Managing Queueing Systems Where Capacity Is Random and Customers Are Impatient

Rouba Ibrahim

UCL School of Management, 1 Canada Square, London E14 5AB, U.K.
rouba.ibrahim@ucl.ac.uk

August 6, 2017

Abstract. One prevalent assumption in queueing theory is that the number of servers in a queueing model is deterministic. However, randomness in the number of available servers often arises in practice, e.g., in virtual call centers where agents are allowed to set their own schedules. In this paper, we study the problems of staffing and controlling queueing systems with an uncertain number of servers and impatient customers. Because randomness in the number of servers creates congestion in the system, the customer abandonment distribution plays an important role. We characterize how it affects both the optimal staffing policy and the cost incurred by the manager. Because of that strong dependence on the abandonment distribution, it is natural to investigate ways of controlling customer abandonment behavior so as to mitigate that cost. Here, we propose doing so by making delay announcements in the system. We characterise how the manager may use three controls in her toolbox, staffing, compensation, and the announcements, to effectively control her system. We show that despite jointly optimizing the usage of those three controls, it may be cost effective for the manager to understaff, overstaff, or match supply and demand in any given shift.

Keywords: delay announcements; many-server queues; random capacity; abandonment.

1 Introduction

There is a broad literature in queueing theory which studies the problems of staffing and controlling large-scale service systems; e.g., for surveys of applications in call-center management, see Gans et al. (2003) and Akşin et al. (2007). Much of that body of research formulates recommendations based on queueing models with several realistic features, such as time-varying parameters and non-standard network structures. However, one prevalent assumption in those models is that the number of servers is deterministic. As such, the realized staffing level in any given time period is assumed to be equal to the planned staffing level for that period. In contrast, this paper studies optimal staffing and control decisions in queueing systems with a random number of servers instead.

Uncertainty in the number of available agents arises in many novel work arrangements. Virtual call centers, such as Liveops *(liveops.com)* or Arise *(arise.com)* hire work-from-home agents who are free to set their own schedules, often at very short time notice. Amazon Flex *(flex.amazon.com)* relies on independent contractors to deliver Amazon Prime Now packages, which have a short delivery deadline, usually 1-2 hours. Those delivery workers enjoy the flexibility of setting their preferred delivery times. Ride-sharing services, such as Uber *(Uber.com)* or Lyft *(lyft.com)*, also allow their drivers to self-schedule. They use "surge pricing" (Uber 2015) to ensure the participation of a sufficient number of drivers in different time periods. Uncertainty in the number of agents also arises in traditional work environments when there is significant non-adherence to planned schedules. For example, it is well known that nurse absenteeism is a considerable problem in healthcare settings (US Bureau of Labor Statistics 2008, BBC 2015). As a result, the actual number of available nurses, in a given shift, is uncertain. Agent absenteeism also remains one of the leading causes of poor customer service levels in brick-and-mortar call centers (TalkDesk 2014). While each of those settings poses unique operational challenges, agents may be viewed as being strategic in each. That is, they are decision makers who choose whether or not to be available for work in a given shift based on their individual preferences or availabilities. We study the operational management of such systems.

Framework. We assume that there are k working shifts, and that agents have inherent, heterogenous, availabilities or preferences for different shifts. We assume that there is a staffing cost c_j per server, depending on the shift j. In a first stage, the system manager decides on the total staffing level, n. Each agent in the pool of size n has a fixed probability, r_j, of showing up to work in the second stage. The show-up probability r_j depends on both the personal preferences of agents and the compensation offered in shift j: An agent shows up to work in a shift if her (random) opportunity cost is less than or equal to the compensation offered for that shift. Customers are assumed to be both impatient and delay sensitive, as is usually the case in practice.

A long-term staffing decision. Virtual call-center platforms, such as Arise or LiveOps, routinely provide training services to agents before matching them with client companies. These training periods typically last a few weeks (up to 10 weeks) [1]. With ride-sharing platforms, such as Uber or Lyft, training and background checks also require advance planning, and typically last around 2 weeks [2]. Thus, staffing decisions in systems with self-scheduling agents cannot, usually, be made "on the fly". Since the agent population is both remote and large, up to hundreds of agents, system managers cannot simply solicit their agents' scheduling preferences ahead of time. For example, hiring decisions in virtual call centers often do not even involve a face-to-face interview [3].

To mimic such practical challenges, we first consider a setting where the manager must decide on n in advance by relying on historical estimates of r_j. The manager may obtain these, e.g., by analyzing human resources data in her firm. For example, based on analyzing that data, she may know that work-from-home parents usually prefer morning shifts while children are at school; however, she would not know whether any specific work-from-home parent will be available for a particular shift. We determine n that minimizes the expected total system cost, which is the sum of all staffing and customer-related (waiting and abandonment) costs, i.e., our focus is on ensuring a sufficiently high quality of service to customers; this is usually a major concern in service systems. Since the staffing problem in our general context is not amenable to exact analysis, we determine optimal staffing levels by solving its fluid approximation.

To quantify the impact of self-scheduling, we consider as benchmark a system where it is optimal, in the absence of self-scheduling, to match the supply (service) and demand (arrival) rates in each shift. Then, there is no congestion in the system, at fluid scale, and customer impatience does not play any role. When servers self-schedule, the resulting uncertainty in the numbers of

[1] http://www.ariseworkfromhome.com/faq/agent-questions/

[2] http://www.businessinsider.com/how-much-you-earn-as-an-uber-driver-2014-6?IR=T

[3] http://workathomemoms.about.com/od/workathomecareers/p/callcenterprofi.htm

servers creates congestion in understaffed shifts. Because of that congestion, the specific customer abandonment distribution now plays an important role.

Short-term controls. We assume that the manager may control the compensation that she offers her agents, and investigate how to optimize this control along with the staffing decision. Indeed, controlling agents through compensation is the rationale behind "surge pricing" in ride-sharing platforms (Uber 2017): By increasing (decreasing) compensation in a given shift, more (less) agents should be willing to participate (Gurvich et al. 2017, Cachon et al. 2016, Riquelme et al. 2016).

However, there is also a need to consider alternative tools, besides compensation and staffing, to control the system. First, the manager may be restricted in how much and how often she can modify compensation. This is certainty the case in virtual call centers because of market transparency and fierce competition between providers. Also, in virtual call centers, compensations are often set in advance by client companies rather than by the virtual call-center platform itself. In this case, the responsibility of the platform is to staff and train agents, and act as an intermediary between client companies and their agents [4] [5]. Second, while pricing influences agents, it cannot always be used to influence the behaviour of customers, e.g., in service-oriented virtual call centers; there is therefore a need to consider other customer-side controls. Third, there is considerable concern about the extent to which pricing should be used as a control in on-demand service platforms, because of extreme and frequent fluctuations. As was noted in Taylor (2017), most on-demand service platforms avoid real-time pricing because of customer resistance to it. As a result, there are numerous calls to consider alternatives (Harvard Business Review 2015). We propose one such alternative in this paper.

Given that customer impatience plays an important role, it is natural to think of ways of controlling customer abandonment behavior so as to alleviate the cost of self scheduling. Here, we propose doing so by communicating to customers information about upcoming delays, in the form of delay announcements. Indeed, delay announcements are known to impact customer abandonment behavior in practice (Mandelbaum and Zeltyn 2013, Aksin et al. 2016, Yu et al. 2016, Ibrahim et al. 2017). When customers respond to the announcements, their behavior alters the performance in the system which, in turn, affects the future announcements given. Therefore, studying customer response requires an equilibrium analysis i.e., one where announced and experienced delays coin-

[4] `http://www.ariseworkfromhome.com/faq/agent-questions/`

[5] `http://join.liveops.com/sales-independent-contractor-better-than-work-from-home-jobs`

cide in the fluid approximation. Communicating announcements allows the manager to alleviate congestion in understaffed shifts, and may also be seen as an alternative to restricting the freedom of agents by capping their access in overstaffed shifts (Gurvich et al. 2017). Since restricting agent flexibility through caps is typically a cause of agent complaint [6], or may not be possible in practice (e.g., it would be easy to send nurses home if they do show up, or to limit the number of Uber drivers that flock to a certain area), it is useful to consider alternatives.

In this paper, we study effective ways of managing service platforms with self-scheduling agents where the manager has three controls in her toolbox: (i) the staffing level, (ii) the announcements, and (iii) the compensation. For completeness, we consider the optimization of such controls individually (fix the other two), pairwise (fix only one), and jointly, because the time scales at which those decisions are made in practice may vary depending on context. When the agent pool size is fixed, the announcements and the compensation may be considered to be short-term controls. Shortly before a given shift, the manager may have at her disposal updated or revised arrival-rate forecasts which are based on some new information, e.g., for ride-sharing services, a concert may have just ended in some region, creating a surge in customer demand for Uber cars. Then, the manager would optimize her compensation and announcement decisions, for that shift, based on those updated arrival rates; we consider this joint optimization problem in §6.2.

Contributions. In this paper, we make the following contributions.

- For a rigorous treatment, we prove the asymptotic (for large arrival rates) accuracy of our fluid approximation to the system with a binomially distributed number of servers and exponential times to abandon. The binomial distribution arises when agents make independent decisions, each with probability r_j, to be available in shift j.

- We characterize the role that customer impatience plays in on-demand service platforms by deriving various stochastic-order relations between different customer abandonment distributions and characterizing how those relations impact the system's cost (extending part of the analysis in Bassamboo and Randhawa (2010) who do not explicitly study the impact on cost and Whitt (2006b) who focuses on a single-shift setting).

[6] `https://www.glassdoor.co.uk/Reviews/Employee-Review-Arise-RVW9860651.htm`

- We determine the optimal long-term staffing policy: We find that it is optimal to either match the supply and demand rates in one of the shifts, thus overstaffing or understaffing the remaining shifts, or to not match the supply and demand rates in any of the shifts, depending on whether the abandonment distribution has a non-decreasing (former) or decreasing (latter) hazard rate. This goes against conventional wisdom in workforce management which supports optimizing controls in order to match supply with customer demand [7].

- We analyze the system with announcements and a self-scheduling capacity across multiple shifts (extending part of the analysis in Armony et al. (2009) who focus on a single shift). Analyzing the system with multiple shifts is not straightforward: Because of self-scheduling, there are different staffing levels in different shifts, leading to different equilibrium announcements and abandonment distributions, depending on the shift. We derive a condition under which the announcements lead to a decrease in the cost of self-scheduling, across all shifts. In the same spirit as Huang et al. (2017), we also position the announcements more generally in a broader context of operational decision-making in service systems: We solve a joint staffing and announcement problem and show that a manager that will be using announcements at a later stage may decide on a different initial staffing level in the first stage.

We formulate the following managerial insights based on optimizing the three controls:

- For a fixed agent pool size, the manager should vary compensation to either incite enough agents to participate (i.e., match demand), or intentionally incite *a smaller* or *larger* supply of agents than the incoming demand. This suggests e.g., that “surge pricing” need not always be used to match supply and demand, which is the guiding principle in managing on-demand service platforms (The Economist 2016).

- By using the announcements, the manager can alleviate costs by altering customer abandonment behavior in understaffed periods. However, if the announcements are sufficiently effective in reducing customer-related costs, then the manager has less incentive to offer high pay to her agents to induce their participation; in other words, agents are worse off because of the announcements.

[7] `http://searchcrm.techtarget.com/tip/Using-workforce-management-software-effectively-in-contact-centers`

- When optimizing all three controls, the *value* of the minimum wage that the manager has to pay her agents plays a major role. To characterize that role, as we do here, is especially important in light of the current debate about the necessity to offer minimum wage to agents in on-demand service platforms[8]. If the minimum wage is “sufficiently low”, then the manager uses *only* the compensation and staffing controls, but does not resort to the announcements: She eliminates all congestion in the system by paying compensation which may be *strictly higher* than the minimum wage. If the minimum wage is “high”, then the manager pays that wage in all shifts and uses the announcements to alleviate congestion in understaffed periods.

Organization. Here is how this paper is organized. In §2, we review the relevant literature. In §3, we describe our modelling framework and the system manager's problem. In §4, we formulate staffing recommendations with self-scheduling agents. In §5, we study the problem with delay announcements. In §6, we investigate the compensation optimization problem, and the joint optimization of all controls. In §7, we establish the asymptotic accuracy of the fluid approximation with a binomial number of servers, and in §8, we draw conclusions. We relegate all proofs to the appendix, and present some additional related material in an online supplement.

2 Related Literature

Our modelling approach is close to the stream of literature initiated by Harrison and Zeevi (2005) which addresses the question of capacity planning under parameter uncertainty. Our paper is also 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, 2006); for other references, see Gans et al. (2003) and Akşin et al. (2007). However, none of those papers considers a random number of servers. 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.

[8] `http://uk.businessinsider.com/british-uber-drivers-entitled-to-minimum-wage-holiday-pay-lodon-tribunal-rules-2`

Our paper is also related to the literature which considers a random arrival rate instead of a random number of servers (Aldor-Noiman et al. 2009, Jongbloed and Koole 2001, Steckley et al. 2005). Bassamboo et al. (2010) rely on a stochastic-fluid approximation to determine optimal staffing levels in many-server queues with random arrival rates, and Whitt (2006b) relies on a fluid approximation to systems with with an uncertain arrival rate, an uncertain number of servers, and a single shift (our focus here is on staffing multiple shifts instead, and on making delay announcements in that setting). It is important to acknowledge that the fluid approximations for systems with random arrival rates or with a random number of servers are equivalent. However, the optimization and control problems that we study in this paper (jointly over both compensation and delay announcements) are especially relevant to a system with a random number of servers, and are not covered by existing results on queues with a random arrival rate. To capture the distinction between those two types of randomness, in number of servers versus in arrivals rates, it is necessary to go beyond the fluid approximation, as we do in a follow-up paper, Dong and Ibrahim (2017).

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 (2016). Our work is related to papers on nurse staffing with absenteeism, such as Green et al. (2013) and Wang and Gupta (2014). However, our self-scheduling staffing context is different because agents self-schedule to different shifts from a single pool, based on their availabilities and preferences, whereas each clinical unit may be staffed separately. This paper is related to research on delay announcements, including Armony et al. (2009), Jouini et al. (2011), Allon and Bassamboo (2011), Aksin et al. (2016), and Yu et al. (2016). However, none of these papers considers multiple shifts, nor the joint staffing and announcement problem which arises in a context with self-scheduling servers.

