Prediction-Driven Surge Planning with Application in the Emergency Department

Yue Hu, Carri W. Chan, Jing Dong
Decision, Risk, and Operations, Columbia Business School
yhu22@gsb.columbia.edu, cwchan@gsb.columbia.edu, jing.dong@gsb.columbia.edu

Determining emergency department (ED) nurse staffing decisions to balance the quality of service and staffing cost can be extremely challenging, especially when there is a high level of uncertainty in patient-demand. Increasing data availability and continuing advancements in predictive analytics provide an opportunity to mitigate demand uncertainty by utilizing demand forecasts. In this work, we study a two-stage prediction-driven staffing framework where the prediction models are integrated with the base (made weeks in advance) and surge (made nearly real-time) staffing decisions in the ED. We quantify the benefit of having the ability to use the more expensive surge staffing and identify the importance of balancing demand uncertainty versus demand stochasticity. We also propose a near-optimal two-stage staffing policy that is straightforward to interpret and implement. Lastly, we develop a unified framework that combines parameter estimation, real-time demand forecasts, and capacity sizing in the ED. High-fidelity simulation experiments for the ED demonstrate that the proposed framework can reduce annual staffing costs by 11%–16% ($2 M–$3 M) while guaranteeing timely access to care.

1. Introduction

Emergency department (ED) crowding is a significant problem in many countries around the world, leading to adverse effects on patient outcomes, patient satisfaction, and staff morale (Bernstein et al. 2009). Nurses provide a substantial portion of patient care and are often a bottleneck resource in the ED (Green 2010). Inadequate nurse staffing is found as a major contributor to significant increase in the waiting time experienced by patients and the percentage of patients who leave without been seen (LWBS) (Ramsey et al. 2018). In addition, nursing costs comprise a substantial fraction of hospital operating budgets. Therefore, developing effective nurse staffing policies to ensure timely access to care is of great importance.

Optimally balancing the ED nurse staffing levels to ensure good quality of service versus increasing staffing costs can be extremely challenging. One of the major complication comes from the high level of uncertainty in patient demand and the relative static nature of ED staffing decisions. Poisson processes have been standard assumptions in modeling the arrival processes in service systems due to their analytical tractability. Their validity has also been statistically verified in some healthcare settings (Kim and Whitt 2014). However, several recent empirical studies suggest the presence of a higher level of uncertainty (dispersion) relative to standard Poisson processes in real ED arrival data (Maman 2009, Armony et al. 2015), and in other service systems such as call centers (Brown et al. 2005, Steckley et al. 2009, Zhang et al. 2014). Random events such as weather conditions, level of flu circulation, and mass causality incidents can cause a high level of fluctuation in ED demand. On the other hand, ED staffing decisions are often made well ahead of time and the staffing level is difficult (or very expensive) to change in real time (Chan et al. 2021). In particular, it is common for EDs to divide a day into multiple nursing shifts. In the United States, there are usually two 12-hour nursing shifts, with the day shift lasting from 7am to 7pm, and the night shift from 7pm to 7am the next day. As a typical practice, a "base" staffing level, which consists of the majority of the staff, is determined several weeks in advance, when the actual demand is largely unknown. This allows the nurses to plan their working schedule ahead of time. As time approaches to several hours before the shift, if the ED manager senses a surge in patient volume, he/she can add an extra level of "surge" staffing by calling in overtime or agency nurses at a higher compensation (e.g., overtime salary). The nurse staffing level is then held more or less at a constant level throughout the shift. The surge staffing provides some flexibility to cope with the demand surge, but there is currently a lack of systematic guidelines in how to optimally utilize this partial flexibility.

Meanwhile, in recent years, increasing data availability and continuing development in statistical learning tools provide an emerging opportunity to mitigate demand uncertainty by building advanced demand forecast models. There have been considerable efforts devoted to developing prediction models for ED patient volume and flow (see, e.g., Marcilio et al. (2013), Calegari et al. (2016), Chang et al. (2018), Whitt and Zhang (2019), Bertsimas et al. (2021)). However, despite the vast amount literature on demand forecasts, how to effectively incorporate the predictive information to improve ED staffing decisions is less studied. In particular, while advanced prediction models that utilize real-time information generate more accurate short-term forecast of the ED demand in comparison to using traditional historical averages (Schweigler et al. 2009), it remains unclear how the increased prediction accuracy can be translated to improved system performance (e.g., reduction in patient waiting time and LWBS rate) and/or reduced staffing costs. In this paper, we study prediction-driven surge planning. The key tradeoff in this two-stage staffing problem is the long-term staffing commitments which have a lower costs but face a higher level of demand uncertainty (larger prediction error) versus the short-term staffing commitments which have a higher cost but face a lower level of demand uncertainty (smaller prediction error).

