. Routing and staffing when servers are strategic. *Operations Research* **64**(4) 1033–1050.

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. 2017. 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.

Harvard Business Review. 2016. How pwc and the washington post are finding and hiring external talent. https://hbr.org/2016/03/how-pwc-and-the-washington-post-are-finding-and-hiring-external-talent. Accessed: 2017-02-08.

Ibrahim, R. 2017. Staffing a service system with a random service capacity and impatient customers. University College London, working paper.

Intuit. 2016. Twenty trends that will shape the next decade. https://http-download.intuit.com/http.intuit/CMO/intuit/futureofsmallbusiness/intuit_2020_report.pdf. Accessed: 2017-02-08.

Jeon, Jongwoo, Subhash Kochar, Chul Gyu Park. 2006. Dispersive orderingsome applications and examples. *Statistical Papers* **47**(2) 227–247.

Jongbloed, Geurt, Ger Koole. 2001. Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry* **17**(4) 307–318.

Lewis, Toby, JW Thompson. 1981. Dispersive distributions, and the connection between dispersivity and strong unimodality. *Journal of Applied Probability* **18**(01) 76–90.

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.

McKinsey&Company. 2015. A labor market that works: Connecting talent with opporunity in the digital age. `http://www.mckinsey.com/global-themes/employment-and-growth/connecting-talent-with-opportunity-in-the-digital-age`. Accessed: 2017-02-08.

New York Times. 2012. Hiring contractors without getting into trouble. `http://www.nytimes.com/2012/02/02/business/smallbusiness/how-to-hire-independent-contractors-without-getting-in-trouble.html`. Accessed: 2017-03-31.

Ozkan, E., A. Ward. 2017. Dynamic matching for real-time ridesharing. University of Southern California, working paper.

Riquelme, C., S. Banerjee, R. Johari. 2017. Pricing in ride-share platforms: A queueing-theoretic approach. Cornell University, working paper.

Roubos, Alex, Oualid Jouini. 2013. Call centers with hyperexponential patience modeling. *International Journal of Production Economics* **141**(1) 307–315.

Shaked, Moshe, J George Shanthikumar. 2007. *Stochastic orders*. Springer Science & Business Media.

Taylor, T. 2017. On-demand service platforms. University of California Berkeley, working paper.

The Economist. 2013. The workforce in the cloud. `http://www.economist.com/news/business/21578658-talent-exchanges-web-are-starting-transform-world-work-workforce.` Accessed: 2017-02-08.

UKCES. 2016. The future of work jobs and skills in 2030. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/303335/the_future_of_work_key_findings_edit.pdf`. Accessed: 2017-02-08.

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. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.

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.

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.

Zhan, D., A. Ward. 2017. Compensation and staffing to trade off speed and quality in large service systems University of Southern California, working paper.

Electronic Companion:

In this e-companion, we present proofs to the theorems and lemmas of the main paper.

EC.1. Proof of Lemma 1

PROOF. Let $X(N)$ denote the steady-state number in a system with N servers. Since $\mu = \theta$, $X(N)$ has the same distribution as the steady-state number in system of an $M/M/\infty$ queue with arrival rate λ and service rate μ. That is, $X(N)$ has a Poisson distribution with rate λ/μ, and we note that $X(N)$ is independent of N and that $Q(N) = (X(N) - N)^+$. We recall the following definition of variability ordering between two random variables X and Y with respective cdf's F_X and F_Y (Shaked and Shanthikumar 2007). The inverse F^{-1} of a cdf F is defined as:

$$F^{-1}(t) = \inf\{x : F(x) > t\} \quad \text{for} \quad 0 < t < 1.$$

DEFINITION EC.1. X is said to be *less dispersed than* Y, denoted by $X \leq_{disp} Y$, if $F_X^{-1}(\beta) - F_X^{-1}(\alpha) \leq F_Y^{-1}(\beta) - F_Y^{-1}(\alpha)$ for all $0 < \alpha < \beta < 1$.

If $\sigma_1 \leq \sigma_2$, then $N_1 \leq_{disp} N_2$. To see why, define the functions $\phi(x) \equiv n + \sigma_1 x$ and $\psi(x) \equiv n + \sigma_1 x$. Recall that $\phi \leq_{disp} \psi$ if, and only if, $\phi' \leq \psi'$. We can write $N_1 = \phi(\epsilon)$ and $N_2 = \psi(\epsilon)$. Now, using Theorem 3.B.5 of Shaked and Shanthikumar (2007):

$$N_1 \leq_{disp} N_2 \quad \text{if, and only if,} \quad \phi \leq_{disp} \psi,$$

yields that $N_1 \leq_{disp} N_2$, as desired. Thus, we also have: $-N_1 \leq_{disp} -N_2$. Since $X \sim$Poisson(λ/μ) has a log-concave probability mass function, and is independent of both N_1 and N_2, we can use Lewis and Thompson (1981) to obtain:

$$X - N_1 \leq_{disp} X - N_2.$$

As $\mathbb{E}[X - N_1] = \mathbb{E}[X - N_2]$ and $u(x) = x^+ \equiv \max\{x, 0\}$ non-decreasing and convex, we obtain that (Jeon et al. 2006):

$$\mathbb{E}[(X - N_1)^+] \leq \mathbb{E}[(X - N_2)^+] \quad \text{i.e.,} \quad \mathbb{E}[Q(N_1)] < \mathbb{E}[Q(N_2)].$$

■

EC.2. Proof of Theorem 1

We first state and prove Lemmas EC.1-EC.3, which will be useful for our proof of Theorem 1.

EC.2.1. Additional Lemmas

LEMMA EC.1. *When $\mu=\theta$,*

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

where $N(m_\lambda,n_\lambda) = m_\lambda + n_\lambda + \sigma_{n_\lambda}\epsilon$. Moreover, if $n_\lambda = \Theta(\lambda)$ and $\sigma_{n_\lambda} > \sqrt{n_\lambda}$ then:

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

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

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

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

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

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 $y \ge 0$:

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

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 $M_\lambda(y) = O(\lambda/\sigma_\lambda)$. Now,

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

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

In addition,

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

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:

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

■

Lemma EC.2.

$$
\bar{\Pi}_\lambda(m,n)\leq\Pi_\lambda(m,n)\leq\bar{\Pi}_\lambda(m,n)+O(\sqrt{\lambda})
$$

Moreover, when $n_\lambda=\Theta(\lambda)$ *and* $\sigma_{n_\lambda}\geq\sqrt{n_\lambda}$*:*

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

Proof. We prove the first statement in detail. The second statement, where $n_\lambda=\Theta(\lambda)$ and $\sigma_{n_\lambda}\geq\sqrt{n_\lambda}$, follows along the same line of arguments.

When $\mu=\theta$, the result follows directly from Lemma EC.1.

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 $A_\lambda^I(N(m_\lambda,n_\lambda))\equiv\mu\mathbb{E}[(X(m_\lambda,n_\lambda;\mu,\mu)-N(m_\lambda,n_\lambda))^+]$ where $X(m_\lambda,n_\lambda;x,y)$ is the steady-state number-in-system with service rate x and abandonment rate y. As $A_\lambda(N(m_\lambda,n_\lambda))\leq A_\lambda^I(N(m_\lambda,n_\lambda))$ (Bassamboo et al. 2010):

$$
\Pi_\lambda(m_\lambda,n_\lambda)=c_0m_\lambda+c_1n_\lambda+(h/\theta+r)A_\lambda(N(m_\lambda,n_\lambda))
$$

$$
\begin{aligned}
&\le c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda^I(N(m_\lambda, n_\lambda)) \\
&\le c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] + \mathcal{O}(\sqrt{\lambda}) \quad \text{by Lemma EC.1} \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda) + O(\sqrt{\lambda}).
\end{aligned}
$$

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 $A_\lambda^{II}(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \theta, \theta) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \ge A_\lambda^{II}(N(m_\lambda, n_\lambda)\mu/\theta)$ (Bassamboo et al. 2010):

$$
\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda\left(N(m_\lambda, n_\lambda)\right) \\
&\ge c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda^{II}\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right) \\
&\ge c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)^+\right] \quad \text{by Lemma EC.1} \\
&= c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}\left[(\lambda/\mu - N(m_\lambda, n_\lambda))^+\right] \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda).
\end{aligned}
$$