This paper is most closely related to recent papers on queues with a self-scheduling capacity. The paper closest to ours is Gurvich et al. (2017), who were the first to study the operational management of systems with self-scheduling agents. They consider a *profit-maximizing* firm which can control the pool size, the compensation, as well as place a cap on agent participation in overstaffed shifts. In contrast, we focus here on minimizing costs when *quality of service* is important and customers are impatient. Modelling system congestion and customer impatience allows us to analyze a *customer-*

side control instead, i.e., the announcements. More generally, there is a growing stream of literature on the management of on-demand service platforms (see for example Ozkan and Ward (2017), Hu and Zhou (2017b), Braverman et al. (2017), Taylor (2017), Cachon et al. (2016), Riquelme et al. (2016), Tang et al. (2017), Bimpikis et al. (2017), Hu and Zhou (2017a,b), Feng et al. (2017), Hu and Chen (2017), etc.). Our paper is related to that stream of literature, but our focus on staffing queues with randomness in capacity, and focusing on the role of customer impatience, are different. Ata et al. (2017) is also relevant to our work, albeit in a different application context (volunteer gleaning operations), and with a different focus.

3 Modelling Framework

In this section, we describe our modelling framework: First, we describe our queueing framework, and then we formulate the optimization problem faced by the system manager.

Queueing Model. There are k shifts and we consider single-class $G/G/N_j^n+GI$ queueing models in steady state, where j indexes the shift and N_j^n is a random variable which depends on the pool size n. As in Gurvich et al. (2017), we assume that agents are statistically identical and have an availability threshold (opportunity cost) T for showing up in shift j. Letting $G(\cdot)$ denote the cumulative distribution function (cdf) of T, an agent shows up in shift j with probability $r_j \equiv G(c_j)$. In particular, $\mathbb{E}[N_j^n] = nr_j$ is the expected number of servers in shift j. We assume that $G(\cdot)$ is log-concave with positive density function $g(\cdot)$. We emphasize that there we make no restriction on whether the agent may appear in multiple and/or successive shifts. Service times are independent and identically distributed (i.i.d.) random variables with a general distribution and mean $1/\mu$. We assume, without loss of generality, that $\mu = 1$.

Each customer will abandon if he is unable to start service before a random amount of time, which we refer to as his patience time. Patience times are i.i.d. across customers, and have a cdf F, complementary cdf (ccdf) $\bar{F}$, density function f, hazard-rate function h_a, and mean $1/\theta$ for some $\theta > 0$. Abandonment makes the system stable, even when N_j^n is random (Whitt 2006b) [9]. Customers arrive to the system according to general stationary processes with rates λ_j, $1 \leq j \leq k$.

[9] Specifically, with abandonment and a deterministic N, a proper steady-state distribution always should exist. Stability with a random N follows by conditioning and unconditioning on N.

We assume that there is no service overlap between the different shifts, i.e., customers who arrive during a shift must be served by agents who are assigned to that shift. While this assumption may not always be justifiable, it is reasonable when the system is sufficiently large. The arrival, service, and abandonment processes are all mutually independent, also independent of N_j^n. There is unlimited waiting space, and we use the first-come-first-served service discipline.

System Manager's Problem. As in Bassamboo and Randhawa (2010), we consider two quality-of-service costs, indexed by the shift j: (i) A delay cost, h_j, per customer for each unit of time that this customer spends waiting to be served, and (ii) an abandonment penalty cost, p_j, incurred per customer who abandons before being served. Let $Q_{N_j^n}$ denote the steady-state queue length and $\alpha_{N_j^n}$ denote the net customer abandonment rate. The system manager can decide on both the staffing level n and the compensation to offer her agents. Because of fierce competition between alternative service providers, and because of mounting pressure to offer a sufficiently high compensation [10] to agents in on-demand service platforms, we assume that there exists a minimum wage, l, i.e., that we must have $c_j \geq l$ in every shift j (Gurvich et al. 2017). Here is the manager's problem:

$$\min_{c_j \geq l, n \in \mathbb{N}} \Pi(n, \mathbf{c}) \equiv \sum_{1 \leq j \leq k} \left(c_j \cdot \mathbb{E}[N_j^n] + p_j \cdot \mathbb{E}[\alpha_{N_j^n}] + h_j \cdot \mathbb{E}[Q_{N_j^n}] \right), \tag{3.1}$$

where $\mathbf{c} \equiv (c_1, c_2, \cdots, c_k)$ is the k-dimensional vector of compensations and $\mathbb{N}$ denotes the set of natural integers. We note in passing that the formulation in (3.1) is a *cost* formulation, i.e., the objective is to minimize the expected cost in the system. An alternative formulation would be to consider a *constraint* formulation instead, e.g., to impose a service-level constraint on the waiting-time. This distinction is carefully treated in Bum Soh and Gurvich (2017) who explore the duality between the two formulations in both single-class and multi-class queues. They find that while the optimal trade-off of capacity and delay can be implemented via a staffing problem with average waiting constraints in a single-class setting, the problem is more complicated with multiple classes where a priority scheme must be implemented as well.

Since the problem in (3.1) is not amenable to exact analysis, we consider a steady-state fluid approximation of the system instead. For an $G/G/s+GI$ system, $\bar{q}_{\rho_s}$ and $\bar{\alpha}_{\rho_s}$ are, respectively, the fluid approximations for the queue-length and net abandonment rates with traffic intensity

[10] `http://www.zerohedge.com/news/2016-12-28/uber-out-drivers-sue-sharing-economy-champion-minimum-wage`

$\rho_s \equiv \lambda/s\mu$. The fluid approximation to problem (3.1) is:

$$\min_{c_j \geq l, n \in \mathbb{N}} C(n, \mathbf{c}) \equiv \sum_{1 \leq j \leq k} nG(c_j)c_j + p_j \cdot \bar{\alpha}_{\rho_j/G(c_j)} + h_j \cdot \bar{q}_{\rho_j/G(c_j)}, \tag{3.2}$$

where $\mathbb{E}[N_j^n] = nG(c_j)$. In §7, we establish the asymptotic accuracy of our fluid approximation and quantify the orders of magnitude of the resulting errors, depending on the prescribed asymptotic regime, when the number of servers has a binomial distribution. There, we show that our fluid approximation is extremely accurate, particularly when the system is heavily congested.

Under the fluid approximation in (3.2), only the *expected* number of agents who show up in a given shift, e.g., based on the offered compensation, matters. In reality, the variance in the number of agents who show up is also important to consider, as it could considerably impact performance measures in the system. For example, larger agent pools may also entail higher variance which must be planned for by the system manager. In this paper, we do not address that issue further as our main goal is to derive insights on how to manage such systems using both short-term (compensation, announcements) and long-term (staffing) controls. We study the impact of the variance in capacity on the operational management of the system in a follow-up paper, Dong and Ibrahim (2017).

4 Long-Term Staffing Policy

In this section, we solve the long-term staffing problem with self-scheduling servers. Here, we fix $\mathbf{c}$ in (3.1) and focus on determining the capacity n directly as a function of $r_j = G(c_j)$. Later, we solve the manager's problem when she can jointly decide on all controls in her toolbox. Since we are particularly interested in describing the role played by the abandonment distribution, and in order to ground our theoretical analysis in common practice, we begin by highlighting some empirical findings on how customers abandon in real-life service systems.

4.1 How Do Customers Abandon in Practice?

Using hazard rates to describe customer patience dates back to Palm (1953), and is common in statistical inference studies on customer abandonment; e.g., see Brown et al. (2005). Mandelbaum and Zeltyn (2013) also advocate the usage of survival functions for patience inference.

Existing empirical evidence, from both the call center and healthcare settings, suggests that customers typically have an abandonment distribution with a decreasing hazard rate. In a call center setting, Mandelbaum and Zeltyn (2013) find that “customers who have already waited for a significant time, tend to remain increasingly patient” (p. 14). In an emergency department setting, Bolandifar et al. (2016) also reach the conclusion that the abandonment distribution of patients has a decreasing hazard rate, and attribute this to the “sunk cost effect” (p. 19). In what follows, we consider abandonment distributions with both increasing and decreasing hazard rates, and determine the optimal staffing policy in each case. Later, we formulate stochastic order relations between abandonment distributions using both survival functions and hazard rates, and study the impact of those relations on the cost in the system.

4.2 Benchmark Case: No Self-Scheduling

Without self-scheduling, the system manager can independently select the optimal staffing levels, n_j^* in each shift j. However, with self-scheduling, she can only choose the total staffing level, n^*, and allow agents in the pool of size n^* to self schedule.

The density of the fluid that has been waiting for exactly u time units, in shift j, is equal to $\lambda_j \bar{F}(u)$. Therefore, the corresponding (unscaled) queue length is given by $q_j = \int_0^{w_j} \lambda_j \bar{F}(u)\,\mathrm{d}u$, where w_j denotes the waiting time given service. The net abandonment rate (unscaled) in shift j is equal to $\lambda_j F(w_j)$. In the absence of self-scheduling, we must have that $n_j^* = \lambda_j \bar{F}(w_j^*) \le \lambda_j$ where w_j^* is the optimal waiting time in shift j; indeed, it is then suboptimal to staff more than λ_j agents in shift j. The fluid approximation to the system manager's problem for shift j is:

$$\min_{w_j \ge 0} \lambda_j \left((c_j - p_j)\bar{F}(w_j) + h_j \int_0^{w_j} \bar{F}(u)\,\mathrm{d}u \right). \tag{4.3}$$

Hereafter, we make the following assumption.

Assumption 4.1. *For all j, $c_j < \min\{h_j/h_a(0) + p_j, h_j/\theta + p_j\}$.*

Assumption 4.1 states that staffing costs are sufficiently inexpensive; this is consistent with the assumptions in Bassamboo and Randhawa (2010). Then, it is easy to establish the following result for the optimal solution to problem (4.3).

Proposition 4.1. *Under Assumption 4.1, in a system with no self-scheduling servers, it is optimal to match the supply and demand rates in every shift, i.e., $n_j^* = \lambda_j$.*

Proposition 4.1 shows that, in a system without self-scheduling, it is optimal to operate every shift in the critically-loaded regime (Halfin and Whitt 1981). Thus, all customers are served immediately upon arrival, and there is no reneging from the system. Intuitively, when the staffing costs are small enough (less than the upper bound in the assumption), it is cost effective for the manager to staff a large enough agent pool to eliminate, at fluid scale, all congestion from her system. In contrast, if the staffing costs are high, then it would be cost effective to allow for congestion instead, i.e., purposely deteriorate the service level because staffing enough agents is too costly.

4.3 Self-Scheduling Capacity

We define the *augmented arrival rate* $\Gamma_j \equiv \lambda_j / r_j$, and let $\Gamma_0 \equiv 0$. Defining the augmented demand rates as such allows us to capture the salient heterogeneity across shifts, which is the key challenge in managing a random capacity. That heterogeneity is due to two factors: (i) time-dependent demand λ_i and (ii) time-dependent availabilities of agents r_i. Without loss of generality, we assume that the alternative shifts are numbered in order of increasing Γ_j values, i.e., $\Gamma_{j-1} \leq \Gamma_j$ for all $j \in \{1, 2, \cdots, k\}$. In other words, we re-index the different shifts so that the Γ_i values are ordered. That is, if $\Gamma_1 > \Gamma_2$ and there are only two periods, we re-index shift 1 as shift 2, and vice-versa. For a total staffing level $\Gamma_{j-1} < n < \Gamma_j$, all shifts with index i where $i \leq j-1$ are underloaded, whereas all shifts where $i \geq j$ are overloaded. Moreover, letting $n = \Gamma_j$ amounts to matching the supply and demand rates in shift j. In an overloaded shift i, we have that $\Gamma_i \bar{F}(w_i) = n$, i.e., $w_i = \bar{F}^{-1}(n/\Gamma_i)$. Since it is never optimal to strictly underload all shifts, the system manager's problem can be defined piecewise over the successive $[\Gamma_{j-1}, \Gamma_j)$ intervals as:

$$\min_{0 \leq n \leq \Gamma_k} C(n) \equiv \left(\sum_{j=1}^{k} \mathbf{1}(\Gamma_{j-1} \leq n < \Gamma_j) u_j(n) \right), \tag{4.4}$$

where $\mathbf{1}(n \in A)$ denotes the indicator function over the set A, and $u_j(n)$ is given by:

$$u_j(n) \equiv \sum_{i=1}^{k} c_i n r_i + \sum_{i=j}^{k} \left(p_i(\lambda_i - n r_i) + h_i \lambda_i \int_0^{\bar{F}^{-1}(n/\Gamma_i)} \bar{F}(u) \, \mathrm{d}u \right), \tag{4.5}$$

i.e., $u_j(n)$ is the total cost incurred if n is chosen in the interval $[\Gamma_{j-1}, \Gamma_j)$. In (4.4), for each value of n, exactly one of the indicator functions will be equal to 1 and the rest will be equal to 0. For example, if $n \in [\Gamma_{i_0-1}, \Gamma_{i_0})$, then $C(n) = u_{i_0}(n)$ for $u_i(n)$ in (4.5). That is, all shifts which are indexed i_0 and above will be congested and the manager incurs customer-related costs in those shifts, whereas the remaining shifts, indexed below i_0, are overstaffed (no congestion). We begin with the following proposition providing necessary and sufficient conditions under which self-scheduling is *not* costly to the manager, relative to the benchmark.

Proposition 4.2. *Self-scheduling is not costly to the system manager if, and only if, the resulting augmented arrival rates are identical across all shifts.*

Proposition 4.2 highlights the importance of considering multiple shifts in our setting. Indeed, with a single shift, it is readily seen that the manager can simply staff a large enough agent pool, equal to λ/r, so as to eliminate the cost of self-scheduling in her system. It is because the manager is confronted with a self-scheduling capacity across multiple shifts that she has to pay a price for self-scheduling. Based on Proposition 4.2, if the manager is able to solicit her agents' scheduling preferences upon hire, then she should make staffing decisions in a way to ensure demand-augmented uniform plans across shifts. For example, if the morning shift (M) typically experiences high demand while the afternoon shift (A) typically experiences low demand, i.e., $\lambda_M > \lambda_A$, then she should hire agents with a stronger preference for the morning shift, i.e., $r_M > r_A$ so that $\Gamma_M = \Gamma_A$. By doing so, she could eliminate the cost of self-scheduling. The problem is, of course, that hiring agents based on their scheduling preferences is not usually possible in our context, e.g., with self-scheduling agents in virtual call centers. Instead, the manager may only have historical estimates of agent preferences, i.e., of the r_j values, and of the resulting Γ_j values, and know the cost structure in her system. Next, we investigate her staffing problem when the augmented arrival rates are not uniform across shifts.

4.4 Monotonically Non-Decreasing Hazard Rate

For abandonment distributions with a monotonically increasing hazard rate or with exponential abandonment, we find that it is optimal to match the supply and demand rates in one of the k shifts when servers self schedule (as opposed to all shifts when they do not self-schedule), with the remaining shifts being either over or under staffed.

