Chapter 1
On Queues with a Random Capacity: Some Theory, and an Application

Rouba Ibrahim

Abstract One standard assumption in workforce management is that the firm can dictate to workers when to show up to work. However, that assumption is challenged in modern business environments, such as those arising in the sharing economy, where workers enjoy various degrees of flexibility, including the right to decide when to work. For example, a ride-sharing service cannot impose on its drivers to be on the road at specific times; similarly, a virtual call-center manager cannot direct her agents to be available for select shifts. When self-scheduling is allowed, the number of workers available in any time period is uncertain. In this chapter, we are concerned with the effective management of service systems where capacity, i.e., the number of available agents, is random. We rely on a queueing-theoretic framework, because customers are time-sensitive and delays are ubiquitous in the services industry, and focus on the performance analysis and control of a queueing system with a random number of servers. In particular, we begin by surveying some theoretical results on the control of queueing systems with uncertainty in parameters (here, the number of servers). Then, we illustrate how to apply those theoretical results to study the problems of staffing and controlling queueing systems with self-scheduling servers and impatient, time-sensitive, customers.

1.1 Introduction

Nowadays, there seems to be "an Uber for everything", e.g., for food delivery, parking, haircuts, domestic cleaning, etc. Such on-demand services are typically provided by online service platforms, which match consumers with workers who are willing to perform said services for a certain fee. The rise of those on-demand platforms has transformed the ways in which services are sought and delivered. For example, rather than managing a team of employees, an on-demand service provider must manage a virtual pool of independent contractors who have a legal right to various degrees of flexibility, e.g., in deciding which tasks to perform, or in setting their own schedules.

The effective management of innovative businesses in the sharing economy presents new challenges for practitioners and academics alike. To illustrate, let us consider the call-center industry. In recent years, virtual call centers have become increasingly prevalent (Vocalcom 2014). In virtual call centers, such as LiveOps (*liveops.com*) or Arise (*arise.com*), self-scheduling agents are usually independent contractors who have no requirement on the least number of working hours to be fulfilled, and are free to choose their own working periods in 30-minute intervals. The standard staffing question (how many agents should the firm have?) is especially relevant in virtual call centers. This is because managing those systems involves two different time scales, and the agent pool size cannot be easily adjusted at very short time notice: (i) weeks ahead of time, typically 4-10 weeks, the system manager selects the total staffing level in the system to allow sufficient time for agent training and qualification; and, (ii) days or, in many cases, just hours ahead of time, agents select their own schedules. 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. Moreover, the promised scheduling flexibility constitutes the main appeal of these jobs, and cannot be simply restricted by the firm. Therefore, hiring the right number of self-scheduling agents which scales appropriately to fit customer needs is a fundamental challenge.

Rouba Ibrahim
University College London, 1 Canada Square, London E14 5AB, U.K., e-mail: rouba.ibrahim@ucl.ac.uk

Similar operational challenges arise in other service contexts as well. 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" to ensure the participation of a sufficient number of drivers in different time periods. 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 are interested in studying the effective operational management of such service systems. To do so, we adopt a queueing-theoretic framework. Relying on queueing models is natural, in our setting, because customers are time-sensitive and delays are ubiquitous in the services industry. To wit, 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, with self-scheduling agents, a firm cannot simply impose on its workers to show up to work at a given time. In other words, the number of agents in any given shift is uncertain, i.e., it must be modelled as a random variable instead.

This chapter studies optimal staffing and control decisions in queueing systems with a random number of servers. Thus, we can position our work, broadly, as being part of the literature on controlling queueing systems with model-parameter uncertainty. This literature can be classified into two main categories: The first category aims at reducing parameter uncertainty through better forecasting (Shen and Huang (2008), Aldor-Noiman et al. (2009), Ibrahim and L'Ecuyer (2013), etc.). The second category, which is more closely related to our approach, investigates effective decision-making in the context of queueing systems with uncertain parameters (Harrison and Zeevi (2005), Whitt (2006b), Bassamboo and Zeevi (2009), Gans et al. (2015), etc.). With that in mind, our aim in this chapter is two-fold. First, we provide some required theoretical background. Specifically, in §1.2, we survey recent papers which propose approximations to queueing systems with uncertain parameters; those approximations are grounded in many-server heavy-traffic limits. Second, we illustrate how those theoretical results may be applied; we devote all remaining sections to that aim. Specifically, we describe some results from Ibrahim (2017a) who studies the operational management of queueing systems with self-scheduling agents, using both short-term and long-term controls.

1.2 Theoretical Background: Queues with Uncertain Parameters

At a high level, the analysis of queueing systems with uncertain parameters is complicated, for the most part, because it involves two "layers" of variability: (i) *stochastic variability*, for any realized value of the underlying uncertain parameter, since e.g., interarrival, service, and patience times are random; and (ii) *parameter uncertainty*, since the parameter itself, e.g., the number of servers in our setting, is random. Because of that analytical complexity, and because we are primarily interested in studying the operations of large service systems, it is useful to rely on many-server heavy-traffic limiting regimes, which typically simplify the analysis and yield valuable insight. In particular, performance measures of interest, e.g., the expected queue length in our setting, are approximated by limits of appropriate sequences, where the arrival rate is allowed to grow without bound. To rigorously justify the appropriateness of such approximations, we have to quantify their corresponding errors, asymptotically in large systems. Specifically, we have to determine how the orders of magnitudes of those errors grow as the system size (or the arrival rate) increases. Here, we focus on two problem formulations, corresponding to two regimes, which have been proposed in the literature.

Stochastic-fluid approximation. The first formulation assumes that uncertainty effects dominate stochastic fluctuations. That is, stochastic fluctuations may be ignored, in large systems, and one can focus solely on uncertainty effects. The intuition is that detailed fluctuations of stochastic processes (item i above) are typically realized on a short time scale, whereas parameter uncertainty (item ii above) is typically realized on a longer time scale. So, if uncertainty effects are "large", then stochastic fluctuations become less critical in describing performance in the system. For example, with self-scheduling servers: Variability in the numbers of servers is realized from shift to shift, i.e., over the hourly time scale. In contrast, variability due to the randomness in arrivals, service times, and times to abandon, is realized on a shorter time scale, e.g., over the course of a few seconds or minutes. In this case, if the uncertainty in the number of servers is larger than the order of stochastic fluctuations, then one can derive e.g., cost-minimizing staffing levels, by solving a *stochastic-fluid* optimization problem which is a fluid-type problem where only parameter uncertainty is accounted for. This type of approximation was first proposed in Harrison and Zeevi (2005) for the control of a multi-class queueing system, and considered in e.g., Bassamboo et al. (2010a) for random arrival rates, and Dong and Ibrahim (2017) for a random number of servers.

For further emphasis, that difference in time scales can be visualised in e.g., computer simulations of the system. In particular, to simulate the system, one first draws a random variate from the distribution of the number of servers, and then simulates a queueing model with that realization for the number of servers. In another (independent) simulation run, one draws another realization from that same distribution, simulates the system for that new realization, and so on. The variability in the number of servers across multiple simulation runs is due to the long-run parameter uncertainty, whereas the variability within a given run, e.g., fluctuations of the queue length around its average in that run, is due to short-term stochastic fluctuations. We illustrate stochastic-fluid approximations in Figures 1.1 and 1.2. In Figure 1.1, we plot simulation sample paths of the queue lengths seen by arriving customers in a queueing system where the number of servers is random (here, assumed to follow a truncated normal distribution), and where the system is, on average, overloaded. The solid horizontal line in each subfigure corresponds to the average queue length in that simulation run. As expected, conditional on the realized number of servers, the magnitude of stochastic fluctuations around the average is on the order of the square-root of the arrival rate (Garnett et al. 2002b), while the average queue-length itself is on the order of the arrival rate. In Figure 1.2, we "ignore" stochastic fluctuations, and approximate the queue length seen upon arrival by the average queue length in that run. We note that this average changes depending on the run, i.e., depending on the specific realization for the number of servers in that particular run. Relying on a stochastic-fluid approximation amounts to focusing on that cross-run variation in the means, while ignoring the more refined stochastic fluctuations within a given simulation run.