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 $A_\lambda^{II}(N(m_\lambda, n_\lambda)) \equiv \theta\mathbb{E}[(X(m_\lambda, n_\lambda; \theta, \theta) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \le A_\lambda^{II}\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)$ (Bassamboo et al. 2010):

$$
\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda)) \\
&\le c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda^{II}\left(\frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right) \\
&\le c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\mathbb{E}\left[\left(\frac{\lambda}{\theta} - \frac{\mu}{\theta} m_\lambda - \frac{\mu}{\theta} N(m_\lambda, n_\lambda)\right)^+\right] + O(\sqrt{\lambda}) \text{ by Lemma EC.1} \\
&= c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda) + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}
$$

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 $A_\lambda^I(N(m_\lambda, n_\lambda)) \equiv \mu\mathbb{E}[(X(N(m_\lambda, n_\lambda); \mu, \mu) - N(m_\lambda, n_\lambda))^+]$. As $A_\lambda(N(m_\lambda, n_\lambda)) \ge A_\lambda^I(N(m_\lambda, n_\lambda))$ (Bassamboo et al. 2010):

$$
\begin{aligned}
\Pi_\lambda(m_\lambda, n_\lambda) &= c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda(N(m_\lambda, n_\lambda)) \\
&\ge c_0 m_\lambda + c_1 n_\lambda + (h/\theta + r) A_\lambda^I(N(m_\lambda, n_\lambda)) \\
&\ge c_0 m_\lambda + c_1 n_\lambda + (h + r\theta)\frac{\mu}{\theta}\mathbb{E}[(\lambda/\mu - N(m_\lambda, n_\lambda))^+] \quad \text{by Lemma EC.1} \\
&= \bar{\Pi}_\lambda(m_\lambda, n_\lambda).
\end{aligned}
$$

■

LEMMA EC.3.

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

PROOF. We notice that by Jensen's inequality,

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

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

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

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

Therefore,

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

■

EC.2.2. Theorem 1

PROOF. From Lemma EC.2, we have

$$\begin{aligned}\Pi_\lambda(\bar m_\lambda,\bar n_\lambda)&\le\bar\Pi_\lambda(\bar m_\lambda,\bar n_\lambda)+\mathcal{O}(\sqrt{\lambda})\\&\le\bar\Pi_\lambda(m^*,n^*)+\mathcal{O}(\sqrt{\lambda})\\&\le\Pi_\lambda(m^*,n^*)+\mathcal{O}(\sqrt{\lambda});\end{aligned}$$

moreover, if $\bar n=\Theta(\lambda)$ and $\sigma_n>\sqrt{n}$:

$$\begin{aligned}\Pi_\lambda(\bar m_\lambda,\bar n_\lambda)&\le\bar\Pi_\lambda(\bar m_\lambda,\bar n_\lambda)+\mathcal{O}(\lambda/\sigma_\lambda)\\&\le\bar\Pi_\lambda(m^*,n^*)+\mathcal{O}(\lambda/\sigma_\lambda)\\&\le\Pi_\lambda(m^*,n^*)+\mathcal{O}(\lambda/\sigma_\lambda).\end{aligned}$$

From Lemmas EC.2 & EC.3, we have:

$$\begin{aligned}\Pi_\lambda(\tilde m_\lambda,\tilde n_\lambda)&\le\bar\Pi_\lambda(\tilde m_\lambda,\tilde n_\lambda)+O(\min\{\sqrt{\lambda},\lambda/\sigma_{\tilde n_\lambda}\})\\&\le\tilde\Pi_\lambda(\tilde m_\lambda,\tilde n_\lambda)+O(\min\{\sqrt{\lambda},\lambda/\sigma_{\tilde n_\lambda}\})+O(\sigma_{\tilde n_\lambda})\\&\le\tilde\Pi_\lambda(m^*_\lambda,n^*_\lambda)+O(\min\{\sqrt{\lambda},\lambda/\sigma_{\tilde n_\lambda}\})+O(\sigma_{\tilde n_\lambda})\\&\le\bar\Pi_\lambda(m^*_\lambda,n^*_\lambda)+O(\min\{\sqrt{\lambda},\lambda/\sigma_{\tilde n_\lambda}\})+O(\sigma_{\tilde n_\lambda})\\&\le\Pi_\lambda(m^*_\lambda,n^*_\lambda)+O(\min\{\sqrt{\lambda},\lambda/\sigma_{\tilde n_\lambda}\})+O(\sigma_{\tilde n_\lambda}).\end{aligned}$$

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

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) \leq \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}(\max\{\sqrt{\lambda}, \sigma_\lambda\}).$$

We complete the proof by noting that we must also have:

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) \geq \Pi_\lambda(m^*_\lambda, n^*_\lambda) \text{ and } \Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) \geq \Pi_\lambda(m^*_\lambda, n^*_\lambda).$$

■

EC.3. Asymptotic Accuracy with Time-Varying Demand

Here, we state and prove Theorem EC.1, which quantifies the asymptotic accuracies of the fluid and stochastic-fluid prescriptions with multiple periods, i.e., with time-varying demand. We note that we obtain reduced accuracy in this case (compared to Theorem 1), but that the orders of magnitude of errors in the theorem are only upper bounds.

THEOREM EC.1. *For large λ and time-varying demand,*

$$\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\sqrt{\lambda}\right) + \mathcal{O}(\sigma_\lambda); \tag{EC.1}$$

and

$$\Pi_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) = \Pi_\lambda(m^*_\lambda, n^*_\lambda) + \mathcal{O}\left(\sqrt{\lambda}\right). \tag{EC.2}$$

PROOF. The proof for (EC.1) and (EC.2) follows directly from Lemmas EC.2 and EC.3. ■

EC.4. Proof of Theorem 2

Recall that $\beta \equiv \left(\frac{h}{\theta} + r\right)\mu$.

PROOF. We first note that,

$$\Pi(m_\lambda, n_\lambda) = c_0 m_\lambda + c_1 n_\lambda + \beta \times \begin{cases} 0 & \text{if } \frac{\lambda}{\mu} \leq m_\lambda + n_\lambda - \sigma_{n_\lambda} \text{ (Case 1)} \\ \frac{\lambda}{\mu} - m_\lambda - n_\lambda & \text{if } \frac{\lambda}{\mu} \geq m_\lambda + n_\lambda + \sigma_{n_\lambda} \text{ (Case 2)} \\ \sigma_{n_\lambda} \int_{-1}^{\frac{1}{\sigma_{n_\lambda}}\left(\frac{\lambda}{\mu} - m_\lambda - n_\lambda\right)} F(x)dx & \text{if } -\sigma_{n_\lambda} \leq \frac{\lambda}{\mu} - m_\lambda - n_\lambda \leq \sigma_{n_\lambda} \text{ (Case 3).} \end{cases}$$

It is readily seen that in an optimal solution we must have that $\bar{m}_\lambda \leq \lambda/\mu$. We begin by fixing m_λ and solving for n_λ. To derive the form of the optimal solution, we consider each case separately.

- Case 1: If $\frac{\lambda}{\mu} - m_\lambda \leq n_\lambda - \sigma_{n_\lambda}$, then we see that $\Pi(m_\lambda, n_\lambda) = c_0 m_\lambda + c_1 n_\lambda$ is increasing in n_λ. Thus, in an optimal solution we must have that $m_\lambda + n_\lambda = \lambda/\mu + \sigma_{n_\lambda}$.
- Case 2: If $\frac{\lambda}{\mu} \geq m_\lambda + n_\lambda + \sigma_{n_\lambda}$, then $\Pi(m_\lambda, n_\lambda) = (c_0 - \beta)m_\lambda + (c_1 - \beta)n_\lambda + \beta\lambda/\mu$ is decreasing in n_λ. We must then have at optimum $m_\lambda + n_\lambda = \lambda/\mu - \sigma_{n_\lambda}$.