Proposition 4.3. *For abandonment distributions with a monotonically non-decreasing hazard rate:*

- *The objective function in problem (4.4) is piecewise concave (piecewise linear with exponential);*
- *There is one shift i_0 where the supply and demand rates must be matched, i.e., $n^* = \Gamma_{i_0}$;*
- *With exponential abandonment, i_0 must satisfy the following condition where $L_j \equiv p_j + \frac{h_j}{\theta}$:*

$$\sum_{j=1}^{k} c_j r_j - \sum_{j=i_0}^{k} L_j r_j < 0 \text{ and } \sum_{j=1}^{k} c_j r_j - \sum_{j=i_0+1}^{k} L_j r_j > 0. \tag{4.6}$$

Proposition 4.3 shows that solving the staffing problem in this case reduces to determining which of the k shifts to critically load. For example, with exponential abandonment, the condition in (4.6) can be interpreted as follows. Starting with a staffing level equal to 0, a unit increase in the staffing level increases the expected staffing cost by $\sum_{j=1}^{k} c_j r_j$. It also decreases the expected congestion cost by $\sum_{j=1}^{k} L_j r_j$. Under Assumption 4.1, adding one server to an empty pool will yield an overall cost decrease in the system, since $\sum_{j=1}^{k} c_j r_j < \sum_{j=1}^{k} L_j r_j$. Condition (4.6) states that the manager must continue increasing the staffing level until the rate of decrease in congestion costs no longer offsets the rate of increase in staffing costs: The supply and demand rates in shift i_0 are then matched. The optimality of overstaffing certain shifts lends some support to the staffing policies adopted in virtual call centers such as LiveOps [11] or Arise [12], where agents regularly complain about the fact that there are "too many other agents on board" and, consequently, "too few calls to answer". However, the compensation structure in those settings is different: There, the manager typically uses volume-dependent pay, e.g., agents earn a piece-rate compensation in addition to some base salary. Under our fixed compensation structure, we find that overstaffing certain shifts can minimize costs, but that this is not true for all shifts.

To illustrate how the staffing policy of Proposition 4.3 may be implemented in practice, we now discuss a simple example with three shifts: morning (M), early afternoon (EA), and (3) late afternoon (LA). Let us assume that $\lambda_{LA} < \lambda_M < \lambda_{EA}$. That is, demand is highest in the early afternoon, followed by morning, and then late afternoon. Let us now compare two different patterns. Under pattern 1, we assume that many agents show up in the morning, few in the early afternoon, and many again in the late afternoon. In particular, we assume that $r_{EA} < r_{LA} < r_M$ and that, as a result, we have $\Gamma_M < \Gamma_{LA} < \Gamma_{EA}$. Recall that we re-index the shifts in order of increasing Γ_i to

[11] https://www.glassdoor.ie/Reviews/Employee-Review-LiveOps-RVW6931455.htm

[12] https://www.glassdoor.co.uk/Reviews/Arise-technical-support-Reviews-EI-IE31617.0,5-PKH6,23.htm

derive the optimal staffing policy. Under pattern 1, since Γ_{EA} is largest, it will be indexed as the third shift. Thus, the shifts are re-indexed as: 1: M, 2: LA, and 3: EA.

Let us now consider another pattern. Under pattern 2, let us assume that few agents show up in the morning, many show up in the early afternoon, and then few in the late afternoon. In particular, $r_M < r_{LA} < r_{EA}$. Let us also assume that we then obtain $\Gamma_{EA} < \Gamma_M < \Gamma_{LA}$. The shifts are now indexed as 1: EA, 2: M, and 3: LA. That is, even though the early afternoon shift has the highest demand, it has the lowest Γ value since many agents show up for that shift. Since the staffing policy in Proposition 4.3 depends on both the indexing of the shifts and the cost structure, we may have different staffing prescriptions under patterns 1 and 2. For example, the morning shift may be overstaffed under pattern 1, and understaffed under pattern 2.

4.5 Monotonically Decreasing Hazard Rate

We now consider abandonment distributions with a monotonically decreasing hazard rate, which is consistent with the way call center customers abandon in practice (§4.1).

Proposition 4.4. *For abandonment distributions with a monotonically decreasing hazard rate:*

- *The objective function in problem (4.4) is piecewise convex;*
- *If there exists $1 \leq i_0 \leq k$ such that $C'(\Gamma_{i_0-1}) < 0$ and $C'(\Gamma_{i_0}^-) > 0$, then the optimal solution to problem (4.4) is $n^* \in (\Gamma_{i_0-1}, \Gamma_{i_0})$, i.e., it is optimal to either under or over staff every shift (no matching). Otherwise, it is optimal to match the supply and demand rates in one of the shifts.*

Interestingly, Proposition 4.4 shows that it may be optimal for the manager to not match the supply and demand rates anywhere, i.e., to effectively under or over load every shift. In the appendix, we derive a sufficient condition on the augmented arrival rates for this to be the case (in the proof of the proposition). At a high level, this sufficient condition shows that if the imbalance between the augmented arrival rates, measured by $\Gamma_{i_0-1}/\Gamma_{i_0}$, is small enough then it may be optimal to "strike a balance" between the two shifts $i_0 - 1$ and i_0, i.e., to underload shift $i_0 - 1$, while overloading shift i_0. This result is different from, e.g., the result in Wang and Gupta (2014) who study the nurse staffing problem with absenteeism and show, under a similar first-order approximation as ours, that "assignments must match average supply to mean demand" (p. 440) in each shift. Indeed, the main difference between our setting and theirs is that agents self-schedule from a single pool, whereas there are distinct pools associated with clinical units, and each may be staffed separately.

In practical terms, Proposition 4.4 shows that it may be optimal for the manager to maintain an imbalance between the average supply and demand rates in each of the shifts. In other words, "having just the right number of staff available within each interval of the day to meet established service levels" [13], which is conventional wisdom for workforce management in call centers, may no longer be the right approach with self-scheduling agents, since it may be optimal not to meet the established service level in any shift, but rather to exceed or fall below it.

4.6 Numerical Example

We now turn to illustrating the impact of the abandonment distribution on the system's cost. In Figures 1 and 2, we solve the staffing problems without and with self-scheduling, respectively, for a Weibull abandonment distribution with fixed mean (equal to 1) and alternative values of the shape parameter, s. Considering Weibull abandonment is convenient because its hazard rate has different monotonicities depending on the value of s: For $s < 1$, it is decreasing (DFR), for $s > 1$ it is increasing (IFR), and for $s = 1$ it is constant. Other parameters are held constant across the two figures, in particular we consider $k = 5$ shifts, and assume equal cost parameters across all shifts: $c = 0.8$, $h = 0.8$, $p = 1$. We also let $r = 0.4$. We let the average arrival rate be equal to 55, and assume equal increments in the arrival rates across the shifts, i.e., $\lambda_{i+1} - \lambda_i$ is constant for $1 \leq i \leq k-1$. We assume that $\lambda_{max}/\lambda_{min} = 5$, so that the shift with the highest arrival rate has an arrival rate which is 5 times larger than the shift with the smallest arrival rate. Figure 1 illustrates that the abandonment distribution plays no role without self scheduling, since the optimal staffing cost is constant, and there is no congestion anywhere. In contrast, Figure 2 shows that, while self-scheduling is always costly, the cost of self scheduling itself depends on the specific shape of the abandonment distribution. Figure 2 suggests that if two abandonment distributions have the same mean, a distribution with a decreasing hazard rate yields a smaller cost than one with an increasing hazard rate. We demonstrate this, along with other properties, next.

[13] http://www.icmi.com/Resources/Workforce-Management/2015/03/How-Remote-Employees-Can-Help-Contact-Centers-Better

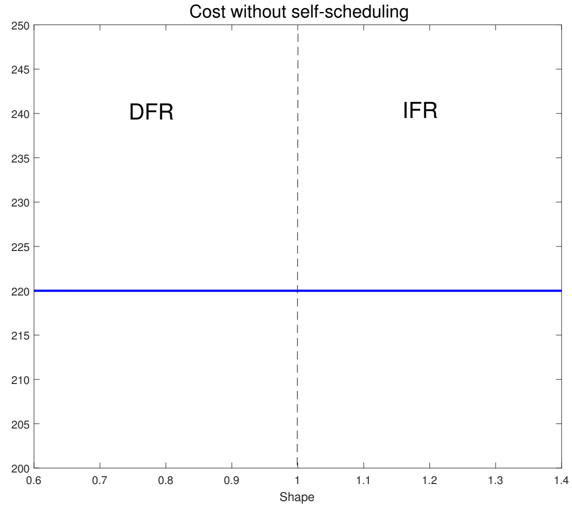

Figure 1: Benchmark cost (Problem (4.3)) without self-scheduling and Weibull abandonment under a constant mean and different shapes.

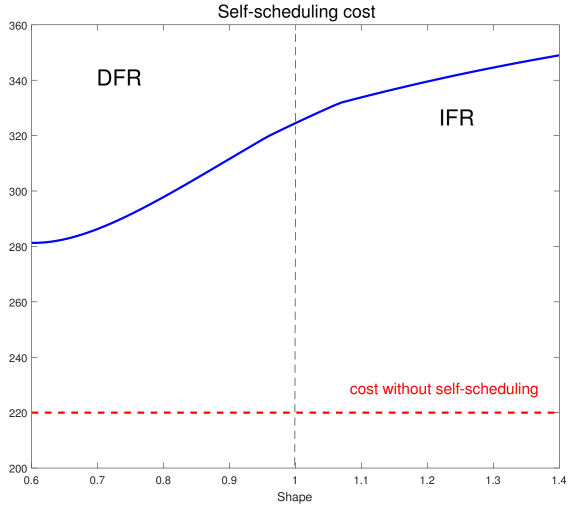

Figure 2: Cost of self-scheduling (Problem (3.2)) with Weibull abandonment under a constant mean and different shapes.

4.7 Stochastic Order Relations

We now study the impact of various stochastic order relations, between abandonment distributions, on the system's cost; for background, see Shaked and Shanthikumar (2007). Let X_1 and X_2 denote two generic times to abandon random variables, with cumulative distributive functions F_1, F_2 and hazard functions h_{a1}, h_{a2}, respectively. Let C_i^* denote the optimal cost under abandonment distribution i, i.e., C_i^* is the optimal objective value in (3.2), for $i = 1, 2$.

Proposition 4.5.

- *Under equal mean times to abandon, if h_{a1} is decreasing and h_{a2} is increasing, then $C_1^* \leq C_2^*$;*
- *Under equal mean times to abandon, if X_1 is new worse than used in expectation (NWUE) and X_2 is new better than used in expectation (NBUE) [14], then $C_1^* \leq C_2^*$;*
- *If $X_1 \leq^{st} X_2$ (first-order stochastic dominance), $X_1 \leq^{LR} X_2$ (likelihood ratio order), $X_1 \leq^{HR} X_2$ (hazard rate order), or $X_1 \leq^{RHR} X_2$ (reverse hazard rate order), then $C_1^* \leq C_2^*$.*

We begin by noting that the results in Proposition 4.5 apply generally to any congested system; since self-scheduling creates understaffed shifts, they apply to our specific context as well. Proposition 4.5

[14] A nonnegative random variable X is said to be NWUE (NBUE) if $\mathbb{E}[X - a|X > a] \geq (\leq)\mathbb{E}[X]$ for all $a \geq 0$.

provides a theoretical justification for the observations made in Figure 2: assuming equal mean times to abandon, the Weibull distribution has a decreasing hazard rate when $s < 1$, and an increasing hazard rate when $s > 1$; thus, the cost of self scheduling must be smaller for $s < 1$. Intuitively, with a decreasing hazard rate, waiting customers are impatient initially and become increasingly patient with time. This leads to a reduction in the average waiting time for all fluid in the system and, consequently, reduces the cost of self-scheduling. In contrast, with an increasing hazard rate, customers grow increasingly impatient with time, but are more patient initially. This leads to an increase in the overall average waiting time for all fluid in the system. The same intuition holds when one of the two distributions is NBUE, while the other is NWUE.

When X_1 and X_2 are ordered in a first-order stochastic dominance sense, we can derive a stronger result: Then, the fluid waiting time given service, waiting time given abandonment, and overall waiting time can all be shown to be smaller with abandonment distribution F_1 than with F_2. As a result, the system's cost is also smaller. Since likelihood ratio, hazard rate, and reverse hazard rate dominance all imply first-order stochastic dominance, the same holds under those types of stochastic orders as well.

5 Controlling Customers: Delay Announcements

Our analysis so far has focused on emphasizing the role played by the customer abandonment distribution in a system with randomness in capacity. Because of this, it is natural to investigate ways of controlling this abandonment behaviour so as to alleviate the system's cost. We now propose to do so via delay announcements in the system. In this section, we first explore the impact of the announcements by assuming that both n and c_j are fixed. The manager may also have updated demand-rate forecasts at her disposal, based on which she would make a decision on whether or not to make announcements. In that sense, the announcements are viewed as a real-time control that is decided upon in a short-time scale. Then, we study a joint staffing and announcement problem.

5.1 How Do Customers React to Delay Announcements in Practice?

We begin by highlighting empirical evidence describing how customers react to delay announcements in practice. That evidence will subsequently guide us in modelling customer response to the announcements. Based on their analysis of call-center data, Mandelbaum and Zeltyn (2013) found that “customers who are promised a short wait become impatient at some point” and “customers with the longest estimated wait seem to be relatively impatient” (p. 22). This evidence suggests that customers grow increasingly impatient with the magnitude of the announced delay, i.e., that their mean time to abandon is decreasing in the announced delay. Similarly, Aksin et al. (2016) found that “callers who receive information that the queue length is long abandon the system sooner and callers who receive information that the queue length is short abandon the system later compared to the case with no information.” (p. 31). Thus, delay announcements incite customers to abandon sooner, thereby reducing congestion in the system. Consistently, we make here the assumption that the mean time to abandon decreases with the magnitude of the announced delay.

5.2 The All-Exponential Model

Because we consider a system with multiple shifts, and different shifts have different congestion levels and therefore different delay announcements, we obtain in each shift a different announcement-dependent abandonment distribution. Herein lies the complexity of considering multiple shifts: The announcements may lead to shorter delays in some shifts, but not in others, and the aggregate effect of those announcements is unclear.