To capture the highly uncertain demand faced by the ED, we assume that patients arrive according to a doubly stochastic Poisson process as in Maman (2009), Bassamboo et al. (2010), Koçağa et al. (2015). The arrival rate for a particular type of shift is a random variable that takes the form of

$$\Lambda = \lambda + \lambda^{\alpha} X, \tag{1}$$

where λ is the mean arrival rate, $\alpha \in (0,1)$ captures the order of arrival-rate uncertainty, and X is a random variable with zero mean. At the base-stage, our prediction model is only able to capture the long-run average pattern that defines the type of the shift, e.g., day of the week effect and day versus night effect. Thus, we we assume the base-stage prediction model predicts $\mathbb{E}[\Lambda] = \lambda$ accurately. At the surge stage, as we gather more real-time information, we can build more sophisticated prediction models. Motivated by value of real-time information identified in Hu et al. (2021), we assume in our main model that the surge-stage prediction model is able to predict the realized arrival rate $\ell = \lambda + \lambda^{\alpha} x$ where x is a particular realization of X for the specific shift. Conditional on ℓ, the ED operates as a Markovian multi-server queue with Poisson arrival process, exponentially distributed service times, and exponentially distributed patience times. Note that even with the predictive information, we still incur a certain level of uncertainty due to the randomness in the interarrival times between patients, patients' service requirements, and their patience times (time before abandoning).

The ED manager makes two staffing decisions for each shift: a base staffing level and a surge staffing level. The base staffing decision is based on the base prediction, i.e., λ, and knowledge of the arrival rate distribution, i.e., the distribution of $\lambda^{\alpha} X$. The surge staffing decision is based on the surge prediction, i.e., ℓ. The surge staff are assumed to be more costly than the base staff. Our objective is to minimize the sum of the staffing cost and the performance cost which consists of the costs incurred by patients' waiting and patients' LWBS. Our main contributions can be summarized as:

The benefit of surge staffing. To quantify the benefit of having the more expensive surge staff, we compare the two-stage stochastic optimization problem to a single-stage benchmark where only base staffing is allowed. We quantify the cost saving of the optimal two-stage staffing rule over the optimal single-stage policy. Our result shows that the magnitude of cost-saving depends on the order of arrival-rate uncertainty captured by α in (1). In particular, the cost saving is $o(\sqrt{\lambda})$ if $\alpha < 1/2$, $O(\sqrt{\lambda})$ if $\alpha = 1/2$, and $\Theta(\lambda^{\alpha})$ if $\alpha > 1/2$. As we will explain in more details, the three regimes of cost saving are divided by the interplay between the order of arrival-rate uncertainty, which is $O(\lambda^{\alpha})$, and stochastic variability in patient arrival and services, which is $O(\sqrt{\lambda})$. The cost-saving quantification suggests that surge staffing is most beneficial when the arrival-rate uncertainty dominates the system stochasticity, i.e., $\alpha > 1/2$. In this regime, the larger the arrival-rate uncertainty, the more cost saving we gain from the flexibility of surge staffing.

Near-optimal two-stage staffing rule. Focusing on the regime where the arrival-rate uncertainty dominates the system stochasticity, i.e., $\alpha > 1/2$, we propose a near-optimal two-stage staffing rule that is easy to interpret and implement. In particular, at the base stage, the base staffing level is set to meet the mean demand, together with a hedging that is of the same order as the arrival-rate uncertainty. After the random arrival rate is realized at the surge stage, the surge staffing level is brought up to meet the realized offered load, together with a hedging against the stochastic variability catered to the realized arrival rate. The parameters of the staffing rule, which dictate the amount of hedging, are the optimal solutions to a two-stage newsvendor problem, which can be viewed as a stochastic-fluid approximation to the optimal staffing problem, and the optimal solutions to a square-root staffing problem based on a diffusion approximation of the queue length process. We prove that our proposed policy has an optimality gap of $o(\sqrt{\lambda})$ compared to the exact two-stage optimum. We also extend the two-stage staffing rule to allow more general prediction errors at the surge stage. In particular, we consider the case where we are not able to predict the realized arrival rate ℓ exactly. Instead, we may incur different levels of prediction error. We quantify how prediction error affect the staffing rule and its corresponding performance.