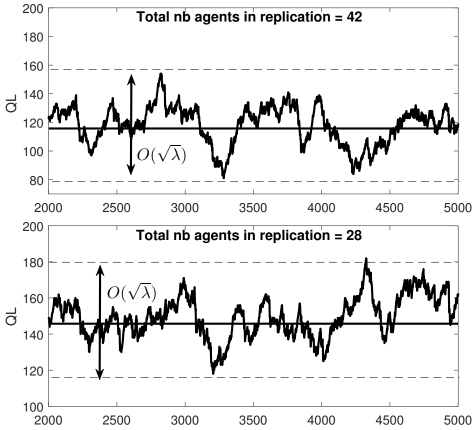

Fig. 1.1 Queue lengths seen by arriving customers in two independent simulation replications of an $M/M/N+M$ model where N follows a truncated Nor$(72,30)$ distribution and $\lambda = 100$.

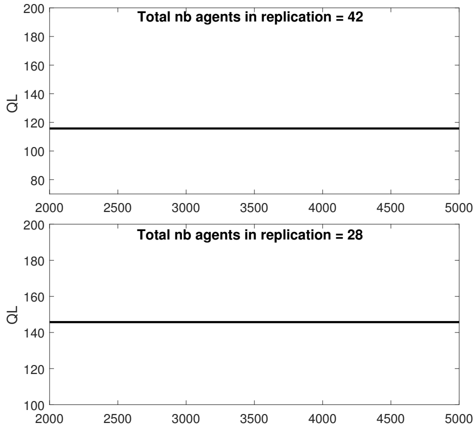

Fig. 1.2 Average queue lengths in two independent simulation replications of an $M/M/N+M$ model where N follows a truncated Nor$(72,30)$ distribution and $\lambda = 100$.

Fluid approximation. The second formulation assumes that both uncertainty effects and stochastic fluctuations are negligible, and can be ignored. In this regime, we derive, e.g., the optimal staffing policy, by solving a deterministic *fluid* optimization problem instead. To illustrate, this corresponds to additionally ignoring the variations in the average queue-lengths across the different runs in Figure 1.2. Whitt (2006c) conjectured the existence of a deterministic fluid limit for general overloaded queueing systems. That fluid limit was later established in Kang et al. (2010) and Zhang (2013). While crude fluid approximations are generally less accurate than their stochastic-fluid counterparts, they remain very useful because they usually have a remarkably simple form. Moreover, they are extremely accurate in many cases, so that there may be no tangible advantage from considering more refined approximations.

Dong and Ibrahim (2017) compare the asymptotic accuracies, i.e., the orders of magnitude of errors, corresponding to both stochastic fluid and fluid approximations in a system with a random number of servers. To summarize their main result, which parallels the result in Bassamboo et al. (2010a) for random arrival rates: When the variance of the number of servers is asymptotically large, in particular larger than the square-root order of stochastic fluctuations, the system may be considered to be in an *uncertainty-dominated regime* where stochastic-fluid approximations are remarkably accurate. Moreover, the more variable the number of servers, the more accurate are those stochastic-fluid approximations. In contrast, if that variance is asymptotically small, in particular at most equal to the square-root order of stochastic fluctuations, then the system may be considered to be in a *variability-dominated regime*, where there is no tangible benefit from using stochastic-fluid approximations over fluid approximations.

1.2.1 Self-Scheduling Servers: A Binomial Distribution

To model self-scheduling agent behavior, it is natural to assume that there is a pool of agents, of size n, and that each agent from that pool makes an independent decision to join a shift j with a given probability, r_j. In this case, the random number of servers in shift j, which we denote by N_j, has a binomial distribution, $Bin(n, r_j)$, where n is the number of trials and r_j is the success probability. Because $\mathrm{Var}[N_j] = \sqrt{nr_j(1-r_j)}$, the variance is on the square-root order, i.e., it is of the same order as stochastic fluctuations in the system. Thus, there should be no advantage in using the stochastic-fluid model, over the fluid model, when the system is large. Intuitively, this is because the binomial distribution "concentrates" around its mean, nr_j, when n is large. We can formally prove this intuition (here, we focus on a single shift since the results easily extend to multiple shifts). To do so, we restrict attention to exponentially-distributed service times and a Poisson arrival process. In particular, service times are independent and identically distributed (i.i.d.) random variables with an exponential distribution and mean $1/\mu$. We assume, without loss of generality, that $\mu = 1$; this amounts to measuring time in units of mean service times. We assume that 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. Abandonment makes the system stable, irrespective of the realized numbers of servers. There is unlimited waiting space, and we use the first-come-first-served (FCFS) service discipline.

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 the traffic intensity $\rho \equiv \lambda/\mathbb{E}[N_\lambda] = \lambda/rn_\lambda$ 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 quality-and-efficiency driven (QED) 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 $\lambda > N_\lambda$. Let $\bar{q}_\rho$ and $\bar{\alpha}_\rho$ be the fluid approximations for the queue length and net abandonment rates with a traffic intensity ρ. The following theorem establishes the asymptotic accuracy of fluid approximations with a binomially-distributed number of servers.

Theorem 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*

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

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

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

(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 1 shows that, in the overloaded system, the fluid approximation for the expected queue length is asymptotically accurate up to $\mathcal{O}(1)$ [1], and the fluid approximation for the net abandonment rate is asymptotically accurate up

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

(a) $f(n) = \mathcal{O}(g(n))$ if there exists $M > 0$ and $C > 0$ such that $|f(n)| \le M|g(n)|$ for $n \ge C$;
(b) $f(n) = o(g(n))$ if for all $\varepsilon > 0$, there exists N such that $|f(n)| \le \varepsilon|g(n)|$ for all $n \ge N$.

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 $\mathscr{O}(\sqrt{\lambda})$, 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. In other words, relying on fluid approximations is justifiable when the number of servers follows a binomial distribution, which is a reasonable model for self-scheduling server behavior.

1.2.2 What Do the Asymptotic Results Mean?

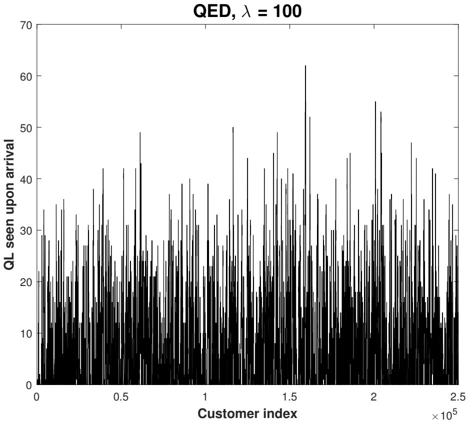

Fig. 1.3 Queue length seen upon arrival in a single simulation run with $\rho = 1$ and $n = 100$. The fluid limit is equal to 0.

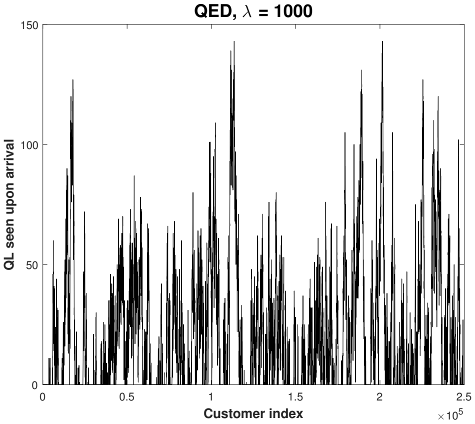

Fig. 1.4 Queue length seen upon arrival in a single simulation run with $\rho = 1$ and $n = 1000$. The fluid limit is equal to 0.

To interpret the asymptotic results of Theorem 1, it is important to clearly distinguish between *stochastic fluctuations* for the queue-length process (which can be observed, e.g., in a given simulation run), and the *accuracy of many-server fluid approximations* for the expected queue-length (obtained by averaging over multiple simulation runs, and quantified by letting λ increase). Figures 1.3-1.8 are based on simulations of an $M/M/n+M$ queueing model with service rate $\mu = 1$ and abandonment rate $\theta = 0.5$. We consider a deterministic number of servers in these simulations because the same intuitions continue to hold when the number of servers is binomially distributed instead.