- Case 3: We first notice that

$$\bar{\Pi}'_\lambda(n_\lambda, m_\lambda) = c_1 - \beta \frac{\sigma_{n_\lambda} + \sigma'_{n_\lambda}(\lambda/\mu - n_\lambda - m_\lambda)}{\sigma_{n_\lambda}} F\left(\frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_{n_\lambda}}\right) + \beta\sigma'_{n_\lambda} \int_{-1}^{\frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_{n_\lambda}}} F(x)dx$$

and

$$\frac{1}{\beta}\bar{\Pi}''_\lambda(n_\lambda) = \sigma''_{n_\lambda} \int_{-1}^{\frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_n}} xf(x)dx + \frac{(\sigma_{n_\lambda} + \sigma'_{n_\lambda}(\lambda/\mu - n_\lambda - m_\lambda))^2}{\sigma^3_{n_\lambda}} f\left(\frac{\lambda/\mu - n_\lambda - m_\lambda}{\sigma_{n_\lambda}}\right) \geq 0$$

Solve for $\bar{\Pi}'_\lambda(n_\lambda, m_\lambda) = 0$, we have $\bar{n}_\lambda$ and $\bar{m}_\lambda$ must satisfy

$$\bar{n}_\lambda + \bar{m}_\lambda = \frac{\lambda}{\mu} - \sigma_{\bar{n}_\lambda} F^{-1}\left(\frac{\sigma_{\bar{n}_\lambda}}{\sigma_{\bar{n}_\lambda} + \sigma'_{\bar{n}_\lambda}(\lambda/\mu - \bar{n}_\lambda - \bar{m}_\lambda)}\left(\frac{c_1}{\beta} - \sigma'_{\bar{n}_\lambda} \int_{-1}^{\frac{\lambda/\mu - \bar{n}_\lambda - \bar{m}_\lambda}{\sigma_{\bar{n}_\lambda}}} F(x)dx\right)\right)$$

Therefore, $\bar{m}_\lambda + \bar{n}_\lambda = \lambda/\mu + \mathcal{O}(\sigma_{\bar{n}_\lambda})$, as desired. Also, note that because of the analysis above it suffices to consider Case 3 where $-\sigma_{n_\lambda} \leq \frac{\lambda}{\mu} - m_\lambda - n_\lambda \leq \sigma_{n_\lambda}$. ■

EC.5. Proof of Lemma 3

Proof. Let $m_\lambda = x\lambda/\mu$, $n_\lambda = y\lambda/\mu$ for $x, y \in \mathbb{R}^+$. Then, optimizing $\bar{\Pi}_\lambda(m_\lambda, n_\lambda)$ is equivalent to optimizing:

$$V(x,y) = \frac{\bar{\Pi}_\lambda(m_\lambda, n_\lambda)}{\lambda/\mu} = c_0 x + c_1 y + \beta\mathbb{E}\left[(1 - x - y - ay\epsilon)^+\right].$$

We denote the optimal solution to $V(x,y)$ as x^*, y^*. As highlighted in the proof of Theorem 2, it suffices to consider x, y such that $-ay \leq 1 - x - y \leq ay$.

That is, we need to solve:

$$\min_{x,y} V(x,y) \tag{EC.3}$$

where

$$V(x,y) = c_0 x + c_1 y + \beta ay \int_{-1}^{\frac{1-x-y}{ay}} F(u)du$$

We first notice that for fixed y, we have

$$\frac{\partial V(x,y)}{\partial x} = c_0 - \beta F\left(\frac{1-x-y}{ay}\right),$$
$$\frac{\partial^2 V(x,y)}{\partial x^2} = \beta \frac{1}{ay} f\left(\frac{1-x-y}{ay}\right) \geq 0 \text{ implying that } V(x,y) \text{ is convex in } x.$$

Let $\alpha_0 \equiv c_0/\beta$. Then, we divide the analysis into two cases:

- **Case a:** If $1 - y - ayF^{-1}(\alpha_0) \geq 0$, $x^*(y) = 1 - y - ayF^{-1}(\alpha_0)$.
- **Case b:** If $1 - y - ayF^{-1}(\alpha_0) < 0$, $x^*(y) = 0$.

In **Case a**, we have $\frac{1-x^*(y)-y}{ay} = F^{-1}(\alpha_0)$. Then, we find y that minimizes:

$$\Xi_1(y) \equiv V(x^*(y), y) = c_0 + \left(c_1 - c_0 - c_0 a F^{-1}(\alpha_0) + a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(u)du \right) y,$$

which is readily seen to be linear in y. Thus, if

$$c_1 > c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(u)du,$$

then $x^* = 1$, $y^* = 0$; otherwise, $x^* = 0$, $y^* = (1 + aF^{-1}(\alpha_0))^{-1}$.

In **Case b**, we find y that minimizes:

$$\Xi_2(y) \equiv V(0, y) = c_1 y + \beta a y \int_{-1}^{1/(ay)-1/a} F(u)du.$$

We notice that

$$\Xi_2'(y) = c_1 + \beta a \int_{-1}^{1/(ay)-1/a} F(x)dx - \frac{\beta}{y} F\left(\frac{1}{ay} - \frac{1}{a}\right)$$

and

$$\Xi_2''(y) = \frac{\beta}{ay^3} f\left(\frac{1}{ay} - \frac{1}{a}\right) > 0 \text{ implying that } V(0, y) \text{ is convex in } y.$$

We also notice that if $\Xi_2'\left(\frac{1}{1+aF^{-1}(\alpha_0)}\right) \geq 0$ then $y^* = \frac{1}{1+aF^{-1}(\alpha_0)}$. Note that $\Xi_2'\left(\frac{1}{1+aF^{-1}(\alpha_0)}\right) \geq 0$ is equivalent to

$$c_1 > c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(u)du, \tag{EC.4}$$

in which case we need to compare $\Xi_1(0)$ to $\Xi_2\left(\frac{1}{1+aF^{-1}(\alpha_0)}\right)$ to find the overall optimal value of $V(x, y)$, assuming that (EC.4) holds. It is readily seen that if (EC.4) holds, then $\Xi_1(0) < \Xi_2\left(\frac{1}{1+aF^{-1}(\alpha_0)}\right)$, so that $y^* = 0$ in this case. Otherwise, $y^* > \frac{1}{1+aF^{-1}(\alpha_0)}$ is the solution of

$$c_1 + \beta a \int_{-1}^{1/(ay)-1/a} F(u)du - \frac{\beta}{y} F\left(\frac{1}{ay} - \frac{1}{a}\right) = 0. \tag{EC.5}$$

In particular, we notice that if

$$\Xi_2'(1) = c_1 + \beta a \int_{-1}^{0} F(u)du - \beta F(0) < 0,$$

then $y^* < 1$; otherwise, $y^* > 1$. Noting that for y^* in (EC.5):

$$\Xi_2(y^*) \leq \Xi_2\left(\frac{1}{1+aF^{-1}(\alpha_0)}\right) = \Xi_1\left(\frac{1}{aF^{-1}(\alpha_0)+1}\right),$$

it must be that the solution to (EC.3) is given by:

- If $c_1 > c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx$, then: $x^* = 1$ and $y^* = 0$;
- Otherwise, $x^* = 0$ and y^* solves (EC.5).

■

EC.6. Proof of Theorem 3

EC.6.1. Case a

If $c_0 < c_1$, then the fluid optimal solution, $\tilde{m}_\lambda$ and $\tilde{n}_\lambda$, as given by Lemma 2 does not involve a flexible capacity. Moreover, it involves matching supply and demand in the period. Thus, we can use the results of Bassamboo and Randhawa (2010) to deduce that letting $\hat{m}_\lambda = \tilde{m}_\lambda$ and $\hat{n}_\lambda = \tilde{n}_\lambda$ yields:

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

EC.6.2. Case b.I

When $c_1 < c_0$, the fluid-optimal solution to $\min_{m,n} \tilde{\Pi}_\lambda(m,n)$ is $\tilde{m}_\lambda = 0$, $\tilde{n}_\lambda = \lambda/\mu$. Letting $\hat{m}_\lambda = \tilde{m}_\lambda$ and $\hat{n}_\lambda = \tilde{n}_\lambda$ yields, making use of the results in Theorem 1, that:

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

EC.6.3. Case b.II

We also consider $c_1 < c_0$. Let $m_\lambda = x\sigma_{\lambda/\mu}$, $n_\lambda = \lambda/\mu + y\sigma_{\lambda/\mu}$ for $x \in \mathbb{R}^+$ and $y \in \mathbb{R}$. Then, optimizing $\bar{\Pi}_\lambda(m,n)$ is equivalent to optimizing:

$$\begin{aligned} C_\lambda(x,y) &= \frac{\bar{\Pi}_\lambda\left(x\sigma_{\lambda/\mu}, \lambda/\mu + y\sigma_{\lambda/\mu}\right) - c_1\lambda/\mu}{\sigma_{\lambda/\mu}} \\ &= c_0 x + c_1 y + \beta\mathbb{E}\left[\left(-(x+y) - \frac{\sigma_{\lambda/\mu + y\sigma_{\lambda/\mu}}}{\sigma_{\lambda/\mu}}\epsilon\right)^+\right] \end{aligned}$$

We denote the optimal solution of $C_\lambda(x,y)$ as x^*_λ, y^*_λ.