Practical insights and ED implementation. To facilitate real-world implementation, we propose an integrated framework to implement the two-stage staffing policy in the ED, which includes 1) parameter estimation, 2) a two-stage prediction model, and 3) a two-stage capacity sizing rule. Using data from the ED in New York Presbyterian Columbia University Medical Center (NYP CUMC), we estimate its arrival-rate uncertainty to be $\alpha = 0.769$. We then build a two-stage prediction model to inform the staffing policy. At the base stage, a simple linear regression model that incorporates the day of the week and day v.s. night effect works well to estimate the mean arrival rates. For the surge stage, we implement a recently developed prediction model in Hu et al. (2021), which utilizes concurrent information such as weather, patient comorbidity profile, ED congestion level, etc. Lastly, we extend the two-stage staffing rule developed based on the parsimonious queueing model to accommodate various realistic patient-flow characteristics in our collaborating ED. These features include lognormal length-of-stay distribution, hour-of-the-day variability in patient arrival rate, and transient system dynamics as the day and night shifts alternate, each lasting for only 12 hours. We leverage the insights from our theoretical analysis to appropriately adapt our staffing approach to this more realistic setting. In this case, we extend our two-stage staffing rule to make the surge staffing decision based on not only the predicted arrival rate (as is the case in the parsimonious theoretical setting) but also on the concurrent queue length information at the beginning of each shift. These two sources of real-time information (i.e., demand forecast and current system state) lead to significant cost savings from the two-stage staffing policy over the benchmark single-stage policy. For example, compared to the newsvendor solution (Bassamboo et al. 2010), our policy achieves a reduction of 16% ($3 M) in the annual staffing cost while the average waiting time is kept below 30 minutes.

1.1. Related Literature

Classic square-root staffing rule. The standard stream of capacity planning problems for service operations focuses on systems where model parameters are exactly known. In this setting, the square-root staffing principle dates back to Erlang (1917) in the study of automatic telephone exchanges. The principle is more recently explained based on an infinite-server queue heuristic in Kolesar and Green (1998). In particular, it is shown that the stochastic fluctuation of the system is of square root order of the offered load. Thus, the square-root staffing can be viewed as an uncertainty hedging against system stochasticity. Halfin and Whitt (1981) establish a formal diffusion limit for $M/M/N$ queues under the square-root staffing as the arrival rate goes to infinity. Borst et al. (2004) further establishes that the square-root staffing rule optimally balances the staffing cost and the service quality. For this reason, the many-server asymptotic scaling under the square root staffing is often referred to as the quality-and-efficiency driven (QED) regime. A few extensions have been considered to incorporate features not captured by the $M/M/N$ model. Garnett et al. (2002) generalize the diffusion limit under the square-root staffing to the $M/M/N+M$ queue where customers abandon the system after an exponentially distributed patience time; more general patience time distributions are considered in Mandelbaum and Zeltyn (2009). Jennings et al. (1996) and Liu and Whitt (2012) extend the square-root staffing rule to systems with time-varying arrival rates. Our work extends this stream of literature by allowing the arrival-rate to be random and considering a two-stage staffing problem in two time scales. Relevantly, after the random arrival rate is realized at the surge stage, our proposed two-stage QED staffing rule brings the total staffing level up to the square-root staffing prescription if the base-stage capacity is inadequate. In addition, similar to the literature, our theoretical analysis takes an asymptotic approach, where we send the mean arrival rate λ to infinity and study how the optimal staffing level scales with λ.