In Figures 1.3 and 1.4, we present queue-length sample paths based each on a single simulation run in a system where $n = \lambda$ i.e., $\rho = 1$. For such parameter values, critical-loading approximations (Garnett et al. 2002a) are known to describe the system well. For Figure 1.3, we let $\lambda = 100$, and for Figure 1.4, we let $\lambda = 1000$. In each figure, the fluid limit is identically equal to 0. It is clear from Figures 1.3 and 1.4 that the magnitude of stochastic fluctuations in the system is on the order of $\sqrt{\lambda}$, as expected. The same continues to hold in an overloaded system, as illustrated in Figures 1.5 and 1.6, where we let $\rho = 1.4$ instead.

The asymptotic accuracy results of Theorem 1 describe how the expected queue length differs from its fluid limit as λ increases. Figure 1.7 considers different λ values, ranging from $\lambda = 100$ to $\lambda = 1000$, in critically-loaded systems where $n = \lambda$. As a function of λ, we plot estimates of the expected queue length which are based on averages over 10 independent simulation runs of length 10 million arrivals each. In Figure 1.8, we do the same but let $\rho = 1.4$ instead, i.e., we consider an overloaded system. Contrasting Figures 1.7 and 1.8 illustrates our asymptotic accuracy results in Theorem 1. In the critically-loaded case, i.e., Figure 1.7, stochastic fluctuations are consistent with those suggested by the **central limit theorem**, i.e., they are on the order of $\sqrt{\lambda}$. In the overloaded case, i.e., Figure 1.8, stochastic fluctuations are better explained by **large deviations theory**: Fluid approximations are practically indistinguishable from the estimates for average queue-lengths; see Bassamboo et al. (2010b) for related additional discussion.

This section presented the theoretical background needed to study performance in queueing systems with randomness in capacity. In the remainder of this chapter, we apply that theoretical framework to formulate managerial recommendations on the operational management of queueing systems with self-scheduling servers.

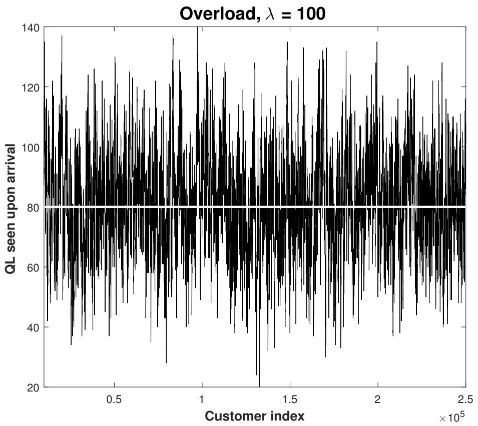

Fig. 1.5 Queue length seen upon arrival in a single simulation run with $\rho = 1.4$ and $n = 100$. The fluid limit is equal to 80.

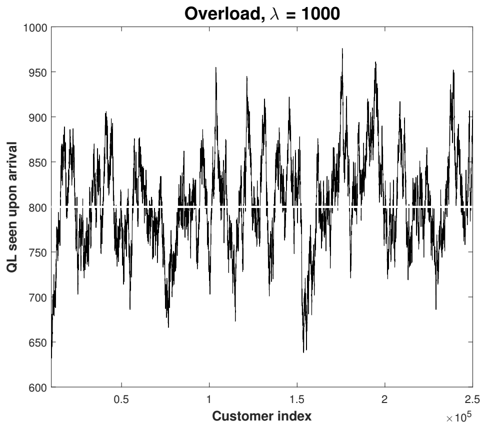

Fig. 1.6 Queue length seen upon arrival in a single simulation run with $\rho = 1.4$ and $n = 1000$. The fluid limit is equal to 800.

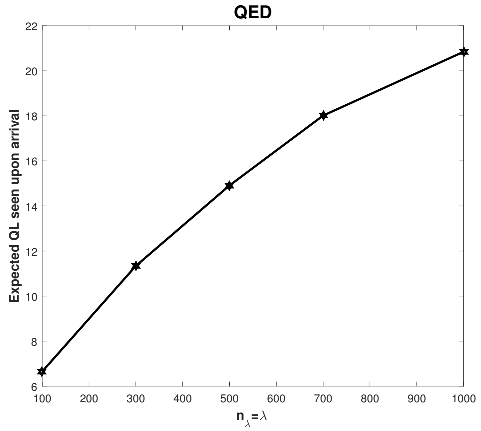

Fig. 1.7 Averages of queue lengths for varying λ where $n_\lambda = \lambda$.

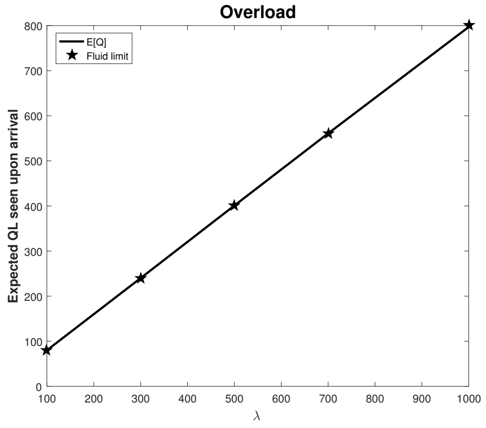

Fig. 1.8 Averages of queue lengths for varying λ where $\lambda / n_\lambda = 1.4$.

1.3 Self-Scheduling Agents: A Long-Term Staffing Decision

1.3.1 The Model

There are k shifts, and agents show up at random for these shifts. In particular, an agent shows up for work in shift j with a given probability, r_j, independently of other agents. We denote the total pool size by n. The number of agents in shift j is a random variable with a binomial distribution, $N_j \sim Bin(n, r_j)$, where n is the number of trials and r_j is the success probability. That is, nr_j is the expected number of agents that show up to work in shift j. Customers arrive to the system according to general stationary processes with rates λ_j. 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. This assumption is reasonable when the system is large enough: In this case, processing any customers who remain in queue at the end of a given shift would not take too long, because there are many servers working in parallel. In each shift, we consider a $G/G/N_j + GI$ model. 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 $\theta > 0$. Service times are assumed to be i.i.d with a general distribution. The arrival, service, and abandonment processes are all mutually independent, also independent of the number of servers. We continue to use the FCFS service discipline.

The system manager must select an appropriate total staffing level, n, to effectively balance staffing and operational costs. With a binomially-distributed number of servers, relying on a fluid approximation is justified in large systems (§1.2.1). In other words, both stochastic variability and parameter uncertainty may be assumed to be of second order, relative to average performance measures in the system. By ignoring stochasticity, the key challenge in managing a random capacity reduces to the salient *heterogeneity across shifts*, in both demand rates and agent availabilities. To illustrate this point, let us compare the settings with a single and two shifts. With a single shift, assume that each agent has a probability 0.25 of showing up for that shift, and that 100 agents are needed. Then, with a pool of $n = 400$ self-scheduling agents, 100 agents will show up on average. Thus, by staffing a large enough agent pool, the manager could induce the desired number of agents to show up, on average. Now, with two shifts, e.g., morning and afternoon, assume that each agent has probabilities 0.25 to show up in the morning shift, and 0.5 to show up in the afternoon shift. Also, assume that 100 agents are needed for the morning shift, and 150 for the afternoon shift. Then, staffing a pool that is large enough to meet demand in one of the two shifts, on average, will lead to either overstaffing or understaffing the other shift. With multiple shifts and heterogeneous arrival rates and show-up probabilities, it is not clear, a priori, how the manager should staff her system: Should she aim to match demand in some shifts, but not others? What should this decision depend on? We investigate such questions in what follows.