To derive insights on system performance, we focus hereafter on an exponential abandonment distribution with an announcement-dependent rate. In particular, letting w be the announcement made, customers abandon according to an exponential abandonment distribution with rate $\theta(w)$. We begin by noting that if the staffing level is fixed, then delay announcements cannot be used to *completely eliminate* the cost of self-scheduling in the system. This is because the rate of abandoning customers is unaffected by the announcements (it is determined solely by the arrival and service completion rates). Thus, the abandonment cost does not decrease. On the other hand, the announcements can be used to control the overall waiting time in the system.

5.2.1 Existence and Uniqueness of Equilibria

In this work, we contend that the announcements made must be truthful, for otherwise customers will learn to mistrust them. Since those announcements alter customer abandonment behavior which, in turn, affects the future announcements made, announcement accuracy in our fluid approximation reduces to investigating the existence of an equilibrium delay for which the waiting time of served customers coincides with the announcement made. Assume that the size of the agent pool is fixed and equal to n. Then, $n_j = nG(c_j)$ is the number of agents available in shift j. We let $w_j^e(n)$ denote the equilibrium delay in shift j, which is dependent on n. Then, we must have:

$$\lambda_j e^{-w_j^e(n)\theta(w_j^e(n))} = n_j, \quad \text{i.e.,} \quad e^{-w_j^e(n)\theta(w_j^e(n))} = \frac{n}{\Gamma_j}, \tag{5.7}$$

by conservation of flow in shift j. The total cost in the system, with the announcements, is

$$C_a(n) \equiv \sum_{i=1}^{k} c_j nG(c_j) + \sum_{i=1}^{k} \left(p_j + \frac{h_j}{\theta(w_j^e(n))} \right) (\lambda_j - nG(c_j))^+. \tag{5.8}$$

Assuming that $\theta(\cdot)$ is continuous and strictly increasing, consistently with the empirical evidence in §5.1, guarantees the existence and uniqueness of an equilibrium $w_j^e(n)$ in every shift j. In what follows, we also assume that $\theta(w)$ is a differentiable function of w and that $\lim_{w\to\infty} \theta(w) > 0$.

5.2.2 When Are the Announcements Effective?

We have different Γ_j values and, consequently, different announcement-dependent abandonment rates given by (5.7). We now derive a simple sufficient condition under which the announcements lead to an overall decrease in the system's cost, across all shifts. We let θ_0 denote the abandonment rate without the announcements, which is constant across all shifts. By Proposition 4.3, it is optimal to critically load one shift, call it i_c, i.e., $n^* = \Gamma_{i_c}$ without the announcements.

Proposition 5.1. *With exponential abandonment with an announcement-dependent rate $\theta(w)$, if*

$$\theta_0 \cdot \theta^{-1}(\theta_0) < \ln\left(\frac{\Gamma_{i_c+1}}{\Gamma_{i_c}}\right), \tag{5.9}$$

then $C_a(n^*) < C^*$ for $C_a(\cdot)$ in (5.8), where C^* is the optimal solution to (3.2) with $n^* = \Gamma_{i_c}$.

Proposition 5.1 shows that for $C_a(n^*) < C^*$ to hold, and under our assumptions on $\theta(\cdot)$, it suffices to impose a condition on customer response at a *single* point only, namely at θ_0, and only *two* shifts, i_c and $i_c + 1$: Only the augmented arrival rates in shifts i_c and $i_c + 1$ matter. The upper bound in (5.9) on θ_0 means that customers do not abandon too fast in the absence of announcements. Since

$\Gamma_{i_c+1}/\Gamma_{i_c}$ measures the "imbalance" in augmented arrival rates, (5.9) shows that the announcements are effective for a larger set of θ_0 values as that imbalance increases. This is desirable, since this is precisely when we would like to control the system. Here is an alternative explanation for the condition in (5.9). By dividing both sides by θ_0, and rewriting the right-hand-side as:

$$\frac{1}{\theta_0}\ln\left(\frac{\Gamma_{i_c+1}}{\Gamma_{i_c}}\right) = \frac{1}{\theta_0}\ln\left(\frac{\lambda_{i_c+1}}{n^* r_{i_c+1}}\right) \equiv w^0_{i_c+1},$$

where $w^0_{i_c+1}$ is the fluid waiting time in period i_c+1 under abandonment rate θ_0, we see that (5.9) can be interpreted as having a "long" waiting time, exceeding $\theta^{-1}(\theta_0)$, in the absence of the announcements, in shift i_c+1. Making the delay announcement would then encourage customers to abandon, thereby alleviating congestion. In other words: If the waiting time is long enough in shift i_c+1, then make the delay announcement.

Since the announcements lead to a decrease in waiting times, it is natural to investigate whether it is optimal for the manager to create additional congestion by understaffing her system in the first stage. The cost increase due to this congestion would subsequently be reduced by the announcements in the second stage. Next, we solve the manager's staffing problem with delay announcements.

5.2.3 A New Staffing Problem

The manager's staffing problem, assuming that she makes announcements in the second stage, is:

$$\min_{n\in\mathbb{N}} \quad \sum_{j=1}^{k}\left(c_j nG(c_j) + \left(p_j + \frac{h_j}{\theta(w^e_j(n))}\right)(\lambda_j - nG(c_j))^+\right), \tag{5.10}$$

where we replace the constant abandonment rate θ_0 by different announcement-dependent rates, $\theta(w^e_j(n))$, depending on both the shift and the staffing level n. That is, in setting her optimal staffing level, the manager needs to consider the subsequent dependence of customer abandonment behavior on the selected pool size. Let n^*_a denote the optimal solution to (5.10) [15], with the announcements, and n^* denote the optimal solution to (3.2), without the announcements.

[15] An optimal solution necessarily exists. If there are multiple optimal solutions, then we pick the smallest one.

Proposition 5.2. *With exponential abandonment with an announcement-dependent rate $\theta(w)$, if*

$$\theta_0 \cdot \theta^{-1}(\theta_0) < \min_{1 \leq i \leq k-1} \ln\left(\Gamma_{i+1}/\Gamma_i\right), \tag{5.11}$$

then $n_a^* < n^*$.

That is, under (5.11), it is optimal for the manager to hire a smaller agent pool than without the announcements. This conclusion is consistent with Huang et al. (2017), who consider a different optimization problem (they minimize the staffing level subject to a quality-of-service constraint) in the context of a single shift with no self-scheduling. They, too, find that the announcements may lead to understaffing. As such, we provide further evidence, in our new context, that management may indeed draw a dual benefit from the delay announcement: first, by reducing the waiting time of served customers and second, by reducing the staffing level. While Proposition 5.2 provides a sufficient condition for the system manager to understaff her system, compared to the no-announcement case, it does not quantify the decrease in cost which results from this. We explore this question, and others, in a numerical study which we relegate to the appendix (§**??** there).

6 Optimizing Staffing, Compensation, and the Announcements

In this section, we study how the manager may use all three controls in her toolbox, staffing, compensation, and the announcements, to effectively manage her system. We focus on the case with exponential abandonment, and begin by studying how compensation and the announcements may be used as short-term controls for a fixed staffing level (§6.1). Then, we study how staffing, compensation, and the announcements may be jointly optimized (§6.2).

6.1 Short-Term Controls: Compensation and Delay Announcements

Compensation. We assume that the staffing level is equal to n, and investigate the optimal compensation to be offered in shift k. In practice, while the manager has to make the staffing decision based on historical estimates of the arrival rates λ_k, she may update the compensation *in each shift* based on some revised estimate of the arrival rate for that shift, e.g., because of additional information at her disposal. The problem for shift k is given by[16]:

$$\min_{c_k \geq l} c_k n G(c_k) + \left(p_k + \frac{h_k}{\theta}\right)(\lambda_k - nG(c_k))^+ . \tag{6.12}$$

For expositional ease, we let $L_k \equiv p_k + h_k/\theta$ capture customer-related costs, and denote $\psi_k^n \equiv G^{-1}\left(\frac{\lambda_k}{n}\right)$. If $c_k = \psi_k^n$, then $nG(c_k) = \lambda_k$: This compensation incites just enough agents to meet demand in shift k. It will also be convenient to define $a_k < L_k$ as follows:

$$G(a_k)\left(1 + (a_k - L_k)\frac{g(a_k)}{G(a_k)}\right) = 0. \tag{6.13}$$

We note that (6.13) is the first-order condition of the unconstrained optimization problem in (6.12), provided that n is such that $n \leq \lambda_k/G(c_k)$; in other words, a_k is the optimizer of that unconstrained optimization problem. The optimal compensation in problem (6.12) is given by the following lemma where we implicitly assume that $l < a_k$; we relax this assumption in the proof of Lemma 6.1, in the appendix.

Lemma 6.1. *The optimal compensation in shift k, solution to (6.12), depends on n as follows:*

(a) *If* $n \geq \frac{\lambda_k}{G(l)}$*, then* $c_k^* = l$ *and shift k is overstaffed;*

(b) *If* $\frac{\lambda_k}{G(a_k)} \leq n < \frac{\lambda_k}{G(l)}$*, then* $c_k^* = \psi_k^n$ *and demand and supply are matched in shift k;*

(c) *If* $n < \frac{\lambda_k}{G(a_k)} < \frac{\lambda_k}{G(l)}$*, then* $c_k^* = a_k$ *and shift k is understaffed.*

Based on Lemma 6.1, we find that the manager uses the minimum wage in shift k when the agent pool size is very large (case (a)). In this case, the manager need not use high compensation to incite sufficient agent participation in the shift. For moderate values of the agent pool size (case (b)), the manager sets compensation to match demand and supply in the shift, i.e., $c_k^* = \psi_k^n$. Finally, when the pool size is very small (case (c)), inciting sufficient agent participation is too costly for the manager, so she sets a compensation that leads to an *understaffed* shift k. We note that the compensation offered to agents is monotonically decreasing in the agent pool size: the larger the pool, the smaller the compensation needed to incite agents to participate.

[16] While the arrival rates may be revised, we retain the same notation, λ_k, for simplicity.

Lemma 6.1 suggests that, even though the manager is able to utilize compensation as a control lever to incite agent participation, she need not use it to match demand and supply in a given shift. In other words, "surge pricing" need not always be used to match supply and demand, as is the current viewpoint with e.g., ride-sharing platforms (The Economist 2016). Since excessive surges in prices usually generate bad press [17], it is insightful that intentionally setting a lower price than needed to match supply and demand may be optimal. This also lends support to recent calls for ride-sharing services to set caps on the prices that they charge their customers, at the expense of potentially inciting fewer drivers to be on the road (Harvard Business Review 2015).

Delay announcements. For tractability, we assume that the announcement-dependent abandonment rate is constant and equal to $\tilde{\theta} > \theta$, where θ is the rate without the announcements. We denote $\tilde{L}_k \equiv p_k + h_k/\tilde{\theta}$ and note that $\tilde{L}_k < L_k \equiv h_k + p_k/\theta$. Thus, it is optimal for the manager to make announcements in every overloaded shift, since doing so would reduce the cost of congestion in that shift. While it is clear that making announcements is beneficial to the manager in that case, it is unclear whether agents will be better or worse off because of the announcements. We now investigate this issue by investigating, for fixed n, the optimal compensation in a shift where the manager is allowed to make delay announcements. Since the announcements are only relevant when the system is congested, we focus on case (c) in Lemma 6.1, i.e., we assume that $n < \lambda_k/G(a_k)$ for a_k in (6.13). We let $\tilde{c}_k^*$ denote the optimal compensation in shift k, assuming that the manager makes announcements in that shift; i.e., $\tilde{c}_k^*$ minimizes $c_k n G(c_k) + \tilde{L}_k \left(\lambda_k - nG(c_k)\right)^+$.

Lemma 6.2. *If $n < \lambda_k/G(a_k)$ for a_k in (6.13), then $\tilde{c}_k^* = \tilde{a}_k < c_k^* = a_k$ where c_k^* is the optimal compensation for the no-announcement problem in (6.12) and $\tilde{a}_k$ is given by*

$$G(\tilde{a}_k)\left(1 + (\tilde{a}_k - \tilde{L}_k)\frac{g(\tilde{a}_k)}{G(\tilde{a}_k)}\right) = 0; \tag{6.14}$$

i.e., agents are worse off because of the announcements.

Lemma 6.2 shows that a manager who uses the announcements to reduce congestion in her system would have less incentive to offer agents a higher compensation to induce their participation. Indeed, she uses the announcements to disincentivize customer waiting, thereby relieving congestion instead.

[17] https://www.nytimes.com/2014/01/12/magazine/is-ubers-surge-pricing-an-example-of-high-tech-gouging.html?_r=1

6.2 Jointly Optimizing All Controls

We now study the manager's problem when she can jointly optimize the staffing level, the compensation offered to her agents, and whether or not to make announcements in any given shift. In the interest of analytical tractability, we continue to assume a constant abandonment rate in response to the announcements, $\tilde{\theta} > \theta$, so that the manager will make announcements in every congested shift. Here is the manager's problem:

$$\min_{c_j \ \geq l, n\in\mathbb{N}} \Pi(n, \mathbf{c}) \equiv \sum_{1\leq j\leq k} \Big(c_j \cdot nG(cj) + \tilde{L}_j(\lambda_j - nG(c_j))^+ \Big), \tag{6.15}$$

where as before $\tilde{L}_j \equiv p_j + h_j/\tilde{\theta}$ is the adjusted congestion cost which accounts for the effect of the announcements. To better position our results, we recall that when capping agents is allowed, the optimal compensation is set equal to the minimum wage in all shifts (Gurvich et al. 2017), irrespective of the value of that wage, and the staffing level high enough to match demand in the highest-demand shift (with the offered minimum wage). In our context, we find that this is no longer the case. Indeed, the optimal compensation depends on the *value* of the minimum wage, in particular whether it is "low" or "high", the manager may offer higher compensation than the minimum wage in some shifts, and may still either understaff or overstaff some shifts. This lends support to recent practices in some sharing-economy platforms which set a minimum compensation to agents that is larger than the minimum wage, e.g., as for TaskRabbit [18]. We begin by establishing the existence and uniqueness of the solution to (6.15).