Managing queues with parameter uncertainty. Motivated by the high level of demand uncertainty in many service systems, more sophisticated models for arrival processes that account for features not captured by standard Poisson processes have been proposed in the literature. Whitt (1999) is one of the first to study a random arrival rate for call centers and its implications on the staffing decision. Chen and Henderson (2001), Avramidis et al. (2004), Brown et al. (2005) and Steckley et al. (2009) provide empirical evidence of arrival-rate uncertainty and explore its modeling implications. Maman (2009) finds empirical evidence of high arrival-rate uncertainty in an Israeli ED. Our work is closely related to works that study the staffing decision in the presence of arrival-rate uncertainty. Whitt (2006) investigates a fluid-based staffing prescription catered to arrival-rate uncertainty and absenteeism of servers. Harrison and Zeevi (2005) and Bassamboo et al. (2010) propose a newsvendor-based solution method whose effectiveness is pronounced when the order of arrival-rate uncertainty is larger than stochastic variability. Their proposed staffing rule is set to meet the mean demand plus a hedging against the arrival-rate uncertainty. More recently, moving from single-stage to two-stage decisions, Koçağa et al. (2015) formulate a joint staffing and co-sourcing problem, where the staffing decision is made before the random arrival-rate is realized, and the co-sourcing decision is made in real time after the arrival-rate uncertainty is resolved. Our two-stage optimization problem has similar decision epochs to those in Koçağa et al. (2015), i.e., before and after the random demand is realized. However, different from Koçağa et al. (2015), we consider a two-stage staffing problem and allow the arrival-rate uncertainty to be of a larger magnitude than stochastic variability. The solution method we use to solve the two-stage stochastic optimization problem leverages the stochastic fluid approximation introduced by Harrison and Zeevi (2005), but we considered a more refined version of this approximation, which takes the system stochasticity into account at the surge stage.

Predictive analytics and data-driven methods in capacity sizing. Several works take a data-driven approach for capacity sizing with demand uncertainty. Zheng et al. (2018) and Sun and Liu (2021) propose statistical methods to estimate the arrival-rate distribution. See also Ibrahim et al. (2016) for a comprehensive review of literature on modeling and forecasting for call center arrivals. Bassamboo and Zeevi (2009) develop a data-driven approach that yields staffing prescriptions that are asymptotically optimal, as both the system scale and data size increase to infinity. There is a large literature on studying demand uncertainty in inventory systems without queueing dynamics (see for example (Chen et al. 2007, Perakis and Roels 2008, Levi et al. 2015, Ban and Rudin 2019, Boada-Collado et al. 2020)). Motivated by the operations of EDs, our work takes into account the arrival-rate distribution at the base stage, the demand visibility at the surge stage, and the stochasticity of queueing dynamics.

ED capacity planning Our work relates to the growing literature on using queueing theory to address capacity planning problems in the ED. Green et al. (2006) models the ED as an $M_t/M/s$

queue and use a Lag SIPP (stationary independent period by period) approach to gain insights into the staffing prescriptions. Yankovic and Green (2011) develop a finite source queueing model with two types of severs—nurses and beds—to study the interplay between bed occupancy level and demand for nursing. Véricourt and Jennings (2011) study nurse staffing using a closed queueing model, where patients alternate between being needy of service and stable without service need. Similar patient reentrant behavior is studied by Yom-Tov and Mandelbaum (2014) using an Erlang-R model in time-varying environments. Chan et al. (2021) use a multiclass queue to study the dynamic assignment of nurses to different areas of the ED at the beginning of each shift. Batt et al. (2019) empirically investigate the impact of discrete work shifts on service rates and patient handoffs (i.e., passing patients in treatment to the next care provider at the end of a shift). Compared to the literature, we focus on studying the effect demand uncertainty on ED staffing, where we combine demand prediction with queueing dynamics to derive good staffing strategies.

Dual sourcing problem in supply chain management. Though our work is motivated by the staffing problem for service systems, a similar core tradeoff between cost and responsiveness arises in dual sourcing inventory systems, in which one supplier is cheaper but slower, while the other is more costly but faster. In this setting, a tailored base-surge (TBS) sourcing policy is found to be effective in both continuous and periodic review models (Allon and Van Mieghem 2010, Janakiraman et al. 2015). Xin and Goldberg (2018) formally prove that the TBS policy is asymptotically optimal as the lead time of the cheaper supplier grows without bound. Different from the dual sourcing problem, our theoretical framework further incorporates queueing dynamics into the optimization problem. We show that the cost saving of our proposed policy increases with the order of arrival-rate uncertainty.