1.3.2 Fluid Formulation

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. Each agent is paid c_j per unit time in shift j, if she is available for work in that shift. The system manager must decide on the staffing level, n. Let $\bar{q}_{\rho_n}$ and $\bar{\alpha}_{\rho_n}$ be the fluid approximations for the queue-length and net abandonment rates with a traffic intensity $\rho_n \equiv \lambda / n\mu$.

Staffing decisions in systems with self-scheduling agents cannot, usually, be made "on the fly". Because of this, a manager must make her staffing decision in advance: She may not know, with certainty, the availability of each agent in her workforce and she cannot enforce attendance but, given current technological advances, she should be able to obtain *historical estimates* of joining probabilities of agents; these are the r_j in our model. For example, based on analyzing human resources data in her firm, she may know that stay-at-home parents usually prefer to work in the morning, while children are at school, but may not know whether a specific work-from-home parent will show up to work on a given morning. She may know that higher compensations are typically offered during certain times of day (ride-sharing) or for certain client companies (virtual call centers), but may not know the exact "surge prices" that are going to be in effect, if any. To capture such challenges, we assume that the compensation, c_j, and the show-up probability, r_j, are fixed (we will relax those assumptions later). In other words, we begin by solving the problem:

$$\min_{n \in \mathbb{N}} C(n) \equiv \sum_{1 \le j \le k} \left(n r_j c_j + p_j \cdot \bar{\alpha}_{\rho_n / r_j} + h_j \cdot \bar{q}_{\rho_n / r_j} \right), \tag{1.1}$$

where $\mathbb{N}$ denotes the set of natural integers.

1.3.3 Optimal Staffing Policy

1.3.3.1 No Self-Scheduling

To quantify the impact of self-scheduling, we need to choose a useful benchmark. Without self-scheduling, the system manager can select the optimal staffing levels, n_j^*, independently for each shift j. To specify n_j^*, we need to make additional assumptions. 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; this is because it is suboptimal to staff more than λ_j agents in shift j, i.e., underload that shift. In shift j, w_j^* is determined by solving:

$$\min_{w_j \ge 0} \lambda_j \left((c_j - p_j) \bar{F}(w_j) + h_j \int_0^{w_j} \bar{F}(u)\,\mathrm{d}u \right). \tag{1.2}$$

Hereafter, we make the following assumption, which states that staffing costs are sufficiently inexpensive.

Assumption 1.3.1 *For all j, $c_j < \min\{h_j/h_a(0)+p_j, h_j/\theta+p_j\}$.*

It is useful to offer a brief comment on the validity of Assumption 1.3.1. To do so, let us assume that c_j corresponds to the minimum wage which is close to \$7 per hour in the United states. Let us also take the time unit to be one hour. Assume that $\theta = 4$, i.e., a customer is willing to wait on average for 15 minutes before abandoning. For a numerical value of h_j, we make use of existing empirical evidence from the call-center literature, e.g., Akşin et al. (2013). Based on their results (see Table 4 in that paper), customers attribute a waiting cost of roughly 1 \$ per minute. This translates into 60 \$ per hour. For such values, the assumption that we make on the staffing cost is satisfied irrespective of the value of p_j. Under Assumption 1.3.1, it is easy to establish the following result for the solution to problem (1.2).

Proposition 1. *Under Assumption 1.3.1, in a system with no self-scheduling servers, it is optimal to match the supply and demand rates in every shift, i.e., $n_j^* = \lambda_j$.*

In other words, we choose as benchmark a setting where it is optimal to match demand and supply in each of the shifts. This is the case when staffing costs are not too high, as per Assumption 1.3.1. When demand and supply are matched in each shift, there is no delay, at fluid scale. Thus, the customer abandonment distribution does not play any role, since customers do not abandon. With self-scheduling, the manager is no longer able to set n_j^* independently for each shift and must decide, instead, on the total pool size n. Because of the ensuing imbalance between demand and supply, some shifts may be congested. Thus, because of self-scheduling, the customer abandonment distribution will now play an important role in congested shifts. Here, we study how to exploit that role to mitigate the cost of self-scheduling.

1.3.3.2 Self-Scheduling Capacity

To capture the heterogeneity across different shifts, we define the *augmented arrival rate* $\Gamma_j \equiv \lambda_j/r_j$, and let $\Gamma_0 \equiv 0$. This will allow us to characterize, in a simple manner, the optimal solution to the staffing problem in (1.1). Letting $n = \Gamma_j$ amounts to matching the supply and demand rates in shift j. This is because the number of agents who show up in shift j is then equal to $n \cdot r_j = \Gamma_j \cdot r_j = \lambda_j$. In a sense, the respective values of Γ_j, across shifts, quantify the degree of self-scheduling imbalance in the system. In particular, if $\Gamma_j \equiv \Gamma$ are identical across all shifts, then it is easy to see that staffing $n = \Gamma$ would eliminate the cost of self-scheduling, on average. However, if the Γ_j's are "very different" across the different shifts, then managing self-scheduling agents becomes increasingly difficult, i.e., leading to a higher cost. In an overloaded shift j, we have that $\Gamma_j \bar{F}(w_j) = n$, i.e., $w_j = \bar{F}^{-1}(n/\Gamma_j)$. Since it is never optimal to strictly underload all shifts, the staffing problem in (1.1) can be defined piecewise:

$$\min_{0 \le n \le \Gamma_k} C(n) \equiv \left(\sum_{j=1}^{k} \mathbf{1}(\Gamma_{j-1} \le n < \Gamma_j) u_j(n) \right), \tag{1.3}$$

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

$$u_j(n) \equiv \sum_{i=1}^{k} c_i n r_i + \sum_{i=j}^{k} \left(p_i(\lambda_i - n r_i) + h_i \lambda_i \int_0^{\bar{F}^{-1}(n/\Gamma_i)} \bar{F}(u)\, \mathrm{d}u \right), \tag{1.4}$$

i.e., $u_j(n)$ is the *total* cost incurred if n is chosen in the interval $[\Gamma_{j-1}, \Gamma_j)$. It turns out that the solution to (1.3) depends on the monotonicity of the hazard rate of the abandonment distribution. Here is how.

Monotonically increasing hazard rate. A monotonically increasing hazard rate corresponds to customer patience "wearing out" as the customers waits longer in queue. For abandonment distributions with a monotonically increasing hazard rate (including the exponential distribution with a constant hazard rate), we find that it is optimal to match the supply and demand rates in one of the k shifts when servers self schedule, with the remaining shifts being either over or under staffed; this lies in contrast to matching supply and demand in all shifts without self-scheduling, as per Proposition 1. In particular, the following proposition holds:

Proposition 2. *For abandonment distributions with a monotonically non-decreasing hazard rate, there is one shift i_0 where the supply and demand rates must be matched, i.e., $n^* = \Gamma_{i_0}$.*

The optimality of overstaffing certain shifts lends some support to the staffing policies adopted in virtual call centers such as LiveOps or Arise, 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 across all shifts.

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. A monotonically decreasing hazard rate for abandonment corresponds to customers becoming increasingly patient as they wait long in queue, e.g., because they feel that "they have waited already for so long, so why not wait a little longer?".

Proposition 3. *For abandonment distributions with a monotonically decreasing hazard rate, it is optimal to either under or over staff every shift (no matching), or to match the supply and demand rates in one of the shifts.*

Interestingly, Proposition 3 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 practical terms, Proposition 3 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, the conventional wisdom for workforce management in call centers, which is to staff just enough agents to meet projected incoming demand, 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.

To summarize, the optimal staffing policy in a system with self-scheduling agents is not straightforward, and strongly depends on both the show-up behavior of agents and the impatience distribution of customers. In particular, it may be optimal to match supply and demand in exactly one shift (monotonically increasing hazard rate), or no shift at all (monotonically decreasing hazard rate). This lies in contrast with the benchmark solution where the abandonment distribution played no role, and it is optimal to match supply and demand across all shifts. Of course, having both understaffed and overstaffed shifts in the optimal staffing policy means that the manager cannot eliminate the imbalance between supply and demand, which is due to self-scheduling, by adjusting the staffing level in her system. Thus, we need to investigate short-term controls in the system as well, in addition to the long-run staffing decision. We do so in what follows.