Lemma EC.4.

$$\limsup_{\lambda\to\infty} x^*_\lambda < \infty, \quad \limsup_{\lambda\to\infty} y^*_\lambda < \infty \text{ and } \liminf_{\lambda\to\infty} y^*_\lambda > -\infty.$$

Proof. We first notice that $C_\lambda(0,0) = \beta\mathbb{E}[(-\epsilon)^+]$. We next prove the lemma by contradiction. Assume that $\limsup_{\lambda\to\infty} y^*_\lambda = \infty$, then for $M \equiv C_\lambda(0,0)/c_1$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x^*_{\lambda_k} > M$. In this case,

$$C_{\lambda_k}\left(x^*_{\lambda_k}, y^*_{\lambda_k}\right) \geq c_1 y^*_{\lambda_k} > c_1 M = C_\lambda(0,0).$$

Thus, we get a contradiction. So, we must have that $\limsup_{\lambda\to\infty} y^*_\lambda < \infty$.

Now, assume that $\liminf_{\lambda\to\infty} y^*_\lambda = -\infty$. Then, we divide the analysis into two cases.

i) If $\liminf_{\lambda\to\infty} x^*_\lambda + y^*_\lambda = -\infty$, then pick $L_1 \equiv C_\lambda(0,0)/(c_1 - \beta) < 0$. For such L, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$., such that $x^*_{\lambda_k} + y^*_{\lambda_k} < L_1$. In this case,

$$C_{\lambda_k}\left(x^*_{\lambda_k}, y^*_{\lambda_k}\right) > (c_1 - \beta)(x^*_{\lambda_k}, y^*_{\lambda_k}) > (c_1 - \beta)L_1 = C_\lambda(0,0).$$

Thus, we get a contradiction.

ii) If $\liminf_{\lambda\to\infty} x^*_\lambda + y^*_\lambda > -\infty$, then we denote $L_2 = \min\{\liminf_{\lambda\to\infty} x^*_\lambda + y^*_\lambda, 0\}$. As $\liminf_{\lambda\to\infty} y^*_\lambda = -\infty$, we must have $\limsup_{\lambda\to\infty} x^*_\lambda = \infty$. Pick $M_2 \equiv \frac{C_\lambda(0,0)-(c_1-\beta)L_2}{c_0-c_1}$. For such M_2, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x^*_{\lambda_k} > M_2$. In this case:

$$C_{\lambda_k}\left(x^*_{\lambda_k}, y^*_{\lambda_k}\right) > (c_0 - c_1)x^*_{\lambda_k} + (c_1 - \beta)(x^*_{\lambda_k} + y^*_{\lambda_k}) > (c_0 - c_1)M_2 + (c_1 - \beta)L_2 = C_\lambda(0,0).$$

Thus, we get a contradiction.

We have already shown that $\limsup_{\lambda\to\infty} y^*_\lambda < \infty$ and $\liminf_{\lambda\to\infty} y^*_\lambda > -\infty$. There remains to show that we must have $\liminf_{\lambda\to\infty} x^*_\lambda < \infty$. For this, let $L_3 \equiv \min\{\liminf_{\lambda\to\infty} y^*_\lambda, 0\}$. Assume if $\limsup_{\lambda\to\infty} x^*_\lambda = \infty$, then for $M_3 \equiv C_\lambda(0,0)/c_0 - c_1 L_3/c_0$, we can find an infinite sequence $\{\lambda_k : k \geq 0\}$, with $\lambda_k \to \infty$ as $k \to \infty$, such that $x^*_{\lambda_k} > M_3$. In this case,

$$C_{\lambda_k}\left(x^*_{\lambda_k}, y^*_{\lambda_k}\right) \geq c_0 x^*_{\lambda_k} + c_1 y^*_{\lambda_k} > c_0 M_3 + c_1 L_3 = C_\lambda(0,0).$$

Thus, we get a contradiction. Combining the above establishes the lemma. ■

To complete the proof of **Case b.II**, let:

$$\hat{C}(x,y) = c_0 x + c_1 y + \beta\mathbb{E}[(-(x+y) - \epsilon)^+].$$

Let $\hat{x}$ and $\hat{y}$ denote the optimal solution of $\hat{C}(x,y)$. As $c_1 < c_0$, it is readily seen that $\hat{x} = 0$ and $\hat{y} = \gamma^*$ where γ^* denotes the optimal solution to $\min_\gamma c_1\gamma + \beta\mathbb{E}[(-\gamma - \epsilon)^+]$.

Lemma EC.5.

$$C_\lambda(x^*_\lambda, y^*_\lambda) \to C_\lambda(0, \gamma^*) \text{ as } \lambda \to \infty.$$

Proof. As $\sigma_{\lambda/\mu + y\sigma_{\lambda/\mu}}/\sigma_{\lambda/\mu} \to 1$ uniformly on compact sets (u.o.c.) as $\lambda \to \infty$,

$$C_\lambda(x,y) \to \hat{C}(x,y) \text{ u.o.c. as } \lambda \to \infty.$$

Then from Lemma EC.4, we have

$$C_\lambda(x^*_\lambda, y^*_\lambda) \to \hat{C}(x^*, y^*) \text{ as } \lambda \to \infty.$$

■

We are now ready to complete the proof of **Case b.II**. We first notice that $\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) > \Pi^*_\lambda$. We also have that:

$$\begin{aligned}
\Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) &\leq \bar{\Pi}_\lambda(\hat{m}_\lambda, \hat{n}_\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.2} \\
&\leq \bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda) + o(\sigma_\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.5} \\
&\leq \Pi^*_\lambda + o(\sigma_\lambda) + O(\sqrt{\lambda}) \quad \text{by Lemma EC.2} \\
&= \Pi^*_\lambda + o(\sigma_\lambda) \text{ because } \sqrt{\lambda} = o(\sigma_\lambda).
\end{aligned}$$

EC.7. Proof of Lemma 5

We define,

$$\bar{\Gamma}_H(m_\lambda, n_\lambda) \equiv c_0 m_\lambda + c_1 n_\lambda + \beta a n_\lambda \int_{-1}^{(\lambda_H/\mu - m_\lambda - n_\lambda)/(a n_\lambda)} F(x)dx,$$
$$\bar{\Gamma}_L(m_\lambda, n_\lambda, \alpha_\lambda) \equiv c_0 m_\lambda + c_1 \alpha_\lambda n_\lambda + \beta \alpha_\lambda a n_\lambda \int_{-1}^{(\lambda_L/\mu - m_\lambda - \alpha_\lambda n_\lambda)/(a \alpha_\lambda n_\lambda)} F(x)dx,$$

so that $\bar{\Pi}(m_\lambda, n_\lambda, \alpha_\lambda) = T_H \bar{\Gamma}_H(m_\lambda, n_\lambda) + T_L \bar{\Gamma}_L(m_\lambda, n_\lambda, \alpha_\lambda)$.

To derive the optimal solution, we first consider $\bar{\Gamma}_L(m_\lambda, n_\lambda, \alpha_\lambda)$ alone and rely on our previous results from the stationary demand case, i.e., Lemma 3. Recall that $\alpha_0 \equiv c_0/\beta$. From Lemma 3, we know that if $c_1 \geq c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx$, i.e., (EC.4) holds, then we will use the fixed capacity only. That is, we would not use any flexible capacity in the low-demand period and must have $\alpha_\lambda = 0$ in any optimal solution. If $c_1 < c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx$, then we will use the flexible capacity only, i.e., we must have $m_\lambda = 0$ in any optimal solution in this case. Let us now analyze each case separately.