1.2. Organization

The rest of the paper is organized as follows. In Section 2 we introduce the model and formulate the two-stage staffing problem. In Section 3 we quantify the cost saving from surge staffing. In Section 4 we propose near-optimal two-stage staffing rules that are easy to interpret and implement. The optimality gap between the proposed policy and the exact two-stage optimum is also derived. The performance of the two-stage staffing rule is further illustrated through numerical experiment in Section 5, where we compare the performance of our proposed staffing rule to several benchmark policies. In Section 6, we extend the two-stage staffing rule to accommodate more general prediction errors at the surge stage. Lastly, in Section 7, we develop a holistic framework to implement the prediction-driven staffing policy in the actual ED, which includes parameter estimation, demand forecast, and capacity sizing that takes the transient shift effect into account. We conclude in Section 8. All the proofs appear in the appendix.

1.3. Notation

For a sequence of positive real numbers $\{a^n : n \in \mathbb{R}_+\}$ and a sequence of real numbers $\{b^n : n \in \mathbb{R}_+\}$, we write (i) $b^n = o(a^n)$ if $|b^n/a^n| \to 0$ as $n \to \infty$, (ii) $b^n = O(a^n)$ if $|b^n/a^n|$ is bounded from above, and (iii) $b^n = \Theta(a^n)$ if $|b^n/a^n|$ is bounded from above and from below by a strictly positive real number, i.e., if $m \le |b^n/a^n| \le M$ for some $0 < m < M < \infty$ for all $n > 0$. For a sequence of random variables $\{X^n : n \in \mathbb{R}_+\}$ and a sequence of positive real numbers $\{a^n : n \in \mathbb{R}_+\}$, we write (i) $X^n = o(a^n)$ if $|X^n/a^n| \to 0$ as $n \to \infty$ with probability 1, and (ii) $X^n = o_{UI}(a^n)$ if $X^n = o(a^n)$ and there exists some random variable Y with $\mathbb{E}[Y] < \infty$ such that $|X^n/a^n| < Y$ for all $n > 0$.

2. The Model

To gain insights into the potential benefits of two-stage staffing, we start with a stylized model of the ED using a parsimonious multi-server queueing system where patients arrive according to a doubly stochastic Poisson process. The arrival rate for a shift Λ is a random variable with cumulative distribution function F_Λ and mean $\mathbb{E}[\Lambda] = \lambda$. Conditional on Λ, the arrival process is a homogeneous Poisson process with that rate. Customers (patients) are served on a first-come first-served (FCFS) basis, and wait in an infinite capacity buffer when all servers (nurses) are busy. While waiting for service, a delayed patient abandons the system (LWBS) after an exponentially distributed amount of time with mean $1/\gamma$. Patients have service requirements that are independently and identically distributed (i.i.d.) exponential random variables with mean $1/\mu$. Hence, conditioned on Λ, the ED operates as an $M/M/N+M$ queue, where the staffing level N is the decision variable.

The ED manager makes two decisions: an upfront base staffing level and a surge staffing level, both of which are non-negative integers. At the base stage, which is often a few weeks/months before the start of the actual shift, the prediction model can only predict the average arrival rate level, λ. We assume the arrival rate distribution is known. Thus, the base staffing level $N_1 := N_1(F_\Lambda) \in \mathbb{N}$ is made before the arrival rate is realized, based on knowledge of the arrival rate distribution, F_Λ, only. At the surge stage, as we gather more real-time information, the prediction model can predict the realized arrival rate ℓ quite accurately. Thus, the surge staffing level $N_2(N_1, \ell) \in \mathbb{N}$ is made based on the base staffing level, N_1, and the realized arrival rate, ℓ. We do not allow $N_2(N_1, \ell)$ to take negative values, because in reality, the ED manager cannot make a last-minute decision to reduce the staffing level, e.g., by canceling shifts for the nurses who are staffed at the base stage. We denote the joint staffing decision as $\pi := (N_1, N_2(N_1, \ell))$, and use Π to denote the set of all feasible staffing rules. Note that in this parsimonious model, the prediction at the base stage is captured by the expected arrival rate, $\lambda := \mathbb{E}[\Lambda]$, and the prediction errors are captured by the distribution of $\Lambda - \lambda$. To start, we assume perfect prediction at the surge stage. We will relax this assumption in Section 6 to explicitly incorporate prediction errors at the surge stage.