1.4 Short-Term Controls

It is natural to investigate how to control the **compensation**, c_j, for each shift j. In particular, we assumed in (1.1) that the agent show-up probability, r_j, was exogenously specified. In practice, r_j usually depends on the compensation offered in shift j. In order to capture how changes in c_j may impact agent show-up behavior, we assume, as in Gurvich et al. (2017), 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 $r_j \equiv G(c_j)$. We also assume that $G(\cdot)$ is log-concave with positive density function $g(\cdot)$; this will be used later to ensure uniqueness of solutions in our optimization problems.

Nevertheless, 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 certainly 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. Second, while pricing influences agents, it cannot always be used to influence the behaviour of customers, e.g., in service-oriented virtual call centers; thus, there is a need to consider 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.

In some settings, it may be possible to cap the participation of agents in certain shifts; e.g., this is the case in virtual call centers where the manager can easily choose which shifts to make available for self-scheduling agents to choose from. However, capping agent participation is restrictive, and agents usually complain when too few shifts are available. Moreover, capping may not be possible in certain settings, e.g., with ride-sharing services where drivers may already be on the road, so that it would be difficult to prohibit them from driving at different times. Thus, we do not consider such a control in this chapter. Instead, we investigate controls on the customer side, as follows.

In §1.3, we characterized the role played by the abandonment distribution in a system with randomness in capacity. Because of this, it is natural to investigate ways of controlling customer impatience to alleviate the system's cost. Here, we propose to do so via **delay announcements** in the system. We assume that the provision of delay announcements is costless for the manager, and that a single announcement is given to each delayed customer immediately upon arrival. The idea of using delay announcements as a control of customer impatience is not new. Indeed, it has been explored both empirically and analytically in several papers, albeit in contexts different from ours; for a survey of those papers, see Ibrahim (2017b). At a high level, while compensation is used to control agent joining behavior in our setting, the announcements are used to control customer behavior instead. However, it should be noted at the onset that the announcements cannot be used to restore balance in the system, i.e., entirely eliminate the cost of self-scheduling. Indeed, while delay information incites impatient customers (who would have abandoned anyway) to abandon earlier, thus leading to a reduction in waiting costs, it does not impact the overall abandonment rate in the system.

In the remainder of this chapter, we study how a manager should decide on staffing, compensation, and announcements both separately and jointly. It is unclear, a priori, what the interaction between our three controls will be, i.e., how one would affect the other. It is also unclear how a manager who has the options of making announcements and controlling compensations would staff her system: Would she use these three controls to consistently match supply and demand in all shifts? Or, would she continue to overstaff/ understaff some shifts? If so, then when?

1.4.1 Delay Announcements: Performance Impact

In this work, we contend that the announcements made must be truthful and accurate, for otherwise customers will learn to mistrust them. Studying the performance impact of delay announcements, when customers respond to these announcements by updating their abandonment distributions, is challenging. Indeed, changes in customer impatience affect system dynamics and, in turn, the future announcements made. For example, if customers abandon faster because of high announcements, then future waiting times, and future announcements which depend on those waiting times, should be shorter. When waiting times decrease, customers are inclined to be patient and wait for service, thus making delays longer again, and so on. In a nutshell, studying the impact of the announcements involves characterizing an equilibrium in the system. At a high level, an equilibrium must correspond to the long-run performance in the system, where the average announced delay coincides with the average experienced delay.

First, it is not clear whether such an equilibrium exists, or if it is unique; indeed, there may be multiple equilibria and the system may exhibit oscillations between those equilibria. We illustrate these possibilities in Figure 1.9, where we assume different customer-response functions, in each subfigure, and plot simulation-based sample paths of waiting times for each such function. (The specifics of the model are not important, and are therefore omitted.) The three subfigures, from left to right, correspond to having no equilibrium, multiple equilibria, and a unique equilibrium, respectively. Second, even when a unique equilibrium exists, it is not clear how to specify that the announcement and the corresponding delay, which are both random variables, coincide in that equilibrium, e.g., this could be in expectation, in distribution, or asymptotically when scaled in an appropriate way. Third, it is not clear how stochastic fluctuations around the equilibrium affect the system's performance. Even under Markovian assumptions, explicit analysis of the underlying birth-and-death process is analytically complex. This is so because the transition rates of the birth-and-death chain would all be dependent on the announcements. Therefore, analysis is typically done in an asymptotic regime instead. Here, we do so in the context of a fluid model, as in Armony et al. (2009) who consider a single shift instead, and a context different from ours.

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, 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 $\theta(w)$. For a fixed agent pool size, n, we let $w_j^e(n)$ denote the equilibrium delay in shift j, which is dependent on n. We must have:

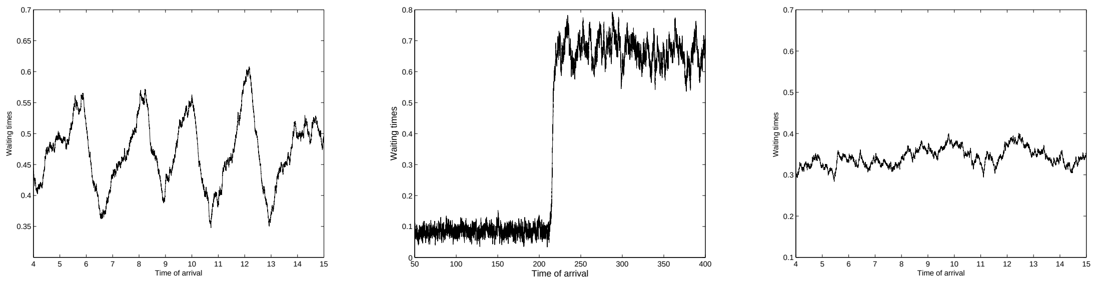

Fig. 1.9 No equilibrium, multiple equilibria, and a single equilibrium under different announcement-dependent abandonment distributions.

$$\lambda_j e^{-w_j^e(n)\theta(w_j^e(n))} = nG(c_j), \quad \text{i.e.,} \quad e^{-w_j^e(n)\theta(w_j^e(n))} = \frac{n}{\Gamma_j}, \tag{1.5}$$

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

$$C_a(n) \equiv \sum_{i=1}^{k} c_j nG(c_j) + \sum_{i=1}^{k} \left(p_j + \frac{h_j}{\theta(w_j^e(n))} \right) (\lambda_j - nG(c_j))^+. \tag{1.6}$$

Assuming that $\theta(\cdot)$ is continuous and strictly increasing, consistently with the empirical evidence in e.g., Mandelbaum and Zeltyn (2013) and Aksin et al. (2016), 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 $\lim_{w\to\infty}\theta(w) > 0$.

1.4.1.1 When Do the Announcements Reduce the Cost of Self-Scheduling?

To gain a deep understanding, it is useful to begin by exploring the performance impact of each short-term control separately. Therefore, we first assume that the manager communicates delay announcements in all congested shifts, and that compensations and the staffing level are fixed. We then ask the question: Do the announcements help in reducing the cost of self-scheduling? Naturally, the announcements are effective if they incite customers to abandon faster than they would have otherwise. In our problem, we have different Γ_j values and, consequently, different announcement-dependent abandonment rates given by (1.5). We let θ_0 denote the abandonment rate without the announcements, which is constant across all shifts. By Proposition 2, because the times to abandon are assumed to be exponentially distributed, it is optimal to critically load one shift, call it i_c, i.e., $n^* = \Gamma_{i_c}$ without the announcements. We now derive a simple sufficient condition under which the announcements lead to an overall decrease in the system's cost.

Proposition 4. *With exponential abandonment with an announcement-dependent rate $\theta(w)$, if*

$$\theta_0 \cdot \theta^{-1}(\theta_0) < \ln\left(\frac{\Gamma_{i_c+1}}{\Gamma_{i_c}}\right), \tag{1.7}$$

then $C_a(n^*) < C^*$ for $C_a(\cdot)$ in (1.6), where C^* is the optimal solution to (1.1) with $n^* = \Gamma_{i_c}$.