When $c_1 \geq c_0 + c_0 a F^{-1}(\alpha_0) - a\beta \int_{-1}^{F^{-1}(\alpha_0)} F(x)dx$, i.e., (EC.4) holds, we set:

$$m_\lambda = \frac{\lambda_L}{\mu} + x\frac{\lambda_H - \lambda_L}{\mu}, \quad n_\lambda = y\frac{\lambda_H - \lambda_L}{\mu} \quad \text{and} \quad \alpha_\lambda = 0 \quad \text{for} \quad x, y \geq 0.$$

Then, we can equivalently minimize:

$$\frac{\bar{\Pi}_\lambda(m_\lambda, n_\lambda, 0) - c_0(T_H + T_L)\lambda_L/\mu - c_1 T_H(\lambda_H - \lambda_L)/\mu}{(\lambda_H - \lambda_L)/\mu}$$
$$= c_0(T_H + T_L)x + c_1 T_H y + \beta T_H \mathbb{E}[(1 - x - y - ay\epsilon)^+];$$

note that this is the same objective as in (EC.3), with modified costs. Let

$$\beta_1 \equiv \frac{c_0(T_H + T_L)}{\beta T_H} = \alpha_0 \frac{T_H + T_L}{T_H} > \alpha_0.$$

Then, again from Lemma 3, we have the following two subcases:

If $c_1 \geq c_0 + c_0 a F^{-1}(\beta_1) - a\beta \int_{-1}^{F^{-1}(\beta_1)} F(x)dx$, then we shall set $x^* = 1$ and $y^* = 0$. In this case,

$$\hat{m}_\lambda = \lambda_H/\mu, \quad \hat{n}_\lambda = 0, \quad \text{and} \quad \hat{\alpha}_\lambda = 0.$$

If $c_1 < c_0 + c_0 a F^{-1}(\beta_1) - a\beta \int_{-1}^{F^{-1}(\beta_1)} F(x)dx$, then we shall set $x^* = 0$ and y^* solves

$$\min_y c_1 y + \beta \mathbb{E}[(1 - y - ay\epsilon)^+],$$

or equivalently, y^* solves

$$c_1 + \beta a \int_{-1}^{1/(ay) - 1/a} F(x)dx - \beta \frac{1}{y} F\left(\frac{1}{ay} - \frac{1}{a}\right) = 0 \quad \text{so that}$$

$$\hat{m}_\lambda = \lambda_L/\mu, \quad \hat{n}_\lambda = y^*(\lambda_H - \lambda_L)/\mu, \quad \text{and} \quad \hat{\alpha}_\lambda = 0.$$

When $c_1 < c_0 + c_0 a F^{-1}(\alpha_0) - a(h+r\theta)\frac{\mu}{\theta}\int_{-1}^{F^{-1}(\alpha_0)} F(x)dx$, i.e., (EC.4) does not hold, we shall only use the flexible capacity. In this case,

$$\bar{\Gamma}_H(0, n_\lambda) = c_1 n_\lambda + \beta\mathbb{E}[(\lambda_H/\mu - n_\lambda - an\epsilon)^+].$$

And, recalling that $\lambda_L = \xi_L\lambda$ and $\lambda_H = \xi_H\lambda$, we can also write:

$$\begin{aligned}
\bar{\Gamma}_L(0, n_\lambda, \alpha_\lambda) &= c_1\alpha_\lambda n_\lambda + \beta\mathbb{E}[(\lambda_L/\mu - \alpha_\lambda n_\lambda - a\alpha_\lambda n_\lambda\epsilon)^+] \\
&= \frac{\xi_L}{\xi_H}\left(c_1\frac{\alpha_\lambda}{\xi_L/\xi_H}n_\lambda + \beta\mathbb{E}\left[\left(\lambda_H/\mu - \frac{\alpha_\lambda}{\xi_L/\xi_H}n_\lambda - a\frac{\alpha_\lambda}{\xi_L/\xi_H}n_\lambda\epsilon\right)^+\right]\right) \\
&= \frac{\xi_L}{\xi_H}\bar{\Gamma}_H\left(0, \frac{\alpha_\lambda}{\xi_L/\xi_H}n_\lambda\right),
\end{aligned}$$

Thus, if $\hat{n}_\lambda$ minimizes $\bar{\Gamma}_H(0, n_\lambda)$, then $(\hat{n}_\lambda, \xi_L/\xi_H)$ minimizes $\bar{\Gamma}_L(0, n_\lambda, \alpha_\lambda)$. Therefore, in this case, we set $\hat{m}_\lambda = 0$, $\hat{n}_\lambda = y^*\lambda_H/\mu$, where y^* solves:

$$c_1 + \beta a\int_{-1}^{(1/(ay)-1/a} F(x)dx - (h+r\theta)\frac{\mu}{\theta}\frac{1}{y}F\left(\frac{1}{ay} - \frac{1}{a}\right) = 0,$$

and $\hat{\alpha}_\lambda = \xi_L/\xi_H$.

EC.8. Proof of Theorem 4

Let $(m^*_\lambda, n^*_\lambda, \alpha^*_\lambda)$ be an optimal solution to $\min_{m_\lambda, n_\lambda, \alpha_\lambda}\Pi_\lambda(m_\lambda, n_\lambda, \alpha_\lambda)$. We also write $(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda)$ as an optimal solution to $\min_{m_\lambda, n_\lambda, \alpha_\lambda}\bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda)$.

EC.8.1. Case a

By Lemma EC.2 and Lemma EC.3, we have:

$$\begin{aligned}
\Pi_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda) &\le \tilde{\Pi}_\lambda(\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda) + \mathcal{O}(\sqrt{\lambda}) \\
&\le \tilde{\Pi}_\lambda(m^*_\lambda, n^*_\lambda, \alpha^*_\lambda) + \mathcal{O}(\sqrt{\lambda}) \\
&\le \Pi_\lambda(m^*_\lambda, n^*_\lambda, \alpha^*_\lambda) + \mathcal{O}(\sqrt{\lambda}) \\
&= \Pi^*_\lambda + \mathcal{O}(\sqrt{\lambda}).
\end{aligned}$$

EC.8.2. Case b.I

Assume that $c_0 \le \frac{T_H}{T_H+T_L}c_1$**.** As for Theorem 3, we explore the appropriate safety-capacity hedge that should be "added" to the corresponding fluid approximation to obtain improved asymptotic accuracy. To this aim, let:

$$\begin{aligned}
\bar{\Gamma}_H(m_\lambda, n_\lambda) &= c_0 m_\lambda + c_1 n_\lambda + \beta\mathbb{E}[(\lambda_H/\mu - m_\lambda - N(n_\lambda))^+] \\
\bar{\Gamma}_L(m_\lambda, n_\lambda, \alpha_\lambda) &= c_0 m_\lambda + c_1\alpha_\lambda n_\lambda + \beta\mathbb{E}[(\lambda_L/\mu - m_\lambda - N(\alpha_\lambda n_\lambda))^+]
\end{aligned}$$

So that,

$$\bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda) = T_H \bar{\Gamma}_H(m_\lambda, n_\lambda, \alpha_\lambda) + T_L \bar{\Gamma}_L(m_\lambda, n_\lambda, \alpha_\lambda).$$

We first notice from Lemma EC.3 that for any $(m_\lambda, n_\lambda, \alpha_\lambda)$:

$$\tilde{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda) \le \bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda),$$

and for $n_\lambda = 0$:

$$\tilde{\Pi}_\lambda(m_\lambda, 0, \alpha_\lambda) = \bar{\Pi}_\lambda(m_\lambda, 0, \alpha_\lambda).$$

We next show that we must have that $\bar{n}_\lambda = 0$. We prove this by contradiction. Assume that $\bar{n}_\lambda > 0$, then for $\bar{m}_\lambda, \bar{\alpha}_\lambda$:

$$\begin{aligned}
\bar{\Pi}_\lambda(\bar{m}_\lambda + \bar{n}_\lambda, 0, \bar{\alpha}_\lambda) &= \tilde{\Pi}_\lambda(\bar{m}_\lambda + \bar{n}_\lambda, 0, \bar{\alpha}_\lambda) \\
&< \tilde{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) \text{ because } c_0 \le \tfrac{T_H}{T_H+T_L} c_1 \text{ by assumption} \\
&\le \bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) \text{ by Lemma EC.3.}
\end{aligned}$$

Thus, we get a contradiction. Since $\bar{n}_\lambda = 0$ in an optimal solution, it must be that the fluid-optimal and stochastic-fluid-optimal solutions coincide. In other words,

$$(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) \equiv (\tilde{m}_\lambda, \tilde{n}_\lambda, \tilde{\alpha}_\lambda).$$

From the solution to the fluid optimization problem in Lemma 4, we can set

$$\hat{m}_\lambda = \lambda_H/\mu, \quad \hat{n}_\lambda = 0, \quad \text{and} \quad \hat{\alpha}_\lambda = 0,$$

which are the (exact) optimal solutions to the stochastic-fluid optimization problem.