There are costs associated with patients' waiting, patients' LWBS (abandonments), and staffing In particular, a holding cost is incurred at a rate of h per patient per unit time spent waiting. Each abandoning patient incurs a fixed cost of a. The staffing cost is c_1 per base server per unit time, and c_2 per surge server per unit time. Let $Q(n, \ell)$ denote the steady-state queue length of an $M/M/n+M$ queue with arrival rate ℓ. Then, we consider the following two-stage cost minimization problem.

$$\min_{\pi \in \Pi} \mathcal{C}_\pi = \min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,\Lambda)} \{c_2 N_2(N_1,\Lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1 + N_2(N_1,\Lambda),\Lambda)|\Lambda\right]\} \right] \right\}. \tag{2}$$

For an $M/M/n+M$ queue with arrival rate ℓ, $\gamma\mathbb{E}[Q(n,\ell)]$ is the steady-state abandonment rate. Thus, $a\gamma\mathbb{E}[Q(n,\ell)]$ captures the abandonment cost while $h\mathbb{E}[Q(n,\ell)]$ captures the holding cost in steady state. Note that there are two expectations in (2). The inner expectation is taken with respect to the stochasticity in the steady-state queue length, i.e., randomness in $Q(n,\Lambda)$ conditional on $\Lambda = \ell$. The outer expectation is taken with respect to the arrival-rate uncertainty, i.e., randomness in Λ.

2.1. Parameter Regime

It makes intuitive sense that if the waiting and abandonment costs are excessively lower than the staffing costs, there is no motivation to staff any server. In addition, if the base staffing cost is higher than the surge staffing cost, i.e., $c_1 > c_2$, it is cost-effective to staff all servers at the surge stage when the arrival-rate uncertainty is resolved. This intuition is formalized in Proposition 1.

Proposition 1 *For the optimal solution $(N_1^*, N_2^*(N_1^*,\Lambda))$ to problem (2):*

(I) If $\min\{c_1, c_2\} > h\mu/\gamma + a\mu$, then $N_1^* = 0$ and $N_2^*(N_1^*,\Lambda) = 0$.

(II) If $\min\{c_1, h\mu/\gamma + a\mu\} > c_2$, then $N_1^* = 0$.

(III) If $c_2 > h\mu/\gamma + a\mu > c_1$, then $N_2^*(N_1,\Lambda) = 0$ for any base staffing level N_1.

Based on Proposition 1, the cost parameters can be divided into four regimes as summarized in Table 1.

Table 1 Optimal staffing combination for different cost parameters

Cost parameters	Staffing decisions
$\min\{c_1, c_2\} > h\mu/\gamma + a\mu$	No staffing
$\min\{c_1, h\mu/\gamma + a\mu\} > c_2$	Complete surge staffing
$c_2 > h\mu/\gamma + a\mu > c_1$	Complete base staffing
$h\mu/\gamma + a\mu > c_2 > c_1$	Base + surge staffing

In this paper, we are interested in the non-trivial regime that provides motivation to staff both base and surge servers.

Assumption 1 *The cost rates satisfy* $h\mu/\gamma + a\mu > c_2 > c_1$.

2.2. Arrival-Rate Uncertainty

Solving (2) explicitly is challenging due to the two sources of randomness. In addition, $\mathbb{E}\left[Q(N_1 + N_2(N_1, \ell), \ell)\right]$ has no closed-form expression. To gain analytical insights, we take an asymptotic approach by sending the mean arrival rate λ to infinity and study how the optimal staffing rule scales with λ.

To facilitate the theoretical development, we assume that the random arrival rate takes the form

$$\Lambda = \lambda + X\lambda^{\alpha}\mu^{1-\alpha}, \tag{3}$$

for some constant $\alpha \in (0,1)$ and random variable X with $\mathbb{E}\left[|X|\right] < \infty$. Note that because $\mathbb{E}[\Lambda] = \lambda$, $\mathbb{E}[X] = 0$. Note that (3) is equivalent to the form of arrival-rate uncertainty introduced in (1) earlier; we factor out $\mu^{1-\alpha}$ to facilitate technical derivations. Let F_X denote its cumulative distribution function (cdf) of X. We also assume that X has a proper probability density function (pdf). The second term in (3) captures the stochastic fluctuation of the arrival rate around its mean. It is further decomposed into two parts: X and $\lambda^{\alpha}\mu^{1-\alpha}$, where the second part captures the order of fluctuation in relation to λ. We refer to the exponent α as the *order of arrival-rate uncertainty*. Random arrival rate of the form (3) is proposed in Maman (2009). Similar arrival rate formula has been used in Bassamboo et al. (2010) and Koçağa et al. (2015).