The condition in (1.7) means that customers do not abandon "too fast" in the absence of the announcements. This is because it can be shown, under our assumption on $\theta(\cdot)$, that the function on the left-hand-side of (1.7) is increasing in $\theta_0 \geq 0$. Thus, the condition may be equivalently interpreted as imposing a threshold, say M, on θ_0.

In Figure 1.10, we plot how the value of the threshold M varies with the degree of self-scheduling "imbalance", as measured by $\Gamma_{i_c+1}/\Gamma_{i_c}$ [2]. The area under the threshold curve corresponds to values of θ_0 for which the provision of delay announcements decreases the overall cost in the system. In other words, this is when the announcements are effective. It is interesting to note that this area increases as $\Gamma_{i_c+1}/\Gamma_{i_c}$ increases, i.e., the announcements are *increasingly* effective as self-scheduling causes a *greater* imbalance in the system.

[2] We assume that $k = 10$; $c = 1.1$; $h = 0.5$; $p = 1.0$; avg. $\lambda = 55$; $r = 0.4$; with announcements: $\theta(w) = 1.5 - e^{-2w}$.

1.4.1.2 A New Staffing Problem

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. This is because this increased congestion would, subsequently, be reduced by the announcements. To explore this, we now assume that the manager can jointly optimize the staffing level in her system, along with the announcements. The manager's staffing problem, assuming that she makes announcements in all shifts at a later stage, is given by:

$$\min_{n\in\mathbb{N}} \quad \sum_{j=1}^{k} \left(c_j nG(c_j) + \left(p_j + \frac{h_j}{\theta(w_j^e(n))} \right) (\lambda_j - nG(c_j))^+ \right), \tag{1.8}$$

where we replace the constant abandonment rate θ_0 by different announcement-dependent rates, $\theta(w_j^e(n))$, 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 (1.8), with the announcements, and n^* denote the optimal solution to (1.1), without the announcements.

Proposition 5. *With exponential abandonment with an announcement-dependent rate $\theta(w)$, if*

$$\theta_0 \cdot \theta^{-1}(\theta_0) < \min_{1\le i\le k-1} \ln(\Gamma_{i+1}/\Gamma_i), \tag{1.9}$$

then $n_a^* < n^*$.

That is, under (1.9), it is optimal for the manager to hire a smaller agent pool than without the announcements. Condition (1.9) could also be interpreted as an upper bound on θ_0, albeit a tighter one than in Proposition 4. Proposition 5 does not give an indication of the extent to which cost can be reduced by staffing a smaller pool. In Figure 1.11, we plot the percent decrease in the cost of self-scheduling as a function of the self-scheduling imbalance in the system, as measured through Γ_k/Γ_1: The higher Γ_k/Γ_1, the larger the imbalance. Figure 1.11 shows that the announcements become *increasingly* effective as the self-scheduling imbalance in the system *increases*.

To summarize, we find that delay announcements are most effective when there is a significant imbalance which arises from having "very" heterogeneous augmented arrival rates, e.g., because the agents have "very" different show-up probabilities across the different periods. This is a desirable property, because we want the announcements to be able to control the system in that case. The announcements are also effective when customers are relatively patient in the benchmark system, because delay information incites them to abandon faster.

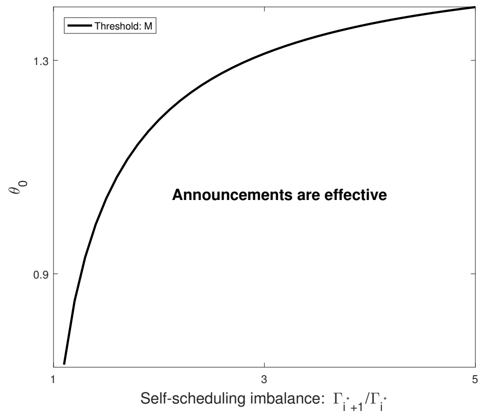

Fig. 1.10 Threshold on θ_0 as given by Proposition 4.

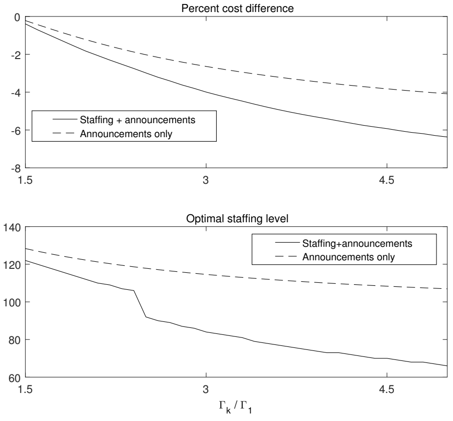

Fig. 1.11 Percent decrease in the system's cost by solving (1.8).

1.5 Joint Control of Compensation and Delay Announcements

In a ride-sharing platform, the manager may know that a concert will end shortly in a given region (say in the next 15-30 mins), and anticipate a surge in demand for Uber cars. She would then use short-term controls, e.g., pricing, to incite more agents to go to that region. She would not, however, be able to control her overall pool of drivers, i.e., hire and train more Uber drivers, because such a decision must be made weeks in advance. To mimic this situation, we assume that the staffing level is equal to n, and investigate the optimal compensation to be offered in shift k. We let l denote the minimum wage allowed in any shift. The manager may decide on c_k separately for each shift k:

$$\min_{c_k \geq l} c_k n G(c_k) + \left(p_k + \frac{h_k}{\theta} \right) (\lambda_k - nG(c_k))^+ . \tag{1.10}$$

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 $nG(c_k) = \lambda_k$. In other words, using compensation $c_k = \psi_k^n$ in shift k incites just enough agents to meet demand in that shift. It will also be convenient to define $a_k < L_k$ as follows:

$$G(a_k)\left(1 + (a_k - L_k)\frac{g(a_k)}{G(a_k)}\right) = 0. \tag{1.11}$$

The optimal compensation in problem (1.10) is given by the following lemma.

Lemma 1. *The optimal compensation in shift k, solution to (1.10), 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.*

It is not surprising that the optimal compensation in a given shift depends on the total pool size, n, and that the larger the pool, the smaller the compensation needed to incite agents to participate. In particular, 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 now turn to the more interesting case where the manager may jointly control both the provision of delay announcements and the compensation offered, in each shift. 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 continue to assume that the staffing level n is fixed, and we investigate 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 1, i.e., we assume that $n < \lambda_k/G(a_k)$ for a_k in (1.11). 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 $c_k n G(c_k) + \tilde{L}_k(\lambda_k - nG(c_k))^+$. In the following lemma, we show that $\tilde{c}_k^* = \tilde{a}_k < c_k^* = a_k$ where

$$G(\tilde{a}_k)\left(1 + (\tilde{a}_k - \tilde{L}_k)\frac{g(\tilde{a}_k)}{G(\tilde{a}_k)}\right) = 0. \tag{1.12}$$

Lemma 2. *If the manager has the option of making delay announcements, then agents receive lower compensation, i.e., they are worse off.*

Intuitively, because the manager is able to reduce congestion in the system by resorting to the announcements, she does not need to incite too many agents to participate. Thus, she offers lower compensation. In other words, instead of using high compensation to incite higher agent participation, she uses the announcements to disincentivize customer waiting instead, thereby relieving the congestion caused by self-scheduling.

1.6 Jointly Optimizing Long and Short-Term Controls

We are now ready to investigate the joint optimization problem, where the manager may use all three controls, staffing, compensation, and the announcements, at once. Here is the manager's problem when she can optimize all controls:

$$\min_{c_j \geq l, n \in \mathbb{N}} \Pi(n, \mathbf{c}) \equiv \sum_{1 \leq j \leq k} \left(c_j \cdot nG(cj) + \tilde{L}_j(\lambda_j - nG(c_j))^+ \right), \tag{1.13}$$

where $\mathbf{c} \equiv (c_1, c_2, \cdots, c_k)$ is the k-dimensional vector of compensations and, 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). Supply is capped in the overstaffed shifts. In our context, because capping is deemed undesirable and not allowed, we find that 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. In understaffed shifts, she uses the announcements. Problem (1.13) may be solved in two stages, first fixing the staffing level n (assuming that the announcements are made in every congested shift) and solving for the optimal compensation, as a function of n and, second, determining the optimal staffing level by exploiting that structure for the optimal compensation. However, the solution to (1.13) is algebraically complex with multiple shifts. Thus, we focus on two special cases: (i) the minimum wage is sufficiently low, and (ii) the minimum wage is sufficiently high.