6.2.1 Existence and Uniqueness of the Solution to (6.15)

We have characterized the optimal solution for the compensation when the staffing level is fixed (Lemma 6.1). To solve problem (6.15), we can make use of the results of that lemma. Indeed, the

[18] `https://newrepublic.com/article/120378/wonolo-temp-worker-app-shows-scary-future-sharing-economy`

optimal agent pool size is the solution to the following problem:

$$\begin{aligned}\min_{n\geq 0}\Pi(n) \;\equiv\; & \sum_{\{k\,:\,n\geq\frac{\lambda_k}{G(l)}\}} nlG(l) \qquad \text{(overstaffed)} \\ + & \sum_{\{k\,:\,\frac{\lambda_k}{G(a_k)}\leq n<\frac{\lambda_k}{G(l)}\}} \lambda_k G^{-1}\left(\frac{\lambda_k}{n}\right) \qquad \text{(supply and demand matched)} \\ + & \sum_{\{k\,:\,n<\frac{\lambda_k}{G(\tilde{a}_k)}\leq\frac{\lambda_k}{G(l)}\}} \tilde{a}_k G(\tilde{a}_k)n + \tilde{L}_k(\lambda_k - nG(\tilde{a}_k)) \qquad \text{(undestaffed/ announcements)},\end{aligned} \tag{6.16}$$

where $\tilde{a}_k$ is given in (6.14) and some intervals for n in (6.16) may be empty. Next, we show that there exists a unique solution to (6.16). Based on this optimal staffing level and the optimal compensation in Lemma 6.1, we can derive the optimal solution to problem (6.15).

Lemma 6.3. *The objective $\Pi(n)$ in (6.16) is piecewise convex, and $\Pi'(n)$ is strictly increasing in n so that $\Pi(n)$ is strictly convex in n. Thus, there exists a unique solution n^* to (6.16). For that n^*, the optimal compensation c_k^* for each shift k is given in Lemma 6.1.*

In general, the solution to problem (6.16) is algebraically tedious to characterize in our multi-shift setting. To formulate meaningful insights, we focus next on two special cases: (i) when the minimum wage is sufficiently low, and (ii) when the minimum wage is sufficiently high.

6.2.2 Low Minimum Wage

We begin by considering the case where the minimum wage is "sufficiently low". We define:

$$l_0 = G^{-1}\left(\frac{\min_i\{\lambda_i\}}{\max_i\{\frac{\lambda_i}{G(\tilde{a}_i)}\}}\right) \quad \text{where} \quad \tilde{a}_k \text{ is given in (6.14).} \tag{6.17}$$

Then, the following lemma holds for $l < l_0$.

Lemma 6.4. *If the minimum wage is sufficiently low, in particular $l < l_0$ in (6.17), then $n^*G(c_k^*) \geq \lambda_k$ for all k, i.e., all shifts are either overstaffed or have matched supply and demand. Moreover, there exists at least one shift i_0 where $c_{i_0}^* = G^{-1}(\lambda_{i_0}/n^*) > l$ where demand and supply are matched.*

Lemma 6.4 shows that the manager need not always resort to using the announcements. In particular, if the minimum wage is "low enough", then she will staff a large enough pool and offer high enough compensations so that, in every shift i, the supply $n^*G(c_i^*)$ is at least as large as the demand λ_i. Moreover, she will offer a compensation that is strictly higher than the minimum wage in at least one of the shifts (with highest demand rates). Intuitively, because the minimum wage is small, the manager is less restricted in the compensation that she has to pay her agents. Therefore, she can afford to staff a larger pool and eliminate congestion in her system. This also explains why she is then able to pay her agents a compensation which is strictly larger than the minimum wage. Because no shift is congested, the manager does not resort to making delay announcements.

6.2.3 High Minimum Wage

We now explore the case where the minimum wage is "sufficiently high". In particular, we assume that $\tilde{a}_i < l < \tilde{L}_i$ for all i. We now show that the manager would then make announcements.

Lemma 6.5. *If the minimum wage is sufficiently high, $l > \bar{l} \equiv \max_{1 \leq i \leq k} \tilde{a}_i$ where $\tilde{a}_i$ is given in (6.14), then $c_i^* = l$ for all i. Moreover, there exists a shift k_0 such that $n^* = \lambda_{k_0}/G(l)$, i.e., supply and demand are matched in shift k_0. All shifts k for which $\lambda_k > \lambda_{k_0}$ are understaffed, and all shifts for which $\lambda_k < \lambda_{k_0}$ are overstaffed; announcements are made in every congested shift.*

Lemma 6.5 shows that the manager must set compensation equal to the minimum wage in every shift, if that minimum wage is sufficiently high. In this case, the manager must staff a smaller agent pool (because it would be too costly to employ many agents), and she will use the announcements to alleviate congestion in understaffed periods. The fact that the manager consistently compensates agents at the minimum wage and operates some shifts overloaded lends support to current practices in virtual call centers [19] where agent compensation is typically set at the minimum wage and there is congestion in shifts which experience peaks in customer demand.

7 Asymptotic Accuracy of the Fluid Approximation

In this section, we prove the asymptotic accuracy of the fluid approximation in (3.2) with a binomially distributed number of servers; this distribution arises when the servers make independent decisions to join the different shifts. We also restrict attention to exponentially distributed service times, and a Poisson arrival process. Conditional on the number of servers in a shift, the queueing

[19] https://www.thespruce.com/how-home-call-centers-pay-3542389

dynamics in different shifts are independent. Thus, to establish the desired asymptotic accuracy, it suffices to focus on a single shift instead. The proof for multiple shifts can then be obtained by a simple argument, exploiting a similar conditioning argument as the one that we use in what follows, along with the conditional independence across shifts. In this section, for clarity of exposition, we consider a single-shift setting.

In this section, we prove the asymptotic accuracy of the fluid approximation in (3.2) with a binomially distributed number of servers; this distribution arises when the servers make independent decisions to join the different shifts. We also restrict attention to exponentially distributed service times, and a Poisson arrival process. Conditional on the number of servers in a shift, the queueing dynamics in different shifts are independent. Thus, to establish the desired asymptotic accuracy, it suffices to focus on a single shift instead. The proof for multiple shifts can then be obtained by a simple argument, exploiting a similar conditioning argument as the one that we use in what follows, along with the conditional independence across shifts. In this section, for clarity of exposition, we consider a single-shift setting.

We consider a sequence of queueing models indexed by the arrival rate λ, and study system performance as λ increases without bound. The number of servers in the λ^{th} system is $N_\lambda \sim Bin(n_\lambda, r)$. We assume that $\rho \equiv \lambda/\mathbb{E}[N_\lambda] = \lambda/rn_\lambda$ remains fixed as λ increases. Let Q_{N_λ} denote the steady-state queue length and α_{N_λ} the net customer abandonment rate in the $M/M/N_\lambda + GI$ queue (abandonment makes the system stable). We refer to the cases with $\rho > 1$, $\rho < 1$, and $\rho = 1$ as the overloaded, underloaded, and critically loaded regimes, respectively. Since N_λ is random, an $M/M/N_\lambda + GI$ system with e.g., $\rho > 1$ may or may not be overloaded, i.e., having $\lambda > N_\lambda$.

Theorem 7.1. *Consider an $M/M/N_\lambda + GI$ queueing model with $N_\lambda \sim Bin(n_\lambda, r)$,*

(a) *If $\rho > 1$ (overloaded regime), then there exists a finite constant $K > 0$ such that*

$$\limsup_{\lambda\to\infty} |\mathbb{E}[Q_{N_\lambda}] - rn_\lambda \bar{q}_\rho| \le K \textit{ and } \lim_{\lambda\to\infty} |\mathbb{E}[\alpha_{N_\lambda}] - rn_\lambda \bar{\alpha}_\rho| \to 0.$$

(b) *If $\rho = 1$ (critically-loaded regime), then there exist finite constants $K_1', K_2' > 0$ such that*

$$\limsup_{\lambda\to\infty} \mathbb{E}[Q_{N_\lambda}] \le K_1'\sqrt{\lambda} \textit{ and } \limsup_{\lambda\to\infty} \mathbb{E}[\alpha_{N_\lambda}] \le K_2'\sqrt{\lambda}.$$

(c) *If $\rho < 1$ (underloaded regime), then*

$$\lim_{\lambda \to \infty} \mathbb{E}[Q_{N_\lambda}] \to 0 \text{ and } \lim_{\lambda \to \infty} \mathbb{E}[\alpha_{N_\lambda}] \to 0.$$

Theorem 7.1 shows that, in the overloaded system, the fluid approximation for the expected queue length is asymptotically accurate up to $\mathcal{O}(1)$ [20], and the fluid approximation for the net abandonment rate is asymptotically accurate up to $o(1)$, i.e., the corresponding error is asymptotically bounded in the former case, and it decreases with the arrival rate in the latter case. In other words, fluid approximations are "extremely accurate" in the overloaded regime. In the critically-loaded system, those fluid-approximation errors are $\mathcal{O}(\sqrt{\lambda})$, i.e., they grow in the square-root of the size of the system. In the underloaded regime, fluid approximations are $o(1)$-accurate since errors for both performance measures decrease with the arrival rate.

While our discussion in Theorem 7.1 is split according to the asymptotic regime prescribed, an alternative approach would be to resort to regime-free universal approximations, as in Gurvich et al. (2013) and Huang and Gurvich (2016). Since such treatment lies outside the scope of this paper, we do not discuss this point further here. The following theorem establishes the asymptotic accuracy of fluid-based staffing prescriptions, by exploiting the results of Theorem 7.1; the proof proceeds along similar lines as Theorem 3 in Bassamboo and Randhawa (2010).

Theorem 7.2. *The fluid-based prescription, n_λ^* is asymptotically optimal in the overloaded, critically-loaded and underloaded regimes in the sense that*

$$\lim_{\lambda \to \infty} \frac{\Pi_\lambda^*}{\Pi_\lambda(n_\lambda^*)} = 1,$$

where Π_λ^* is the optimal objective value for (3.1) and $\Pi_\lambda(n_\lambda^*)$ is the value of its objective evaluated at n_λ^*. If, in addition, n_λ^* is such that the system is overloaded, then there exists $K'' > 0$ such that

$$\limsup_{\lambda \to \infty} |\Pi_\lambda^* - \Pi_\lambda(n_\lambda^*)| \leq K'',$$

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

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

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

i.e., the fluid staffing prescription is asymptotically $\mathcal{O}(1)$-accurate in the overloaded regime.

Intuitively, because the binomial random variable "concentrates" around its mean asymptotically, so that fluctuations in the number of servers are asymptotically negligible, intuitions similar to those in Bassamboo and Randhawa (2010), who consider a deterministic number of servers, continue to hold in our setting. In particular, in the overloaded case, stochastic fluctuations are better explained by large deviations theory. Thus, fluid approximations are practically indistinguishable from the estimates for, e.g., average queue-lengths. This translates into the $\mathcal{O}(1)$ accuracy for the fluid-based staffing prescriptions in Theorem 7.2. In the critically-loaded regime, stochastic fluctuations are consistent with those suggested by the central limit theorem, i.e., they are on the order of $\sqrt{\lambda}$. In other words, they are also asymptotically negligible since the magnitude of the optimal objective in our original problem is $\mathcal{O}(\lambda)$: This is because the staffing cost is linear in the staffing pool, and the staffing pool size itself is $\mathcal{O}(\lambda)$; this is the first part of Theorem 7.2. In the online supplement, we describe the results of some numerical experiments validating our asymptotic results.

8 Conclusions

In this paper, we studied the problem of staffing and controlling large-scale service systems with a random number of servers. This randomness arises with strategic agent behavior, e.g., as with self-scheduling agents in virtual call centers. Our asymptotic accuracy results support the usefulness of fluid approximations in this context. Because of congestion in understaffed shifts, the customer abandonment distribution plays an important role. We characterized this role, and proposed using delay announcements as a control of customer abandonment behavior; this is especially useful in settings where other tools, such as pricing, cannot be used to influence customers. We studied the optimization of three controls (staffing, compensation, and the announcements), and found that the optimal control policy may be non-standard, in that it may be optimal to match supply and demand in one of the shifts, or to not do so in any shift, effectively understaffing and overstaffing all shifts. This suggests that managers may need, in this new context, to shift from the traditional paradigm of workforce management, or from the current viewpoint dominating sharing-economy applications, where it is believed that matching supply and demand, at least in first order, is always desirable.

The current growth of the sharing economy has motivated several recent papers in the academic literature. Nevertheless, further exploration of the dynamics of such systems remains of interest for future research. In this paper, we explored the staffing and control question using a fluid approximation. There remains to establish supporting many-server heavy-traffic limits for different stochastic processes in the system, such as the queue length (corresponding to a functional law of large numbers). Such an investigation would lead to a deeper understanding of the system's dynamics. Several modelling extensions (multiplicity of customer classes, time-variability in the demand rates, etc.) remain to be explored. Finally, alternative system design questions, e.g., optimal priority rules when both the available capacity and incoming demand are random, as is the case with two-sided platforms in the sharing economy, would be interesting to explore as well.

9 Acknowledgments

The author gratefully acknowledges Vasiliki Kostami, Yunan Liu, Amy Ward, Dongyuan Zhan, three referees, the senior editor, and the department editor for their useful feedback on earlier versions of the manuscript.

References

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.

Aksin, Z., B. Ata, S. Emadi, C. Su. 2016. Impact of delay announcements in call centers: An empirical approach. *Operations Research* .

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.

Allon, Gad, Achal Bassamboo. 2011. The impact of delaying the delay announcements. *Operations research* **59**(5) 1198–1210.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.

Ata, B., D. Lee, E. Sonmez. 2017. Dynamic staffing of volunteer gleaning operations. University of Chicago, working paper.

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

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., M. J. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research* **54**(3) 419–435.

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.

BBC. 2015. Hospital staff absences for mental health reasons double. URL `http://www.bbc.co.uk/news/uk-england-32022114`.

Bimpikis, Kostas, Ozan Candogan, Saban Daniela. 2017. Spatial pricing in ride-sharing networks. Stanford University, working paper.

Bolandifar, E., N DeHoratius, T. L. Olsen, J. Wiler. 2016. Modeling the behavior of patients who leave the ed without being seen The Chinese University of Hong Kong , working paper.

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

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.

Bum Soh, Seung, Itai Gurvich. 2017. Duality in staffing problems: Between holding costs and waiting constraints .

Cachon, G., K. Daniels, R. Lobel. 2016. 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.

Dong, Jing, Rouba Ibrahim. 2017. Flexible workers or full-time employees? on staffing service systems with a blended workforce. Northwestern University, working paper.

Feng, Guiyun, Guangwen Kong, Zizhuo Wang. 2017. We are on the way: Analysis of on-demand ride-hailing systems. University of Minnesota, working paper.

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, R., S. Doroudi, A. Ward, A. Wierman. 2016. Routing and staffing when servers are strategic. *arXiv preprint arXiv:1402.3606* .

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.

Gurvich, Itai, Junfei Huang, Avishai Mandelbaum. 2013. Excursion-based universal approximations for the erlang-a queue in steady-state. *Mathematics of Operations Research* **39**(2) 325–373.

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. 2015. Everyone hates uberâĂŹs surge pricing âĂŞ hereâĂŹs how to fix it. `https://hbr.org/2015/12/everyone-hates-ubers-surge-pricing-heres-how-to-fix-it`. Accessed: 2017-03-12.