In what follow, we use the superscript λ to denote quantities that scale with λ. To simplify notations, we sometimes suppress the superscript when it is clear from the context.

3. When is Surge Staffing Beneficial?

As mentioned in Section 1, implementing the two-stage staffing requires knowing the realized arrival rate with high precision. In practice, this often involves investing in sophisticated prediction models, which can be costly to develop and maintain. In addition, even though surge staffing is paid at a higher rate, it may not be a desirable working mode for nurses. Therefore, it is important to know how much cost saving we can gain by having the flexibility of surge staffing.

In resonance with the two-stage optimization problem (2), we define the single-stage optimal staffing problem as

$$\min_{\pi \in \Pi} \mathcal{C}_{\pi} = \min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[(h + a\gamma)\, Q(N_1, \Lambda)\right] \right\}. \tag{4}$$

Note that the single-stage problem is equivalent to the two-stage staffing problem (2) by imposing the surge staffing level to be $N_2(N_1, \Lambda) = 0$ for any base staffing level N_1.

For the sequence of systems indexed by λ, we use $\mathcal{C}^\lambda_{1,*}$ to denote the optimal total cost for the single-stage optimization problem (4). Correspondingly, we use $\mathcal{C}^\lambda_{2,*}$ to denote the optimal total cost for the two-stage optimization problem (2).

Theorem 1 (benefit of surge staffing) *Given the order of uncertainty α, the difference in optimal costs for the single-stage versus two-stage optimization problem can be summarized as:*

(I) If $\alpha < 1/2$, then $\mathcal{C}^\lambda_{1,*} - \mathcal{C}^\lambda_{2,*} = o(\sqrt{\lambda})$.

(II) If $\alpha = 1/2$, then $\mathcal{C}^\lambda_{1,*} - \mathcal{C}^\lambda_{2,*} = O(\sqrt{\lambda})$.

(III) If $\alpha > 1/2$, then $\mathcal{C}^\lambda_{1,*} - \mathcal{C}^\lambda_{2,*} = \Theta(\lambda^\alpha)$.

We next provide some intuition behind Theorem 1. We first note that when $\gamma = \mu$, for a given realization of the arrival rate, i.e., $\Lambda = \ell$, the steady-state number of patients in the system follows a Poisson distribution with mean ℓ/μ. Its standard deviation is equal to $\sqrt{\ell/\mu} = O(\sqrt{\lambda})$. This system stochasticity cannot be resolved by the prediction model. On the other hand, the arrival-rate uncertainty characterized by (3) is of order λ^α. This parameter uncertainty can be resolved by the prediction model at the surge stage. When $\alpha < 1/2$, the system stochasticity dominates the parameter uncertainty. The gain by conducting two-stage staffing is restricted to $o(\sqrt{\lambda})$. The cost saving is $O(\sqrt{\lambda})$ if the parameter uncertainty and system stochasticity are of the same order, i.e., $\alpha = 1/2$. When $\alpha > 1/2$, the parameter uncertainty dominates the system stochasticity. This is when we gain the most cost savings from the flexibility offered by surge staffing. In this regime, the larger the order of arrival-rate uncertainty is, the larger magnitude of cost saving we gain from surge staffing. Above all, Theorem 1 suggests the ED manager should only consider surge staffing when the is high arrival-rate uncertainty.

4. Near-Optimal Surge Staffing Policy

As derived in Section 3, when the order of arrival-rate uncertainty is strictly larger than that of system stochasticity, the cost saving of implementing the two-stage staffing optimally is significant, i.e., $\Theta(\lambda^\alpha)$. We thus consider this regime as the most meaningful scenario to execute the two-stage staffing, and assume throughout this section that $\alpha > 1/2$. We next derive solutions to the two-stage staffing problem.

Due to the convoluted system dynamics, solving the two-stage stochastic optimization problem (2) explicitly is prohibitively hard. Part of the difficulty lies in characterizing the expected steady-state queue length which depends intricately on the staffing decisions. The two-stage decisions, before and after the realization of the arrival rate, further exacerbate the complexity of the problem. While the problem can be solved numerically, e.g., via simulation optimization, limited insights about the optimal policy can be generated. Hence, we take the approach of solving more tractable approximations of the two-stage optimization problem (2). These approximations can be viewed as asymptotic limits of (2) under appropriate scalings as the system scale λ grows to infinity. Thus, policies derived based on them work really well for relatively large systems and provide insights into how the optimal policy scales with λ. We also discuss small system adaptions in Section 4.3.