EC.8.3. Case b.II

Assume that $\frac{T_H}{T_H+T_L} c_1 < c_0 < c_1$.

We first show that $\bar{m}_\lambda \ge \lambda_L/\mu$ and $\bar{\alpha}_\lambda = 0$. We prove this by contradiction. Suppose that $\bar{m}_\lambda < \lambda_L/\mu$ or $\bar{\alpha}_\lambda > 0$. We consider two subcases: **(i)** $\bar{m}_\lambda + \bar{n}_\lambda < \lambda_L/\mu$, and **(ii)** $\bar{m}_\lambda + \bar{n}_\lambda \ge \lambda_L/\mu$.

In **case (i)**: $\tilde{\Gamma}_L(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) > \tilde{\Gamma}_L(\bar{m}_\lambda, \bar{n}_\lambda, 1)$ by the linearity of the objective (since the period is understaffed). But, $\tilde{\Gamma}(\bar{m}_\lambda, \bar{n}_\lambda, 1) > \tilde{\Gamma}(\bar{m}_\lambda + \bar{n}_\lambda, 0, 0) = \bar{\Gamma}(\bar{m}_\lambda + \bar{n}_\lambda, 0, 0)$ since $c_0 < c_1$. Thus, we reach a contradiction, and **case (i)** cannot happen.

We now focus on **case (ii)**. Set $m'_\lambda = \lambda_L/\mu$, $n'_\lambda = (\bar{n}_\lambda + \bar{m}_\lambda - \lambda_L/\mu)$, and $\alpha'_\lambda = 0$. We notice that $\bar{m}_\lambda + \bar{n}_\lambda = m'_\lambda + n'_\lambda$ and $\bar{n}_\lambda > n'$. For any fixed values of λ, μ, and s, we have that $\mathbb{E}[(\lambda/\mu - s - \sigma_n \epsilon)^+]$ is increasing in n. Thus,

$$\bar{\Gamma}_H(\bar{m}_\lambda, \bar{n}_\lambda) \ge \bar{\Gamma}_H(m'_\lambda, n'_\lambda).$$

We also notice that $\tilde{\Gamma}_L(m, n, \alpha)$ is minimized at $m' = \lambda_L/\mu$, $\alpha' = 0$, by Lemma 4, i.e.,

$$\bar{\Gamma}_L(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) \ge \tilde{\Gamma}_L(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) \ge \tilde{\Gamma}_L(m', n', 0) = \bar{\Gamma}_L(m', n', 0).$$

Thus, $\bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) \ge \bar{\Pi}_\lambda(m', n', \alpha')$, and we get a contradiction. So, **case (ii)** cannot happen. We conclude that we must have $\bar{m}_\lambda \ge \lambda_L/\mu$ and $\bar{\alpha}_\lambda = 0$.

Asymptotic accuracy. Now, let $m_\lambda = \lambda_L/\mu + x\sigma_{\lambda_H/\mu-\lambda_L/\mu}$ and $n_\lambda = (\lambda_H/\mu - \lambda_L/\mu) + y\sigma_{\lambda_H/\mu-\lambda_L/\mu}$ where $x \geq 0$ (based on the analysis above) and $y \in \mathbb{R}$. Recall also that we must have $\bar{\alpha}_\lambda = 0$. Then, we set:

$$\begin{aligned} C_{2,\lambda}(x,y) &\equiv \frac{\bar{\Pi}_\lambda(m_\lambda, n_\lambda, 0) - c_0(T_H+T_L)\frac{\lambda_L}{\mu} - c_1 T_L \frac{\lambda_H-\lambda_L}{\mu}}{\sigma_{\lambda_H/\mu-\lambda_L/\mu}} \\ &= c_0(T_H+T_L)x + c_1 T_H y \\ &\quad + \beta T_H \mathbb{E}\left[\left(-x-y-\frac{\sigma_{(\lambda_H/\mu-\lambda_L/\mu)+y\sigma_{\lambda_H/\mu-\lambda_L/\mu}}}{\sigma_{\lambda_H/\mu-\lambda_L/\mu}}\epsilon\right)^+\right] \end{aligned}$$

Let x^*_λ, y^*_λ denote the optimal solution to

$$\min_{x\geq 0, y\in\mathbb{R}} C_{2,\lambda}(x,y). \tag{EC.6}$$

We also set:

$$\hat{C}_2(x,y) \equiv c_0(T_H+T_L)x + c_1 T_H y + \beta T_H \mathbb{E}\left[(-x-y-\epsilon)^+\right],$$

and denote x^*, y^* as the optimal solution to $\min_{x\geq 0, y\in\mathbb{R}} \hat{C}_2(x,y)$. For $c_0 < c_1 < \frac{T_H+T_L}{T_H}c_0$, we have $x^* = 0$ and y^* solves $\hat{H}_2(y) \equiv c_1 y + \beta\mathbb{E}\left[(-y-\epsilon)^+\right]$.

Lemma EC.6.

$$C_{2,\lambda}(x^*_\lambda, y^*_\lambda) \to \hat{C}_2(x^*, y^*) \text{ as } \lambda \to \infty.$$

Proof. The proof Lemma EC.6 follows exactly the same lines as Lemma EC.5. We shall omit it here. ■

Now, for $\hat{m}_\lambda = \lambda_L/\mu$, $\hat{n}_\lambda = (\lambda_H/\mu - \lambda_L/\mu) + y^*\sigma_{\lambda_H/\mu-\lambda_L/\mu}$, where y^* is the solution to (EC.6), and $\hat{\alpha}_\lambda = 0$, we have:

$$\begin{aligned} \Pi_\lambda(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda) &\leq \bar{\Pi}_\lambda(\hat{m}_\lambda, \hat{n}_\lambda, \hat{\alpha}_\lambda) + \mathcal{O}(\sqrt{\lambda}) \text{ by Lemma EC.2} \\ &\leq \bar{\Pi}_\lambda(\bar{m}_\lambda, \bar{n}_\lambda, \bar{\alpha}_\lambda) + o(\sigma_\lambda) + \mathcal{O}(\sqrt{\lambda}) \text{ by Lemma EC.6} \\ &\leq \Pi^*_\lambda + o(\sigma_\lambda), \end{aligned}$$

which concludes the proof for this case.