Hu, Ming, Yiwei Chen. 2017. Pricing and matching with forward-looking buyers and sellers. University of Toronto, working paper.

Hu, Ming, Yun Zhou. 2017a. Dynamic type matching. University of Toronto, working paper.

Hu, Ming, Yun Zhou. 2017b. Price, wage and fixed commission in on-demand matching. University of Toronto, working paper.

Huang, Junfei, Itai Gurvich. 2016. Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue .

Huang, Junfei, Avishai Mandelbaum, Hanqin Zhang, Jiheng Zhang. 2017. Refined models for efficiency-driven queues with applications to delay announcements and staffing Technion, working paper.

Ibrahim, Rouba, Mor Armony, Achal Bassamboo. 2017. Does the past predict the future? the case of delay announcements in service systems. *Management Science* .

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

Jouini, Oualid, Zeynep Aksin, Yves Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* **13**(4) 534–548.

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., S. Zeltyn. 2013. Data-stories about (im) patient customers in tele-queues. *Queueing Systems* **75**(2-4) 115–146.

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

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

Riquelme, C., S. Banerjee, R. Johari. 2016. Pricing in ride-share platforms: A queueing-theoretic approach. Cornell University, working paper.

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

Steckley, Samuel G, Shane G Henderson, Vijay Mehrotra. 2005. Performance measures for service systems with a random arrival rate. *Proceedings of the 37th conference on Winter simulation*. Winter Simulation Conference, 566–575.

TalkDesk. 2014. 10 causes of low service level in the call center. URL `https://www.talkdesk.com/blog/10-causes-of-low-service-level-in-the-call-center/`.

Tang, Christopher S, Jiaru Bai, Kut C So, Xiqun Michael Chen, Hai Wang. 2017. Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio. University of California, Los Angeles, working paper.

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

The Economist. 2016. A fare shake. `http://www.economist.com/news/finance-and-economics/21698656-jacking-up-prices-may-not-be-only-way-balance-supply-and-demand-taxis`. Accessed: 2017-03-12.

Uber. 2015. What is surge pricing? URL `https://help.uber.com/h/6c8065cf-5535-4a8b-9940-d292ffdce119`.

Uber. 2017. What is surge? `https://help.uber.com/h/e9375d5e-917b-4bc5-8142-23b89a440eec`. Accessed: 2017-03-12.

US Bureau of Labor Statistics. 2008. Industry injury and illness data. URL `http://www.bls.gov/iif/oshwc/osh/os/osnr0032.pdf`.

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.

Yu, Q., G. Allon, A. Bassamboo. 2016. How do delay announcements shape customer behavior? an empirical study. *Management Science* .

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. 2016. Compensation and staffing to trade off speed and quality in large service systems University of Southern California, working paper.

TECHNICAL APPENDIX

10 Asymptotic Accuracy of the Fluid Approximation

10.1 The Overloaded Regime

10.1.1 $\mathcal{O}(1)$-Accuracy for the Fluid Queue Length.

We begin by establishing the asymptotic $\mathcal{O}(1)$-accuracy for the expected queue length. Let $0 < \epsilon < r$ and define $k_1 \equiv r - \epsilon$ and $k_2 \equiv r + \epsilon$. Assume that ϵ is small enough so that $\rho r/(r+\epsilon) > 1$. Denote $\mathbb{E}[Q_{N_\lambda}|N_\lambda = s] \equiv \mathbb{E}[Q_s]$ where Q_s is the steady-state queue length in the corresponding $M/M/s{+}GI$ queue with the same arrival rate.

Conditioning and unconditioning on N_λ. Conditioning on N_λ, we can write:

$$\begin{aligned}
|\mathbb{E}[Q_{N_\lambda}] - rn_\lambda\bar{q}_\rho| &= \left|\sum_{s\geq 0}\mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) - rn_\lambda\bar{q}_\rho\right| \\
&= \left|\sum_{s\geq 0}(\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s)\right| \qquad \text{since } \mathbb{E}[N_\lambda] = rn_\lambda = \sum_{s\geq 0} s\mathbb{P}(N_\lambda = s), \\
&\leq \left|\sum_{s<k_1n_\lambda \text{ or } s>k_2n_\lambda}(\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s)\right| + \left|\sum_{k_1n_\lambda\leq s\leq k_2n_\lambda}(\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s)\right|.
\end{aligned}$$

We now turn to establishing asymptotic bounds for A_λ and B_λ, defined as follows:

$$A_\lambda \equiv \left|\sum_{s<k_1n_\lambda \text{ or } s>k_2n_\lambda}(\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s)\right| \text{ and } B_\lambda \equiv \left|\sum_{k_1n_\lambda\leq s\leq k_2n_\lambda}(\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s)\right|.$$

Asymptotic bound for N_λ far from $n_\lambda r$. We begin by showing that A_λ is asymptotically negligible.

Lemma 10.1. $\lim_{\lambda\to\infty} A_\lambda = 0$.

Proof. We can write,

$$\begin{aligned}
A_\lambda &= \left|\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}\mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) - \sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda} s\bar{q}_\rho\mathbb{P}(N_\lambda = s)\right|, \\
&\leq \mathbb{E}[Q_0]\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}\mathbb{P}(N_\lambda = s) + \sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda} s\bar{q}_\rho\mathbb{P}(N_\lambda = s).
\end{aligned}$$

Also, define $A_\lambda^{(1)} \equiv \mathbb{E}[Q_0]\sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda}\mathbb{P}(N_\lambda = s)$ and $A_\lambda^{(2)} \equiv \sum_{s>k_2n_\lambda \text{ or } s<k_1n_\lambda} s\bar{q}_\rho\mathbb{P}(N_\lambda = s)$. Note that Q_0 has the same distribution as the steady-state number in the system in an $M/GI/\infty$ model with Poisson arrivals at rate $\lambda = rn_\lambda\rho$ and i.i.d. generally distributed service times having the same distribution, F, as the abandonment times in our original model. Therefore, exploiting standard results for the infinite-server queue, Q_0 has a Poisson distribution with mean

$\lambda/\theta = rn_\lambda\rho/\theta$, i.e., $\mathbb{E}[Q_0] = \mathcal{O}(\lambda)$. Applying Hoeffding's inequality to the binomial distribution: $\mathbb{P}\left(k_1 n_\lambda \leq N_\lambda \leq k_2 n_\lambda\right) \geq 1 - 2e^{-2\epsilon^2 n_\lambda}$; equivalently, $\mathbb{P}\left(k_1 n_\lambda > N_\lambda \text{ or } N_\lambda > k_2 n_\lambda\right) \leq 2e^{-2\epsilon^2 n_\lambda}$. Thus,

$$A_\lambda^{(1)} = \mathbb{E}[Q_0] \sum_{s > k_2 n_\lambda \text{ or } s < k_1 n_\lambda} \mathbb{P}(N_\lambda = s) = \mathbb{E}[Q_0] \cdot \mathbb{P}\left(k_1 n_\lambda > N_\lambda \text{ or } N_\lambda > k_2 n_\lambda\right) \to 0 \text{ as } \lambda \to \infty.$$

We now turn to showing that $A_\lambda^{(2)}$ is asymptotically negligible as well. Note that:

$$A_\lambda^{(2)} = \bar{q}_\rho \sum_{s > k_2 n_\lambda \text{ or } s < k_1 n_\lambda} s\mathbb{P}(N_\lambda = s) = \bar{q}_\rho \mathbb{E}[N_\lambda \mathbb{1}\{N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda\}],$$

where $\mathbb{1}\{\cdot\}$ denotes an indicator random variable. By the Cauchy-Schwarz inequality:

$$\begin{aligned}\mathbb{E}[N_\lambda \mathbb{1}\{N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda\}] &\leq \sqrt{\mathbb{E}[N_\lambda^2]\mathbb{P}(N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda)} \\ &= \sqrt{(n_\lambda r(1-r) + n_\lambda^2 r^2)\mathbb{P}(N_\lambda > k_2 n_\lambda \text{ or } N_\lambda < k_1 n_\lambda)} \to 0 \text{ as } \lambda \to \infty.\end{aligned}$$

Therefore, $A_\lambda^{(2)} \to 0$ as $\lambda \to \infty$. Combining the above, we obtain that $A_\lambda \to 0$ as well.

Asymptotic bound for N_λ close to $n_\lambda r$. We now characterize B_λ for large λ.

Lemma 10.2. *There exists a finite constant $C > 0$ such that* $\limsup_{\lambda\to\infty} B_\lambda \leq C$.

Proof. We begin by writing B_λ as follows,

$$B_\lambda \leq \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} \left|\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}\right| \mathbb{P}(N_\lambda = s) + \left| \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} s(\bar{q}_{\rho_s} - \bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right|, \tag{10.18}$$

where $\rho_s \equiv n_\lambda r\rho/s$ and $\bar{q}_{\rho_s}$ is the fluid limit for the queue length in the $M/M/s + GI$ queue with traffic intensity ρ_s (the arrival rate is $\lambda = rn_\lambda\rho$ and the number of servers is s). Let,

$$B_\lambda^{(1)} \equiv \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|\mathbb{P}(N_\lambda = s) \text{ and } B_\lambda^{(2)} \equiv \left| \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} s(\bar{q}_{\rho_s} - \bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right|.$$

First, we consider $B_\lambda^{(1)}$ and show that it is asymptotically bounded. Fix n_λ and note that to each $k_1 n_\lambda \leq s \leq k_2 n_\lambda$ corresponds a traffic intensity ρ_s in the $M/M/s + GI$ system, where $\rho_s = n_\lambda r\rho/s$ and $1 < \rho r/(r+\epsilon) \leq \rho_s \leq \rho r/(r-\epsilon)$. By Theorem 5 of Bassamboo and Randhawa (2010), assuming that f is strictly positive and continuously differentiable,

$$\limsup_{\lambda\to\infty} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \leq \sqrt{f(\bar{w}_{\rho_s})} \left(\frac{3|f'(\bar{w}_{\rho_s})|}{\rho_s f^2(\bar{w}_{\rho_s})} + 1/2 \right), \tag{10.19}$$

where $\bar{w}_{\rho_s}$ is the fluid limit for the steady-state waiting time in the overloaded $M/M/s + GI$ queue with traffic intensity ρ_s. Note that for $\rho r/(r+\epsilon) \leq \rho_s \leq \rho r/(r-\epsilon)$, we have that $\bar{w}_{\rho r/(r+\epsilon)} \leq \bar{w}_{\rho_s} \leq \bar{w}_{\rho r/(r-\epsilon)}$. By the continuity of the bounding function in (10.19) and the boundedness theorem, we conclude that there exists a finite constant $C_1 > 0$ such that

$$\sup_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} \sqrt{f(\bar{w}_{\rho_s})} \left(\frac{3|f'(\bar{w}_{\rho_s})|}{\rho' f^2(\bar{w}_{\rho_s})} + 1/2 \right) \leq C_1. \tag{10.20}$$

Since $B_\lambda^{(1)} = \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \mathbb{P}(N_\lambda = s) \le \sup_{k_1 n_\lambda \le s \le k_2 n_\lambda} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} \mathbb{P}(N_\lambda = s) \le \sup_{k_1 n_\lambda \le s \le k_2 n_\lambda} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|$, combining (10.19) and (10.20) yields that $\limsup_{\lambda\to\infty} B_\lambda^{(1)} \le C_1$ by taking limits on both sides. There remains to study the asymptotic behaviour of $B_\lambda^{(2)}$. Note that $\bar{q}_{\rho_s} = \rho_s \int_0^{(\bar{F})^{-1}(1/\rho_s)} \bar{F}(u)\,\mathrm{d}u$, e.g., by equations (3.6) and (3.7) in Whitt (2006a). Consider,

$$\begin{aligned}
&\left| \sum_{s\ge 0} s \left(\rho_s \int_0^{(\bar{F})^{-1}(1/\rho_s)} \bar{F}(x)\,\mathrm{d}x - \rho \int_0^{(\bar{F})^{-1}(1/\rho)} \bar{F}(u)\,\mathrm{d}u \right) \mathbb{P}(N_\lambda = s) \right| \\
&= \left| \sum_{s\ge 0} \left(n_\lambda r\rho \int_0^{(\bar{F})^{-1}(s/n_\lambda r\rho)} \bar{F}(u)\,\mathrm{d}u - s\rho \int_0^{(\bar{F})^{-1}(1/\rho)} \bar{F}(u)\,\mathrm{d}u \right) \mathbb{P}(N_\lambda = s) \right|, \\
&= \left| \mathbb{E}\left[\left(n_\lambda r\rho \int_0^{(\bar{F})^{-1}(N_\lambda/n_\lambda r\rho)} \bar{F}(u)\,\mathrm{d}u - N_\lambda\rho \int_0^{(\bar{F})^{-1}(1/\rho)} \bar{F}(u)\,\mathrm{d}u \right) \right] \right|, \\
&= \left| n_\lambda\rho r \mathbb{E}\left[\left(\int_{(\bar{F})^{-1}(1/\rho)}^{(\bar{F})^{-1}(N_\lambda/n_\lambda r\rho)} \bar{F}(u)\,\mathrm{d}u \right) \right] \right|.
\end{aligned}$$

We now show that there must exist a finite constant $C_2 > 0$ such that

$$\left| n_\lambda\rho r \mathbb{E}\left[\left(\int_{(\bar{F})^{-1}(1/\rho)}^{(\bar{F})^{-1}(N_\lambda/n_\lambda r\rho)} \bar{F}(u)\,\mathrm{d}u \right) \right] \right| \le C_2$$

for λ large enough. To this aim, define the function

$$g_\lambda(x) = n_\lambda\rho r \int_{(\bar{F})^{-1}(1/\rho)}^{(\bar{F})^{-1}(x/n_\lambda r\rho)} \bar{F}(u)\,\mathrm{d}u \text{ for } x \ge 0.$$

For a given λ, we use a Taylor-series expansion of $\mathbb{E}[g_\lambda(N_\lambda)]$ around $\mathbb{E}[N_\lambda] = n_\lambda r$ (we can do this since g_λ is sufficiently differentiable and the moments of N_λ are finite):

$$|\mathbb{E}[g_\lambda(N_\lambda)]| = \left| \mathbb{E}\left[g_\lambda(n_\lambda r) + g'_\lambda(n_\lambda r)\,(N_\lambda - n_\lambda r) + \frac{1}{2} g''_\lambda(n_\lambda r)(N_\lambda - r n_\lambda)^2 \right] \right| + \mathcal{O}(1/\lambda).$$