1.6.1 Low Minimum Wage

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

$$l_0 = G^{-1}\left(\frac{\min_i\{\lambda_i\}}{\max_i\{\frac{\lambda_i}{G(\tilde{a}_i)}\}}\right) \quad \text{where} \quad \tilde{a}_k \text{ is given in } (1.12). \tag{1.14}$$

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

Lemma 3. *If the minimum wage is sufficiently low, then all shifts are either overstaffed or have matched supply and demand. Moreover, there exists at least one shift where demand and supply are matched. The manager does not resort to using delay announcements.*

Lemma 3 shows that the manager need not always resort to using the announcements. In particular, if the minimum wage is "low enough", then she will hire enough agents and offer high compensations so that $n^* G(c_i^*) \geq \lambda_i$ in each shift i, i.e., no period is congested and there is no need to resort to the announcements. 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.

1.6.2 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. In the following lemma, we show that the manager would then make announcements.

Lemma 4. *If the minimum wage is sufficiently high, then the manager uses the minimum wage in all shifts. Moreover, there exists a shift where supply and demand are matched, with the remaining shifts either under or over staffed; announcements are made in every congested shift.*

Lemma 4 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.

1.7 Conclusions

The recent and ongoing growth of the sharing economy has motivated several recent papers in the academic literature; indeed, this book is testament to that growing interest. In this chapter, we surveyed some theoretical results on the analysis of queueing systems with uncertain parameters, and described how such results may be applied for the effective management of queueing systems with self-scheduling agents. Because of the analytical complexity in such settings, queueing-theoretic approximations, which are grounded in many-server heavy-traffic limits, are useful in generating valuable insight.

Nevertheless, there remains numerous directions that are interesting to explore. For example, several modelling extensions (e.g., multiplicity of customer classes) remain to be explored. Our modelling approach was based on approximating system dynamics by using a fluid model. This is justifiable when the number of servers is binomially-distributed. In general, e.g., when there is considerable variability in agent show-up behavior or when the binomial distribution is not appropriate, there is a need to explore more refined approximations, jointly with dynamic compensation decisions and other controls. In studying the effect of delay announcements, we focused on a setting where the announcement-dependent abandonment rate is constant. In practice, customer response to the announcements tends to be non-regular, exhibiting jumps at the epochs of announcements. Developing tools to study such a response, in a setting where there is randomness in capacity, would be interesting to explore as well.

Technical Appendix

1.8 Proof of Theorem 1

1.8.1 The Overloaded Regime

1.8.1.1 $\mathscr{O}(1)$-Accuracy for the Fluid Queue Length.

We begin by establishing the asymptotic $\mathscr{O}(1)$-accuracy for the expected queue length. Let $0 < \varepsilon < r$ and define $k_1 \equiv r - \varepsilon$ and $k_2 \equiv r + \varepsilon$. Assume that ε is small enough so that $\rho r/(r+\varepsilon) > 1$. 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:

$$\begin{aligned}
|\mathbb{E}[Q_{N_\lambda}] - rn_\lambda \bar{q}_\rho| &= \left| \sum_{s\geq 0} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) - rn_\lambda \bar{q}_\rho \right| \\
&= \left| \sum_{s\geq 0} (\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right| \qquad \text{since } \mathbb{E}[N_\lambda] = rn_\lambda = \sum_{s\geq 0} s\mathbb{P}(N_\lambda = s), \\
&\leq \left| \sum_{s<k_1 n_\lambda \text{ or } s>k_2 n_\lambda} (\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right| + \left| \sum_{k_1 n_\lambda \leq s \leq k_2 n_\lambda} (\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right| .
\end{aligned}$$

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

$$A_\lambda \equiv \left| \sum_{s<k_1 n_\lambda \text{ or } s>k_2 n_\lambda} (\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right| \text{ and } B_\lambda \equiv \left| \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} (\mathbb{E}[Q_s] - s\bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right|.$$

Asymptotic bound for N_λ far from $n_\lambda r$.

We begin by showing that A_λ is asymptotically negligible.

Lemma 5. $\lim_{\lambda\to\infty} A_\lambda = 0$.

Proof. We can write,

$$\begin{aligned} A_\lambda &= \left| \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) - \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} s\bar{q}_\rho \mathbb{P}(N_\lambda = s) \right|, \\ &\le \mathbb{E}[Q_0] \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} \mathbb{P}(N_\lambda = s) + \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} s\bar{q}_\rho \mathbb{P}(N_\lambda = s). \end{aligned}$$

Also, define $A_\lambda^{(1)} \equiv \mathbb{E}[Q_0]\sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} \mathbb{P}(N_\lambda = s)$ and $A_\lambda^{(2)} \equiv \sum_{s>k_2 n_\lambda \text{ or } s<k_1 n_\lambda} s\bar{q}_\rho \mathbb{P}(N_\lambda = s)$. 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., $\mathbb{E}[Q_0] = \mathscr{O}(\lambda)$. Applying Hoeffding's inequality to the binomial distribution: $\mathbb{P}(k_1 n_\lambda \le N_\lambda \le k_2 n_\lambda) \ge 1 - 2e^{-2\varepsilon^2 n_\lambda}$; equivalently, $\mathbb{P}(k_1 n_\lambda > N_\lambda \text{ or } N_\lambda > k_2 n_\lambda) \le 2e^{-2\varepsilon^2 n_\lambda}$. Thus,

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

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:

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

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

Proof. We begin by writing B_λ as follows,

$$B_\lambda \le \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} \left| \mathbb{E}[Q_s] - s\bar{q}_{\rho_s} \right| \mathbb{P}(N_\lambda = s) + \left| \sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} s(\bar{q}_{\rho_s} - \bar{q}_\rho)\mathbb{P}(N_\lambda = s) \right|, \tag{1.15}$$

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,

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

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

$$\limsup_{\lambda \to \infty} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}| \le \sqrt{f(\bar{w}_{\rho_s})} \left(\frac{3|f'(\bar{w}_{\rho_s})|}{\rho_s f^2(\bar{w}_{\rho_s})} + 1/2 \right), \tag{1.16}$$

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 $\rho r/(r+\varepsilon) \le \rho_s \le \rho r/(r-\varepsilon)$, we have that $\bar{w}_{\rho r/(r+\varepsilon)} \le \bar{w}_{\rho_s} \le \bar{w}_{\rho r/(r-\varepsilon)}$. By the continuity of the bounding function in (1.16) and the boundedness theorem, we conclude that there exists a finite constant $C_1 > 0$ such that

$$\sup_{k_1 n_\lambda \le s \le k_2 n_\lambda} \sqrt{f(\bar{w}_{\rho_s})} \left(\frac{3|f'(\bar{w}_{\rho_s})|}{\rho' f^2(\bar{w}_{\rho_s})} + 1/2 \right) \le C_1. \tag{1.17}$$

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 (1.16) and (1.17) 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):

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

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 $\mathscr{O}(1/\lambda)$. Also, $g_\lambda(n_\lambda r) = 0$ and

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

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:

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

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 $\mathbb{E}[g_\lambda^2(N_\lambda)] = \mathscr{O}(\lambda^2)$ since $\int_{(\bar{F})^{-1}(1/\rho)}^{(\bar{F})^{-1}(N_\lambda/n_\lambda r\rho)} \bar{F}(u)\,\mathrm{d}u \leq 1/\theta$. 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 $\limsup_{\lambda\to\infty} B_\lambda \leq C$.

$\mathscr{O}(1)$-accuracy.

Since both A_λ and B_λ are asymptotically bounded, there must exist $K > 0$ such that, as desired:

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

1.8.1.2 $o(1)$-Accuracy for the Fluid Net Abandonment Rate.