EC.8.4. Case b.III

Assume that $c_0 \geq c_1$. Let $m_\lambda = x\sigma_{\lambda_H/\mu}$, $n_\lambda = \frac{\lambda_H}{\mu} + y\sigma_{\lambda_H/\mu}$, and $\alpha_\lambda = \frac{\lambda_L}{\lambda_H} + z\frac{\sigma_{\lambda_H/\mu}}{\lambda_H/\mu}$.

$$\begin{aligned} &\frac{\bar{\Pi}_\lambda(m_\lambda, n_\lambda, \alpha_\lambda) - c_1 T_H \lambda_H/\mu - c_1 T_L \lambda_L/\mu}{\sigma_{\lambda_H/\mu}} \\ &= T_H\left(c_0 x + c_1 y + \beta\mathbb{E}\left[\left(-x-y-\frac{\sigma_{\lambda_H/\mu+y\sigma_{\lambda_H/\mu}}}{\sigma_{\lambda_H/\mu}}\epsilon\right)^+\right]\right) \end{aligned}$$

$$
\begin{aligned}
&+T_L\left(c_0x+c_1\left(\frac{\xi_L}{\xi_H}y+z\right)+c_1yz\frac{\sigma_{\lambda_H/\mu}}{\lambda_H/\mu}+\beta\right.\\
&\left.\times\mathbb{E}\left[\left(-x-\frac{\xi_L}{\xi_H}y-z-yz\frac{\sigma_{\lambda_H/\mu}}{\lambda_H/\mu}-\frac{\sigma_{\xi_L/\xi_H(\lambda_H/\mu+y\sigma_{\lambda_H/\mu})+z\sigma_{\lambda_H/\mu}+yz\sigma^2_{\lambda_H/\mu}/(\lambda_H/\mu)}}{\sigma_{\lambda_H/\mu}}\epsilon\right)^+\right]\right)\\
&\equiv C_{3,\lambda}(x,y,z).
\end{aligned}
$$

Let x^*_λ, y^*_λ, z^*_λ denote the optimal solution to $\min_{x\geq 0,y\in\mathbb{R},z\in\mathbb{R}} C_{3,\lambda}(x,y,z)$. We also define

$$
\begin{aligned}
\hat{C}_3(x,y,z)\equiv\ & T_H\left(c_0x+c_1y+\beta\mathbb{E}\left[(-x-y-\epsilon)^+\right]\right)\\
&+T_L\left(c_0x+c_1\left(\frac{\xi_L}{\xi_H}y+z\right)+\beta\mathbb{E}\left[\left(-x-\frac{\xi_L}{\xi_H}y-z-\left(\frac{\xi_L}{\xi_H}\right)^q\epsilon\right)^+\right]\right)
\end{aligned}
$$

Let x^*, y^*, z^* denote the optimal solution to $\min_{x,y,z}\hat{C}_3(x,y,z)$. For $c_0>c_1$, we have $x^*=0$, and y^* and z^* solve:

$$
\min_{y,z} c_1\left(T_H+T_L\frac{\xi_L}{\xi_H}\right)y+c_1T_Lz+\beta T_H\mathbb{E}\left[(-y-\epsilon)^+\right]+\beta T_L\mathbb{E}\left[\left(-\frac{\xi_L}{\xi_H}y-z-\left(\frac{\xi_L}{\xi_H}\right)^q\epsilon\right)^+\right].
$$

Lemma EC.7.

$$
C_{3,\lambda}(x^*_\lambda,y^*_\lambda,z^*_\lambda)\to\hat{C}_3(x^*,y^*,z^*)\ \textit{as}\ \lambda\to\infty.
$$

Proof. The proof Lemma EC.6 follows exactly the same line of analysis as Lemma EC.5. We shall omit it here. ■

Now let $\hat{m}_\lambda=0$, $\hat{n}_\lambda=\lambda_H/\mu+y^*\sigma_{\lambda_H/\mu}$, and $\hat{\alpha}_\lambda=\frac{\xi_L}{\xi_H}+z^*\sigma_{\lambda_H/\mu}/(\lambda_H/\mu)$. Then, we have:

$$
\begin{aligned}
\Pi_\lambda(\hat{m}_\lambda,\hat{n}_\lambda,\hat{\alpha}_\lambda)&\leq\bar{\Pi}_\lambda(\hat{m}_\lambda,\hat{n}_\lambda,\hat{\alpha}_\lambda)+\mathcal{O}(\sqrt{\lambda})\text{ by Lemma EC.2}\\
&\leq\bar{\Pi}_\lambda(\bar{m}_\lambda,\bar{n}_\lambda,\bar{\alpha}_\lambda)+o(\sigma_\lambda)+\mathcal{O}(\sqrt{\lambda})\text{ by Lemma EC.6}\\
&\leq\Pi^*_\lambda+o(\sigma_\lambda),
\end{aligned}
$$

which concludes the proof of this case.

EC.9. Proof of Theorem 5

Our proof proceeds along similar lines as Theorem 1, so we will be brief. First, we establish the following lemma, paralleling Lemma EC.1.

Lemma EC.8. *When $\theta=\mu$,*

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

Furthermore, if $p>q$, $p>1/2$, *and* $m_\lambda+n_\lambda=\mathcal{O}(\lambda)$,

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

If $q>p$, $q>1/2$, *and* $n_\lambda=\Theta(\lambda)$,

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

Proof. When $\theta=\mu$, $[X(\Lambda_\lambda,N(m_\lambda,n_\lambda))|\Lambda_\lambda=\xi]\sim\text{Poisson}(\xi/\mu)$. We first notice that:

$$\begin{aligned}(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+&\le\mathbb{E}\left[(X(\Lambda_\lambda,m_\lambda,n_\lambda)-N(m_\lambda,n_\lambda))^+|\Lambda_\lambda,N(m_\lambda,n_\lambda)\right]\\&\le(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+\sqrt{\frac{4\pi\Lambda_\lambda}{\mu}}\exp\left(-\frac{\mu}{4\Lambda_\lambda}\left(\frac{\Lambda_\lambda}{\mu}-N(m_\lambda,n_\lambda)\right)^2\right)\\&\quad+1/\log 2,\qquad\text{(EC.7)}\\&\le(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+\sqrt{\frac{4\pi\Lambda_\lambda}{\mu}}+1/\log 2.\qquad\text{(EC.8)}\end{aligned}$$

Then, we have:

$$\mathbb{E}\left[(X(\Lambda_\lambda,N(m_\lambda,n_\lambda))-N(m_\lambda,n_\lambda))^+\right]\ge\mathbb{E}\left[(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+\right],$$

and, from (EC.8), we have that:

$$\begin{aligned}&\mathbb{E}\left[(X(\Lambda_\lambda,m_\lambda,n_\lambda)-N(m_\lambda,n_\lambda))^+\right]\\&\le\mathbb{E}\left[(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+\right]+\mathbb{E}\left[\sqrt{\frac{4\pi\Lambda_\lambda}{\mu}}\right]+1/\log 2\\&\le\mathbb{E}\left[(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+\right]+\sqrt{\frac{4\pi\mathbb{E}[\Lambda_\lambda]}{\mu}}+1/\log 2\quad\text{by Jensen's inequality}\\&=\mathbb{E}\left[(\Lambda_\lambda-N(m_\lambda,n_\lambda))^+\right]+\mathcal{O}(\sqrt{\lambda}).\end{aligned}$$

We next analyze some refined upper bounds based on (EC.7) for some special cases. In what follows, we let $f_N^\lambda(s)$ denote the pdf of $N(m_\lambda,n_\lambda)$ and $g_\Lambda^\lambda(s)$ denote the pdf of Λ.

When $p>q$, $p>1/2$ and $m_\lambda+n_\lambda=\mathcal{O}(\lambda)$:

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

Let $H_\lambda(y) = \sup_{y-\sqrt{\lambda}\log\lambda<\xi<y+\sqrt{\lambda}\log\lambda} \lambda g_\Lambda^\lambda(\xi)$. We notice that $H_\lambda(y) = \mathcal{O}(\lambda/\eta_\lambda)$ and

$$\begin{aligned}
&\int_{s\mu-\sqrt{\lambda}\log\lambda}^{s\mu+\sqrt{\lambda}\log\lambda} \sqrt{\frac{4\pi\xi}{\mu}} \exp\left(-\frac{\mu}{4\xi}\left(\frac{\xi}{\mu}-s\right)^2\right) g_\Lambda^\lambda(\xi)d\xi \\
&\leq H_\lambda(s\mu) \int_{\lambda-\sqrt{\lambda}\log\lambda}^{\lambda+\sqrt{\lambda}\log\lambda} \sqrt{\frac{4\pi}{\mu}} \frac{\sqrt{s\mu+\sqrt{\lambda}\log\lambda}}{\lambda} \exp\left(-\frac{\mu}{4\left(s\mu+\sqrt{\lambda}\log\lambda\right)}\left(\frac{\xi}{\mu}-s\right)^2\right) d\xi \\
&\leq H_\lambda(s\mu) \int_{\lambda-\sqrt{\lambda}\log\lambda}^{\lambda+\sqrt{\lambda}\log\lambda} \frac{K_1}{\sqrt{\lambda}} \exp\left(-\frac{K_2}{\lambda}(\xi-s\mu)^2\right) d\xi \text{ for some } K_1, K_2 > 0, \text{ as } s = \mathcal{O}(\lambda) \\
&= \mathcal{O}(\lambda/\eta_\lambda).
\end{aligned}$$