Indeed, by computing the centralized moments of N_λ and higher-order derivatives of g_λ, it can be shown that the remainder term in the Taylor series is $\mathcal{O}(1/\lambda)$. Also, $g_\lambda(n_\lambda r) = 0$ and

$$g'_\lambda(n_\lambda r) = -\frac{1/\rho}{f\left(\bar{F}^{-1}(1/\rho)\right)} \text{ and } g''_\lambda(n_\lambda r) = -\frac{1}{r n_\lambda \rho} \frac{h_1(\rho) + (1/\rho)h_2(\rho)/h_1(\rho)}{h_1^2(\rho)},$$

where $h_1(\rho) = f(\bar{F}^{-1}(1/\rho))$ and $h_2(\rho) = f'(\bar{F}^{-1}(1/\rho))$. Thus, there exists $C_2 > 0$ such that:

$$|\mathbb{E}[g_\lambda(N_\lambda)]| \approx |\frac{1}{2} g''_\lambda(n_\lambda r) n_\lambda r(1-r)| \le C_2 \text{ for } \lambda \text{ large enough.}$$

We now turn to the asymptotic behaviour of $B_\lambda^{(2)}$. Note that:

$$B_\lambda^{(2)} = |\mathbb{E}[g_\lambda(N_\lambda) \mathbb{1}\{N_\lambda \in [k_1 n_\lambda, k_2 n_\lambda]\}]|\,, \text{ and}$$

$$|\mathbb{E}[g_\lambda(N_\lambda)]| = |\mathbb{E}[g_\lambda(N_\lambda) \mathbb{1}\{N_\lambda \in [k_1 n_\lambda, k_2 n_\lambda]\}] + \mathbb{E}[g_\lambda(N_\lambda) \mathbb{1}\{N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]\}]|\,.$$

Bounding the second term in the last equality,

$$
\begin{aligned}
\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]\}] &\leq |\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]\}]| \\
&\leq \sqrt{\mathbb{E}[g_\lambda^2(N_\lambda)]}\mathbb{P}(N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda]) \quad \text{(Cauchy Schwarz inequality)} \\
&\to 0,
\end{aligned}
$$

since $\mathbb{P}(N_\lambda \notin [k_1 n_\lambda, k_2 n_\lambda])$ vanishes exponentially fast as $\lambda \to \infty$, and $\mathbb{E}[g_\lambda^2(N_\lambda)] = \mathcal{O}(\lambda^2)$ since $\int_{(\bar{F})^{-1}(1/\rho)}^{(\bar{F})^{-1}(N_\lambda/n_\lambda r\rho)} \bar{F}(u)\, du \leq 1/\theta$. Thus, $\limsup_{\lambda\to\infty} B_\lambda^{(2)} = \limsup_{\lambda\to\infty}|\mathbb{E}[g_\lambda(N_\lambda)\mathbb{1}\{N_\lambda \in [k_1 n_\lambda, k_2 n_\lambda]\}]| \leq C_2$. Combining the above, there exists $C > 0$ such that $\limsup_{\lambda\to\infty} B_\lambda \leq C$. ■

$\mathcal{O}(1)$-accuracy. Since both A_λ and B_λ are asymptotically bounded, there must exist $K > 0$ such that, as desired:

$$
\limsup_{\lambda\to\infty} |\mathbb{E}[Q_{N_\lambda}] - rn_\lambda \bar{q}_\rho| \leq K.
$$

10.1.2 $o(1)$-Accuracy for the Fluid Net Abandonment Rate.

The proof for the net abandonment rate proceeds along similar lines, so we will be brief. Paralleling (10.19), and denoting $\mathbb{E}[\alpha_{N_\lambda}|N_\lambda = s] \equiv \mathbb{E}[\alpha_s]$, we can exploit Theorem 5 in Bassamboo and Randhawa (2010) to show that $\sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda}(\mathbb{E}[\alpha_s] - s\bar{\alpha}_{\rho_s})\mathbb{P}(N_\lambda = s) \to 0$ as $\lambda \to \infty$. Moreover, by equation (3.3) in Whitt (2006a): $\bar{\alpha}_{\rho_s} = \rho_s - 1$; thus, $s(\bar{\alpha}_{\rho_s} - \bar{\alpha}_\rho) = \rho(n_\lambda r - s)$. We can then write:

$$
\sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} s(\bar{\alpha}_{\rho_s} - \bar{\alpha}_\rho)\mathbb{P}(N_\lambda = s) = \rho\mathbb{E}[(nr - N_\lambda)\mathbb{1}(k_1 n_\lambda \leq N_\lambda \leq k_2 n_\lambda)],
$$

and deduce that $\mathbb{E}[(nr - N_\lambda)\mathbb{1}(k_1 n_\lambda \leq N_\lambda \leq k_2 n_\lambda)] \to 0$ since $\mathbb{E}[N_\lambda] = rn_\lambda$.

10.2 The Underloaded Regime

Let $0 < \epsilon < r$ be small enough so that $\rho r/(r - \epsilon) < 1$, and recall that $k_1 \equiv r - \epsilon$ and $k_2 \equiv r + \epsilon$. Then, conditioning on N_λ:

$$
\begin{aligned}
\mathbb{E}[Q_{N_\lambda}] &= \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \sum_{k_1 n_\lambda > s \text{ or } s > k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s), \\
&\leq \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \mathbb{E}[Q_0] \sum_{k_1 n_\lambda > s \text{ or } s > k_2 n_\lambda} \mathbb{P}(N_\lambda = s).
\end{aligned}
$$

As in the proof of Theorem 7.1, we can show that: $\mathbb{E}[Q_0]\sum_{k_1 n_\lambda > s \text{ or } s > k_2 n_\lambda} \mathbb{P}(N_\lambda = s) \to 0$ as $\lambda \to \infty$. Also, $\sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) \leq \mathbb{E}[Q(k_1 n_\lambda)]\sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} \mathbb{P}(N_\lambda = s)$. Since $\mathbb{E}[Q(k_1 n_\lambda)]$ is the expected steady-state queue length in an underloaded queue, it converges to 0 as $\lambda \to \infty$, e.g, see Theorem 5.1 in Zeltyn and Mandelbaum (2005). The limit for the net abandonment follows similarly.

10.3 The Critically-Loaded Regime

We condition on N_λ:

$$\begin{aligned}
\mathbb{E}[Q_{N_\lambda}] &= \sum_{k_1 n_\lambda \le s < n_\lambda r} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \mathbb{E}[Q_{n_\lambda r}]\mathbb{P}(N_\lambda = n_\lambda r), \\
&\le \sum_{k_1 n_\lambda \le s < n_\lambda r} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|\mathbb{P}(N_\lambda = s) + \sum_{k_1 n_\lambda \le s < n_\lambda r} s\bar{q}_{\rho_s}\mathbb{P}(N_\lambda = s) + \sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) \\
&\quad + \mathbb{E}[Q(n_\lambda r)]\mathbb{P}(N_\lambda = n_\lambda r), \qquad (10.21)
\end{aligned}$$

where $\rho_s = r\rho n_\lambda / s$. Paralleling (10.19) and (10.20), we can show that there exists a finite constant C_1' such that for large λ: $\sum_{k_1 n_\lambda \le s < n_\lambda r} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|\mathbb{P}(N_\lambda = s) \le C_1'$ since $\rho_s > 1$ for all $k_1 n_\lambda \le s < n_\lambda r$. Also,

$$\begin{aligned}
\sum_{k_1 n_\lambda \le s < n_\lambda r} s\bar{q}_{\rho_s}\mathbb{P}(N_\lambda = s) &= \sum_{k_1 n_\lambda \le s < n_\lambda r} n_\lambda r \left(\int_0^{(\bar{F})^{-1}(s/n_\lambda r)} \bar{F}(x)\,\mathrm{d}x \right) \mathbb{P}(N_\lambda = s) \qquad (10.22) \\
&= \mathbb{E}\left[\left(n_\lambda r \int_0^{(\bar{F})^{-1}(N_\lambda/n_\lambda r)} \bar{F}(x)\,\mathrm{d}x \right) \mathbb{1}(N_\lambda \in [k_1 n_\lambda, n_\lambda r)) \right].
\end{aligned}$$

Using arguments as in Theorem 7.1 (noting e.g., that $g_\lambda(n_\lambda r) = \int_0^{\bar{F}^{-1}(1)} \bar{F}(x)dx = 0$), we can show that there exists a finite $C_2' > 0$ such that

$$\limsup_{\lambda\to\infty} \mathbb{E}\left[\left(n_\lambda r \int_0^{(\bar{F})^{-1}(N_\lambda/n_\lambda r)} \bar{F}(x)\,\mathrm{d}x \right) \mathbb{1}(N_\lambda \in [k_1 n_\lambda, n_\lambda r)) \right] \le C_2'.$$

By Theorem 4.1 of Zeltyn and Mandelbaum (2005), there exists $K' > 0$ such that $\mathbb{E}[Q_{n_\lambda r}] \le K'\sqrt{\lambda}$ for large enough λ. Given that $\sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) \to 0$ as $\lambda \to \infty$ (underloaded regime), we obtain that the entire expression in (10.21) is $\mathcal{O}(\sqrt{\lambda})$. The proof for the abandonment rate follows along similar lines, so we omit the relevant details. ■

11 Proofs of Propositions

Proposition 4.2 If $\Gamma_i \equiv \Gamma$, then letting $n^* = \Gamma$ yields $n_i^* = \lambda_i$ in each shift i. Thus, there is no congestion anywhere, and the overall cost $C(n^*) = \sum_{j=1}^k \lambda_i$, which is the optimal benchmark cost.

Proposition 4.3 If the abandonment distribution is exponential, then for $\Gamma_{i-1} \le n < \Gamma_i$, $u_i(n) = \sum_{j=1}^k c_j r_j n + \sum_{j=i}^k (p_j + h_j/\theta)(\lambda_j - nr_j)$. Clearly, under condition (4.6), $C(n)$ is piecewise linear with piecewise negative slopes for $n \le \Gamma_{i_0}$, and strictly positive slopes for $n > \Gamma_{i_0}$.
With a monotonically increasing hazard rate, we have

$$u_i(n) = \sum_{j=1}^k c_j r_j n + \sum_{j=i}^k \left(p_j(\lambda_j - nr_j) + h_j \lambda_j \int_0^{\bar{F}^{-1}(nr_j/\lambda_j)} \bar{F}(u)\,\mathrm{d}u \right).$$

Thus,

$$u_i'(n) = \sum_{j=1}^{k} c_j r_j - \sum_{j=i}^{k} r_j \left[p_j + \frac{h_j}{h_a\left(\bar{F}^{-1}\left(\frac{n}{\Gamma_j}\right)\right)} \right],$$

which is strictly decreasing in n, i.e., $u_i^{''}(n) < 0$. Thus, the objective is piecewise strictly concave. The minimum must be achieved at some $\Gamma_{i'}$, at which we critically load shift i'.

Proposition 4.4 In $[\Gamma_{i-1}, \Gamma_i)$, $u_i'(n)$ is as in the proof of Proposition 4.3, so that $u_i^{''}(n) > 0$ and the function is piecewise convex. It also follows that $u_i'(n_1) < u_{i+1}'(n_2)$ for $n_1 \in [\Gamma_i, \Gamma_{i+1})$ and $n_2 \in [\Gamma_{i+1}, \Gamma_{i+2})$. In other words, if $C'(x) > 0$, then $C'(y) > 0$ for $y \geq x$. Thus, the minimum n^* will be at the interior of an interval $(\Gamma_{i_0-1}, \Gamma_{i_0})$ if $u_{i_0}^{'}(\Gamma_{i_0-1}) < 0$ and $u_{i_0}^{'}(\Gamma_{i_0}-) > 0$. Here is a sufficient condition for this to be the case.

Sufficient condition. There exists i_0, $\beta, \gamma > 0$ such that:

$$\frac{\Gamma_{i_0-1}}{\Gamma_{i_0}} < \beta; \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a(\bar{F}^{-1}(\beta))} \right) < 0; \frac{\Gamma_{i_0}}{\Gamma_k} > \gamma; \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a(\bar{F}^{-1}(\gamma))} \right) > 0.$$

To see why this implies an interior point solution, note that:

$$\begin{aligned} u_{i_0}'(\Gamma_{i_0-1}) &= \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a\left(\bar{F}^{-1}\left(\frac{\Gamma_{i_0-1}}{\Gamma_i}\right)\right)} \right) < \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a\left(\bar{F}^{-1}\left(\frac{\Gamma_{i_0-1}}{\Gamma_{i_0}}\right)\right)} \right) \\ &< \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a\left(\bar{F}^{-1}(\beta)\right)} \right) < 0 \text{ by assumption.} \end{aligned}$$

Furthermore,

$$\begin{aligned} u_{i_0}'(\Gamma_{i_0}^-) &= \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a\left(\bar{F}^{-1}\left(\frac{\Gamma_{i_0}^-}{\Gamma_i}\right)\right)} \right) > \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a\left(\bar{F}^{-1}\left(\frac{\Gamma_{i_0}^-}{\Gamma_k}\right)\right)} \right) \\ &> \sum_{i=1}^{k} c_i r_i - \sum_{i=i_0}^{k} r_i \left(p_i + \frac{h_i}{f_a\left(\bar{F}^{-1}(\gamma)\right)} \right) > 0. \end{aligned}$$

Combining both, we get that $u_{i_0}'(\Gamma_{i_0-1}) < 0$ and $u_{i_0}'(\Gamma_{i_0}^-) > 0$ which, combined with the fact that $C'(\cdot)$ increases across intervals, implies that the minimizer must lie strictly in the interval $(\Gamma_{i_0-1}, \Gamma_{i_0})$. In words, if the imbalance between the augmented arrival rates $\Gamma_{i_0}/\Gamma_{i_0+1}$ is small enough, it is optimal to "strike a balance" between the two shifts, i.e., underloading a shift, while overloading the other.

Proposition 4.5. Let $m_l(t) = \mathbb{E}[X_l - t | X_l > t]$ denote the mean residual life (MRL) under abandonment distribution l, $l = 1, 2$.

- Note that $m_1(0) = m_2(0) = \mathbb{E}[X_1] = \mathbb{E}[X_2]$, and $m_1(\cdot)$ increasing while $m_2(\cdot)$ decreasing (by the respective monotonicities of the hazard rates). Thus, $m_1(t) \geq m_2(t)$ for all $t \geq$

0. Fix n, and write $u^l_j(n) = \sum_{i=1}^k c_i r_i + \sum_{i=j}^k \left(p_i(\lambda_i - nr_i) + h_i\lambda_i \int_0^{w^l_i} \bar{F}_l(u)du\right)$ where l indexes the distribution and w^l_i is the fluid waiting time in shift i. Since $\int_0^{w^l_i} \bar{F}_l(u)du = m_l(0) - \int_{w^l_i}^{\infty} \bar{F}_l(u)du = m_l(0) - \bar{F}_l(w^l_i)m_l(w^l_i)$ and $\bar{F}_l(w^l_i) = n/\Gamma_i$, we get that $\int_0^{w^1_i} \bar{F}_1(u)du \leq \int_0^{w^2_i} \bar{F}_2(u)du$ so that $u^1_j(n) \leq u^2_j(n)$ for every n fixed. In particular, this holds at optimal n^*.