The proof for the net abandonment rate proceeds along similar lines, so we will be brief. Paralleling (1.16), 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 $\mathbb{E}[N_\lambda] = rn_\lambda$.

1.8.2 The Underloaded Regime

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

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

As in the proof of Theorem 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, $\sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) \le \mathbb{E}[Q(k_1 n_\lambda)]\sum_{k_1 n_\lambda \le s \le k_2 n_\lambda} \mathbb{P}(N_\lambda = s)$. 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.

1.8.3 The Critically-Loaded Regime

We condition on N_λ:

$$\begin{aligned}\mathbb{E}[Q_{N_\lambda}] &= \sum_{k_1 n_\lambda \le s < n_\lambda r} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s) + \mathbb{E}[Q_{n_\lambda r}]\mathbb{P}(N_\lambda = n_\lambda r),\\ &\le \sum_{k_1 n_\lambda \le s < n_\lambda r} |\mathbb{E}[Q_s] - s\bar{q}_{\rho_s}|\mathbb{P}(N_\lambda = s) + \sum_{k_1 n_\lambda \le s < n_\lambda r} s\bar{q}_{\rho_s}\mathbb{P}(N_\lambda = s) + \sum_{n_\lambda r < s \le k_2 n_\lambda} \mathbb{E}[Q_s]\mathbb{P}(N_\lambda = s)\\ &\quad + \mathbb{E}[Q(n_\lambda r)]\mathbb{P}(N_\lambda = n_\lambda r), \end{aligned} \tag{1.18}$$

where $\rho_s = r\rho n_\lambda / s$. Paralleling (1.16) and (1.17), 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,

$$\begin{aligned}\sum_{k_1 n_\lambda \le s < n_\lambda r} s\bar{q}_{\rho_s}\mathbb{P}(N_\lambda = s) &= \sum_{k_1 n_\lambda \le s < n_\lambda r} n_\lambda r \left(\int_0^{(\bar{F})^{-1}(s/n_\lambda r)} \bar{F}(x)\,\mathrm{d}x\right)\mathbb{P}(N_\lambda = s)\\ &= \mathbb{E}\left[\left(n_\lambda r \int_0^{(\bar{F})^{-1}(N_\lambda/n_\lambda r)} \bar{F}(x)\,\mathrm{d}x\right)\mathbb{1}(N_\lambda \in [k_1 n_\lambda, n_\lambda r))\right].\end{aligned} \tag{1.19}$$

Using arguments as in Theorem 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 (1.18) is $\mathscr{O}(\sqrt{\lambda})$. The proof for the abandonment rate follows along similar lines, so we omit the relevant details.

1.9 Proofs of Propositions

Proposition 2:

If the abandonment distribution is exponential, then for $\Gamma_{i-1} \le n < \Gamma_i$, $u_i(n) = \sum_{j=1}^k c_j r_j n + \sum_{j=i}^k (p_j + h_j/\theta)(\lambda_j - n r_j)$. We assume:

$$\sum_{j=1}^k c_j r_j - \sum_{j=i_0}^k L_j r_j < 0 \text{ and } \sum_{j=1}^k c_j r_j - \sum_{j=i_0+1}^k L_j r_j > 0. \tag{1.20}$$

Clearly, under condition (1.20), $C(n)$ is piecewise linear with piecewise negative slopes for $n \leq \Gamma_{i_0}$, and strictly positive slopes for $n > \Gamma_{i_0}$.
With a monotonically increasing hazard rate, we have

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

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 3:

In $[\Gamma_{i-1}, \Gamma_i)$, $u_i'(n)$ is as in the proof of Proposition 2, 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:

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.

Proposition 5:

It suffices to show that, for all n, $C'_a(n) > C'(n)$. If this holds, then $C'(n) > 0$ would imply $C'_a(n) > 0$, so that $C_a(\cdot)$ increases whenever $C(\cdot)$ increases, which leads to $n^*_a < n^*$. Fix $n \in [\Gamma_{i_0-1}, \Gamma_{i_0})$, for some i_0. Then, $C_a(n) = \sum_{i=1}^k c_i n r_i + \sum_{i=i_0}^k \left(p_i + \frac{h_i}{\theta(w^e_i)}\right)(\lambda_i - nr_i)$ where $e^{-w^e_i\theta(w^e_i)} = n/\Gamma_i$. That is, for $i \geq i_0$,

$$w^e_i\theta(w^e_i) = \ln\left(\frac{\Gamma_i}{n}\right) > \ln\left(\frac{\Gamma_i}{\Gamma_{i_0}}\right) > \min_{1\leq i\leq k-1} \ln\left(\frac{\Gamma_{i+1}}{\Gamma_i}\right) > \theta_0 \cdot \theta^{-1}(\theta_0),$$

under condition (1.9). This implies that $\theta(w^e_i) > \theta_0$ for all $i \geq i_0$. Note that for $n \in [\Gamma_{i_0-1}, \Gamma_{i_0})$,

$$C'_a(n) = \sum_{i=1}^k c_i r_i - \sum_{i=i_0}^k p_i r_i - \sum_{i=i_0}^k \frac{r_i h_i}{\theta(w^e_i)} + \sum_{i=i_0}^k h_i(\lambda_i - nr_i)\frac{\theta'(w^e_i)}{n\theta^2(w^e_i)(\theta'(w^e_i)w^e_i + \theta(w^e_i))}.$$

Thus, assuming condition (1.9) implies that $C'_a(n) > C'(n)$, for every $n \in [\Gamma_{i_0-1}, \Gamma_{i_0})$. Since we can let i_0 denote any period index, we obtain that $n^*_a < n^*$.

Lemma 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 2:

We let $\tilde{a}_k$ be the solution to (1.12). 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 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 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 3:

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 4:

In this case, problem (1.13) simplifies to:

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

Note that $\Pi(n)$ is piecewise linear. Then, $\Pi'(n) = klG(l) - \sum_{\{i:n<\frac{\lambda_i}{G(l)}\}} \tilde{L}_i G(l)$. 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 $n^* = \frac{\lambda_{k_0}}{G(l)}$.

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* .

Akşin, Zeynep, Barış Ata, Seyed Morteza Emadi, Che-Lin Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Science* **59**(12) 2727–2746.

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.

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

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. 2010a. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Bassamboo, Achal, Ramandeep S Randhawa, Assaf Zeevi. 2010b. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Bassamboo, Achal, Assaf Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57**(3) 714–726.

Dong, Jing, Rouba Ibrahim. 2017. Fexible workers or full-time exmployees? on staffing service systems with a blended workforce. Northwestern University, 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.

Gans, Noah, Haipeng Shen, Yong-Pin Zhou, Nikolay Korolev, Alan McCord, Herbert Ristock. 2015. Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing & Service Operations Management* **17**(4) 571–588.

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

Garnett, Ofer, Avishai Mandelbaum, Martin Reiman. 2002b. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Gurvich, I., M. Lariviere, T. Moreno-Garcia. 2017. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Northwestern University, working paper.

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

Ibrahim, Rouba. 2017a. Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management* .

Ibrahim, Rouba. 2017b. Sharing delay information in service systems: A literature survey. *Queueing Systems* .

Ibrahim, Rouba, Pierre L'Ecuyer. 2013. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management* **15**(1) 72–85.

Kang, Weining, Kavita Ramanan, et al. 2010. Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* **20**(6) 2204–2260.

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

Shen, Haipeng, Jianhua Z Huang. 2008. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management* **10**(3) 391–410.

Vocalcom. 2014. Virtual call center industry projected to more than quadruple - are you ready? URL `http://www.vocalcom.com/en/blog/customer-service/virtual-call-center-industry-projected-to-more-than-quadruple-are-you-ready/`.

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

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

Whitt, Ward. 2006c. Fluid models for multiserver queues with abandonments. *Operations research* **54**(1) 37–54.

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.

Zhang, Jiheng. 2013. Fluid models of many-server queues with abandonment. *Queueing Systems* **73**(2) 147–193.