In addition,

$$\begin{aligned}
&\int_{s\mu+\sqrt{\lambda}\log\lambda}^{\infty} \sqrt{\frac{4\pi\xi}{\mu}} \exp\left(-\frac{\mu}{4\xi}\left(\frac{\xi}{\mu}-s\right)^2\right) g_\Lambda^\lambda(\xi)d\xi \\
&\leq \sqrt{\frac{4\pi(s+\sqrt{\lambda}\log(\lambda))}{\mu}} \exp\left(-\frac{\mu}{4(s+\sqrt{\lambda}\log(\lambda))}\lambda(\log\lambda)^2\right) \\
&= o(1) \quad \text{as } s = \mathcal{O}(\lambda).
\end{aligned}$$

Similarly,

$$\int_{-\infty}^{s\mu-\sqrt{\lambda}\log\lambda} \sqrt{\frac{4\pi\xi}{\mu}} \exp\left(-\frac{\mu}{4\xi}\left(\frac{\xi}{\mu}-s\right)^2\right) g_\Lambda^\lambda(\xi)d\xi = o(1).$$

Thus,

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

When $q > p$, $q > 1/2$ and $n = \Theta(\lambda)$:

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

Let $M_N^\lambda(y) = \sup_{y-\sqrt{\lambda}\log\lambda<\xi<y+\sqrt{\lambda}\log\lambda} \lambda f_N^\lambda(s)$. We notice that, when $n_\lambda = \Theta(\lambda)$, we have that: $M_N^\lambda(y) = \mathcal{O}(\lambda/\sigma_\lambda)$ and:

$$\int_{\xi/\mu-\sqrt{\lambda}\log\lambda}^{\xi/\mu+\sqrt{\lambda}\log\lambda} \sqrt{\frac{4\pi\xi}{\mu}} \exp\left(-\frac{\mu}{4\xi}\left(\frac{\xi}{\mu}-s\right)^2\right) f_N^\lambda(s)ds$$

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

In addition,

$$
\begin{aligned}
&\int_{\xi/\mu+\sqrt{\lambda}\log\lambda}^{\infty}\sqrt{\frac{4\pi\xi}{\mu}}\exp\left(-\frac{\mu}{4\xi}\left(\frac{\xi}{\mu}-s\right)^2\right)f_{m,n}(s)ds\\
&\leq\sqrt{\frac{4\pi\xi}{\mu}}\exp\left(-\frac{\mu}{4\xi}\lambda(\log\lambda)^2\right)\\
&=o(1)\quad \text{as } \xi=\mathcal{O}(\lambda).
\end{aligned}
$$

Similarly,

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

Thus,

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

■

To complete the proof of the theorem, we prove analogs of Lemmas EC.2 and EC.3. Since the proofs for these lemmas proceeds along almost the same lines as those previous lemmas, we omit the relevant details.

EC.10. Proof of Theorem 6

The proof for **Case a** follows from the analysis in Bassamboo et al. (2010), since we know that $\bar{m}=0$ if $c_0<c_1$, so that the problem reduces to one with only randomness in arrivals. We next prove **Case b**. The proof proceeds along similar lines as Case b in Theorem 4, so we will be brief.

EC.10.1. Case b.I

Let $m_\lambda=x\eta_\lambda$, $n_\lambda=\lambda/\mu+y\eta_\lambda$, and define:

$$
U_\lambda(x,y)\equiv\frac{\bar{\Pi}_\lambda(m_\lambda,n_\lambda)-c_1\lambda/\mu}{\eta_\lambda}=c_0x+c_1y+\beta\mathbb{E}\left[\left(\frac{\delta}{\mu}-x-y-\frac{\sigma_{\lambda/\mu+y\eta_\lambda}}{\eta_\lambda}\epsilon\right)^+\right].
$$

We denote the optimal solution to $\min_{x\geq0,y}U_\lambda(x,y)$ as $(x_\lambda^*,y_\lambda^*)$. We also write:

$$
\hat{U}(x,y)=c_0x+c_1y+\beta\mathbb{E}\left[(\delta/\mu-x-y)^+\right],
$$

and let (x^*, y^*) be the optimal solution to $\min_{x\geq 0,y} \hat{U}(x,y)$. When $c_1 < c_0$, we obtain that $x^* = 0$ and that y^* solves:

$$\min_y c_1 y + \beta \mathbb{E}\left[(\delta/\mu - y)^+\right].$$

We can also establish the following lemma, proceeding as before.

LEMMA EC.9.

$$U_\lambda\left(x_\lambda^*, y_\lambda^*\right) \to \hat{U}(x^*, y^*) \text{ as } \lambda \to \infty.$$

EC.10.2. Case b.II

Let $m = x\sigma_{\lambda/\mu}$, $n = \lambda/\mu + y\sigma_{\lambda/\mu}$, and

$$V_\lambda(x,y) \equiv \frac{\bar{\Pi}_\lambda(m_\lambda, n_\lambda) - c_1\lambda/\mu}{\sigma_{\lambda/\mu}} = c_0 x + c_1 y + \beta \mathbb{E}\left[\left(\frac{\eta_\lambda}{\sigma_{\lambda/\mu}\mu}\delta - x - y - \frac{\sigma_{\lambda/\mu + y\sigma_{\lambda/\mu}}}{\sigma_{\lambda/\mu}}\epsilon\right)^+\right].$$

We denote the optimal solution to $\min_{x\geq 0,y} V_\lambda(x,y)$ as $(x_\lambda^*, y_\lambda^*)$. We also write:

$$\hat{V}(x,y) = c_0 x + c_1 y + \beta \mathbb{E}\left[(-x - y - \epsilon)^+\right]$$

and let (x^*, y^*) be the optimal solution to $\min_{x\geq 0,y} \hat{V}(x,y)$. When $c_1 < c_0$, we obtain that $x^* = 0$ and y^* solves:

$$\min_y c_1 y + \beta \mathbb{E}\left[(-y - \epsilon)^+\right].$$

We can then establish the following lemma.

LEMMA EC.10.

$$V_\lambda\left(x_\lambda^*, y_\lambda^*\right) \to \hat{V}(x^*, y^*) \text{ as } \lambda \to \infty.$$

EC.10.3. Case b.III

Let $m = x\sigma_{\lambda/\mu}$, $n = \lambda/\mu + y\sigma_{\lambda/\mu}$, and

$$W_\lambda(x,y) \equiv \frac{\bar{\Pi}_\lambda(m_\lambda, n_\lambda) - c_1\lambda/\mu}{\sigma_{\lambda/\mu}} = c_0 x + c_1 y + \beta \mathbb{E}\left[\left(\frac{\eta_\lambda}{\sigma_{\lambda/\mu}\mu}\delta - x - y - \frac{\sigma_{\lambda/\mu + y\sigma_{\lambda/\mu}}}{\sigma_{\lambda/\mu}}\epsilon\right)^+\right]$$

We denote the optimal solution to $W_\lambda(x,y)$ as $(x_\lambda^*, y_\lambda^*)$. We also write

$$\hat{W}(x,y) = c_0 x + c_1 y + \beta \mathbb{E}\left[(a\delta/\mu - x - y - \epsilon)^+\right],$$

and let (x^*, y^*) be the optimal solution to $\min_{x\geq 0,y} \hat{W}(x,y)$. We get that $x^* = 0$ and y^* solves

$$\min_y c_1 y + \beta \mathbb{E}\left[(a\delta/\mu - y - \epsilon)^+\right].$$

We can then establish the following lemma.

LEMMA EC.11.

$$W_\lambda\left(x_\lambda^*, y_\lambda^*\right) \to \hat{W}(x^*, y^*) \text{ as } \lambda \to \infty.$$