- For every $a \geq 0$, we have that $m_1(a) = \mathbb{E}[X_1 - a|X_1 > a] \geq \mathbb{E}[X_1] = \mathbb{E}[X_2] \geq \mathbb{E}[X_2 - a|X_2 > a] = m_2(a)$. Writing the objective as in part 1 of the proposition, and proceeding similarly, yields that the expected cost under the NWUE abandonment distribution is lower.

- With first-order stochastic dominance, since $\Gamma_i\bar{F}(w^l_i) = 1$ under both distributions, we must have that $w^1_i \leq w^2_i$ for all i (if n is fixed). Thus, $\int_0^{w^l_i} \bar{F}_1(x)dx \leq \int_0^{w^2_i} \bar{F}_2(x)dx$ for all i under each fixed n, and we must have $C^*_1 \leq C^*_2$. The remaining stochastic order relations all imply first-order stochastic dominance, so the same holds under each as well.

Proposition 5.1. For convenience, we drop the dependence of w^e on n. It suffices to show that $\theta(w^e_{i+1}(\Gamma_{i_c})) > \theta_0$ for $i \geq i_c$. To see this, note that: $\lambda_{i+1}e^{-w^e_{i+1}\theta(w^e_{i+1})} = \Gamma_{i_c}r_{i+1}$. This implies: $e^{-w^e_{i+1}\theta(w^e_{i+1})} = \Gamma_{i_c}/\Gamma_{i+1}$, for $i \geq i_c$, i.e., $w^e_{i+1}\theta(w^e_{i+1}) = \ln\left(\frac{\Gamma_{i+1}}{\Gamma_{i_c}}\right)$. Assume that $\theta_0 \cdot \theta^{-1}(\theta_0) < \ln\left(\frac{\Gamma_{i_c+1}}{\Gamma_{i_c}}\right)$. Then, $\theta_0 \cdot \theta^{-1}(\theta_0) < w^e_{i+1}\theta(w^e_{i+1})$ for $i \geq i_c$ since $\Gamma_{i_c+1} \leq \Gamma_{i+1}$ for $i \geq i_c$. Since $x\theta^{-1}(x)$ is increasing in x, we obtain that $w^e_{i+1} > \theta^{-1}(\theta_0)$, which implies that $\theta(w^e_{i+1}) > \theta_0$ for $i \geq i_c$, as desired. Then, $C_a(\Gamma_{i_c}) < C(\Gamma_{i_c}) = C^*$, and we get strict reduction in cost due to the announcements.

Lemma 6.1. We derive the optimal compensation for a fixed value of the pool size n. Since c^*_i can be decided upon separately for each shift, we focus on a single shift setting in what follows, i.e., we fix the shift i. The solution depends on the specific value of n.

1. $n \geq \frac{\lambda_i}{G(l)}$: $c^*_i = l$, i.e., offer minimum wage and overstaff shift i (under-loaded).

2. $n < \frac{\lambda_i}{G(l)}$. Note that we must have that $\lambda_i \geq nG(c_i)$ i.e., $c_i \leq G^{-1}\left(\frac{\lambda_i}{n}\right)$ because it will not be cost effective for the manager to incite more supply than the demand in the shift.

   Subcase 1: We assume that $L_i \leq l$. In this case, the problem becomes:

$$\min_{L_i \leq l \leq c_i \leq G^{-1}\left(\frac{\lambda_i}{n}\right)} nc_iG(c_i) + L_i(\lambda_i - nG(c_i))$$

   which is equivalent to

$$\min_{L_i \leq l \leq c_i \leq G^{-1}\left(\frac{\lambda_i}{n}\right)} t_i(c_i) \equiv (c_i - L_i)G(c_i).$$

   Since $c_i > L_i$, it is readily seen that the objective is increasing in c_i. Thus, we must have that $c^*_i = l$. That is, we offer minimum wage and understaff shift i (over-loaded).

   Subcase 2: We now assume that $L_i > l$. In this case, $\frac{\lambda_i}{G(L_i)} < \frac{\lambda_i}{G(l)}$. We then consider the two intervals: (a) $n \leq \frac{\lambda_i}{G(L_i)} < \frac{\lambda_i}{G(l)}$ and (b) $\frac{\lambda_i}{G(L_i)} < n < \frac{\lambda_i}{G(l)}$.

(a) $n \leq \frac{\lambda_i}{G(L_i)} < \frac{\lambda_i}{G(l)}$. The problem is now: $\min_{l \leq c_i \leq \min\{G^{-1}\left(\frac{\lambda_i}{n}\right), L_i\}} nc_iG(c_i)+L_i(\lambda_i - nG(c_i))$ which is equivalent to solving:

$$\min_{l \leq c_i \leq \min\{L_i, G^{-1}\left(\frac{\lambda_i}{n}\right)\}} t_i(c_i) \equiv (c_i - L_i)G(c_i).$$

Note that $t_i'(c_i) = G(c_i)\left(1 + (c_i - L_i)\frac{g(c_i)}{G(c_i)}\right)$. In this case, we have $L_i \leq G^{-1}\left(\frac{\lambda_i}{n}\right)$. Since $t'(L_i) \geq 0$, and $t_i(\cdot)$ is convex under log-concavity of G, we obtain that:

i. If $t_i'(l) < 0$ i.e., $\left(1 + (l - L_i)\frac{g(l)}{G(l)}\right) < 0$, then there exists an optimal $c_i^* = a_i \in (l, L_i)$ where $t'(a_i) = 0$;

ii. If $t_i'(l) \geq 0$ i.e., $\left(1 + (l - L_i)\frac{g(l)}{G(l)}\right) \geq 0$, then we have $c_i^* = l$.

In both cases (i) and (ii), the system is overloaded, i.e., the manager incites a smaller supply than the demand in shift i.

(b) Now, consider: $\frac{\lambda_i}{G(L_i)} < n < \frac{\lambda_i}{G(l)}$. Let $0 < a_i < L_i$ be such that $t_i'(a_i) = 0$ i.e.,

$$G(a_i)\left(1 + (a_i - L_i)\frac{g(a_i)}{G(a_i)}\right) = 0.$$

The optimization problem is

$$\min_{l \leq c_i \leq G^{-1}\left(\frac{\lambda_i}{n}\right) < L_i} t_i(c_i).$$

Note that if $a_i < l$, then $c_i^* = l$ (by the convexity of the objective); in other words, the manager offers the minimum wage and runs shift i overloaded. Now, assume that $a_i \geq l$. We then have the following two cases:

i. $t'\left(G^{-1}\left(\frac{\lambda_i}{n}\right)\right) \leq 0$ i.e., $G^{-1}\left(\frac{\lambda_i}{n}\right) \leq a_i$ i.e., $\frac{\lambda_i}{G(L_i)} < \frac{\lambda_i}{G(a_i)} \leq n < \frac{\lambda_i}{G(l)}$. In this case, $c_i^* = G^{-1}\left(\frac{\lambda_i}{n}\right)$ which means that the manager incites a supply equal to the demand, i.e., she critically loads her shift.

ii. $t'\left(G^{-1}\left(\frac{\lambda_i}{n}\right)\right) > 0$ i.e., $G^{-1}\left(\frac{\lambda_i}{n}\right) > a_i$ i.e., $\frac{\lambda_i}{G(L_i)} < n < \frac{\lambda_i}{G(a_i)} \leq \frac{\lambda_i}{G(l)}$. In this case, $c_i^* = a_i$ and the manager incites a supply that is smaller than the demand, i.e., she overloads her shift.

Lemma 6.2. We let $\tilde{a}_k$ be the solution to (6.14). Then, $\tilde{t}'(x) \equiv G(x)\left(1 + (x - \tilde{L}_k)\frac{g(x)}{G(x)}\right)$ is increasing for $x \leq \tilde{L}_k$ by the log-concavity of $G(\cdot)$. If $a_k > \tilde{L}_k$, then it must be that $a_k > \tilde{a}_k$ since $\tilde{a}_k < \tilde{L}_k$. Let us now assume that $a_k \leq \tilde{L}_k$. Since $\tilde{L}_k < L_k$, we must have that

$$G(a_k)\left(1 + (a_k - \tilde{L}_k)\frac{g(a_k)}{G(a_k)}\right) > G(a_k)\left(1 + (a_k - L_k)\frac{g(a_k)}{G(a_k)}\right) = G(\tilde{a}_k)\left(1 + (\tilde{a}_k - \tilde{L}_k)\frac{g(\tilde{a}_k)}{G(\tilde{a}_k)}\right) 0.$$

Because $\tilde{t}'(x)$ is increasing in x for $x \leq \tilde{L}_k$, and we have both $a_k, \tilde{a}_k \leq \tilde{L}_k$, we also obtain that $a_k > \tilde{a}_k$. If $n < \lambda_k/G(a_k)$, then we must also have that $n < \lambda_k/G(\tilde{a}_k)$, so that the optimal compensation as per Lemma 6.1 is to set $\tilde{c}_k^* = \tilde{a}_k < c_k^* = a_k$. We note that if n is as in cases (a) and (b) of Lemma 6.1, then the compensation offered to agents is unchanged since compensation is set so that there is no congestion in the shift. We also note that if $\tilde{a}_k < l < a_k$ then $\tilde{c}_k^* = l$ so that $\tilde{c}_k^* < c_k^*$ as well. In other words, agents are worse off in all cases.

Lemma 6.3. The derivative of the objective is given by:

$$\begin{aligned}\Pi'(n) &= \sum_{\{i:n\geq\frac{\lambda_i}{G(l)}\}} lG(l) \\ &- \sum_{\{i:\frac{\lambda_i}{G(\tilde{a}_i)}\leq n<\frac{\lambda_i}{G(l)}\}} \frac{\lambda_i^2}{n^2 g\left(G^{-1}\left(\frac{\lambda_i}{n}\right)\right)} \\ &+ \sum_{\{i:n<\frac{\lambda_i}{G(\tilde{a}_i)}\leq\frac{\lambda_i}{G(l)}\}} (\tilde{a}_i - \tilde{L}_i)G(\tilde{a}_i).\end{aligned}$$

Note that for values of n such that the shift groupings in (6.16) do not change, the derivative $\Pi'(n)$ is increasing in n under log-concavity of $G(\cdot)$. Indeed, $-\frac{(\lambda_i/n)^2}{g(G^{-1}(\lambda_i/n))} = -\frac{\lambda_i}{n}\frac{G(G^{-1}(\lambda_i/n))}{g(G^{-1}(\lambda_i/n))}$ is increasing in n (this is for the critically-loaded shifts). Also, $t'(\tilde{a}_i) = 0$ for $\tilde{t}'(x) \equiv G(x)\left(1 + (x - \tilde{L}_i)\frac{g(x)}{G(x)}\right)$. Thus, assuming that $G(\tilde{a}_i) > 0$, we get that $1 + (\tilde{a}_i - \tilde{L}_i)\frac{g(\tilde{a}_i)}{G(\tilde{a}_i)} = 0$. Since $u(n) = \frac{(\lambda_i/n)^2}{g(G^{-1}(\lambda_i/n))}$ is decreasing in n, we must have that $u(n) \leq u(\lambda_i/G(\tilde{a}_i)) = \frac{G(\tilde{a}_i)^2}{g(\tilde{a}_i)} = -G(\tilde{a}_i)(\tilde{a}_i - \tilde{L}_i)$ if $n \geq \lambda_i/G(\tilde{a}_i)$. This implies that $-\frac{(\lambda_i/n)^2}{g(G^{-1}(\lambda_i/n))} \geq G(\tilde{a}_i)(\tilde{a}_i - \tilde{L}_i)$ for all $\frac{\lambda_i}{G(\tilde{a}_i)} \leq n < \frac{\lambda_i}{G(l)}$.

In other words, if n increases such that a shift i that was understaffed becomes critically-loaded, then the corresponding part of the derivative of the objective increases too. Thus, $\Pi'(n)$ strictly increases as n increases, so that the function is overall strictly convex. Since $\Pi'(0) < 0$ and $\Pi'\left(\max\{\frac{\lambda_i}{G(l)}\}\right) > 0$, there must exist a unique solution n^*. Together with the optimal compensation results in Lemma 6.1, we can obtain the optimal solution (n^*, c^*) to the original problem.

Lemma 6.4. Note that if $l < l_0$ then $\max\{\frac{\lambda_i}{G(\tilde{a}_i)}\} < \min\frac{\lambda_i}{G(l)}$. For $\max\{\frac{\lambda_i}{G(\tilde{a}_i)}\} < n < \min\frac{\lambda_i}{G(l)}$, we must have that $\Pi'(n) < 0$. Thus, $n^* \geq \min\frac{\lambda_i}{G(l)} > \max\{\frac{\lambda_i}{G(\tilde{a}_i)}\}$, and we do not overload or use the announcements in any shift (since $\Pi'(n)$ is strictly increasing in n). It is readily seen that we cannot, for an optimal n^*, have all shifts strictly underloaded. Thus, there must exist i_0 as specified in the lemma.

Lemma 6.5. In this case, problem (6.15) simplifies to:

$$\begin{aligned}\min_{n\geq 0}\Pi(n) &\equiv \sum_{\{i:n\geq\frac{\lambda_i}{G(l)}\}} nlG(l) \qquad \text{(underloaded)} \\ &+ \sum_{\{i:n<\frac{\lambda_i}{G(l)}\}} lnG(l) + \tilde{L}_i(\lambda_i - nG(l)) \qquad \text{(overload+announcements)}\end{aligned}$$

Note that $\Pi(n)$ is piecewise linear. Then, $\Pi'(n) = klG(l) - \sum_{\{i:n<\frac{\lambda_i}{G(l)}\}}\tilde{L}_iG(l)$. Clearly, as n increases, $\Pi'(n)$ increases too. Under our assumptions, there must exist a unique k_0 such that $\Pi'(n) < 0$ for $n < \frac{\lambda_{k_0}}{G(l)}$ and $\Pi'(n) > 0$ for $n > \frac{\lambda_{k_0}}{G(l)}$. The optimal solution is to set $n^* = \frac{\lambda_{k_0}}{G(l)}$.