4.1. Stochastic-Fluid Based Solution

Since the parameter uncertainty is of a larger order than system stochasticity, we start by approximating the objective function in (2) via suppressing the system stochasticity and focusing solely on the uncertainty in the arrival rate. This relaxation is known as the *stochastic-fluid approximation* (Harrison and Zeevi 2005, Bassamboo et al. 2010). In particular, conditional on the arrival rate Λ, we approximate the steady-state queue length of the $M/M/n+M$ queue via $(\Lambda - n\mu)/\gamma$, which is the equilibrium queue length of a deterministic fluid model with the same arrival rate, service rate, and abandonment rate.

Before introducing the stochastic-fluid approximation for the two-stage optimization problem (2), we illustrate the idea by reviewing the single-stage newsvendor policy (denoted by $u_{1,NV}$) proposed by Bassamboo et al. (2010). Given the staffing level N_1, the steady-state abandonment rate is approximately $(\Lambda - \mu N_1)$ and the steady state queue length is approximately $(\Lambda - N_1\mu)/\gamma$. Then, the single-stage optimization problem (4) can be approximated by

$$\min_{N_1} \left\{ c_1 N_1 + (h\mu/\gamma + a\mu)\,\mathbb{E}\left[(\Lambda/\mu - N_1)^+\right] \right\}. \tag{5}$$

Note that (5) is a typical newsvendor problem, with unit capacity cost c_1, unit sales price $h\mu/\gamma + a\mu$, random demand Λ/μ, and capacity decision N_1. The optimal solution is given by

$$N_1 = \bar{F}^{-1}_{\Lambda/\mu}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right),$$

where $\bar{F}_{\Lambda/\mu} := 1 - F_{\Lambda/\mu}$ is the complementary cumulative distribution function (ccdf) of Λ/μ, and $\bar{F}^{-1}_{\Lambda/\mu}$ is its inverse. Equivalently, we can write

$$N_1 = \frac{\lambda}{\mu} + \bar{F}^{-1}_X\left(\frac{c_1}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\alpha}, \tag{6}$$

where $\bar{F}_X$ is the ccdf of X. We remark that for all staffing rules discussed in the paper, we do not explicitly restrict N_1 and N_2 to satisfy the integer constraints. Since rounding becomes immaterial as we consider the asymptotic performance of the policy as $\lambda \to \infty$, we assume without loss of generality that each staffing prescription is ceiled up to its nearest integer.

Let $\mathcal{C}^{\lambda}_{1,NV}$ denote the expected total cost defined in (2) under the one-stage newsvendor solution. Recall that $\mathcal{C}^{\lambda}_{1,*}$ is the optimal total cost for the single-stage optimization problem (4). Theorem 1 in Bassamboo et al. (2010) establishes that

$$\mathcal{C}^{\lambda}_{1,NV} - \mathcal{C}^{\lambda}_{1,*} = O(\lambda^{1-\alpha}). \tag{7}$$

Note that when $\alpha > 1/2$, $O(\lambda^{1-\alpha}) = o(\sqrt{\lambda})$. Thus, the single-stage newsvendor solution works remarkably well in the single-stage optimal staffing problem.

We next extend the single-stage newsvendor solution to the two-stage newsvendor solution where surge staffing is allowed after we observed the realized arrival rate. The stochastic-fluid approximation of the two-stage optimization problem (2) takes the form

$$\min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,\Lambda)} \left\{ c_2 N_2(N_1,\Lambda) + (h/\gamma + a)\left(\Lambda - \mu(N_1 + N_2(N_1,\Lambda))\right)^+ \right\} \right] \right\}. \tag{8}$$

Given N_1, Assumption 1 implies that the optimal surge-stage staffing level in (8) is given by

$$N_2(N_1,\Lambda) = (\Lambda/\mu - N_1)^+ .$$

Hence, the optimal base-stage staffing level is the optimal solution to

$$\min_{N_1} \left\{ c_1 N_1 + c_2 \mathbb{E}\left[(\Lambda/\mu - N_1)^+ \right] \right\}. \tag{9}$$

Similar to (5), (9) is a newsvendor problem, with unit capacity cost c_1, unit sales price c_2, random demand Λ/μ, and capacity decision N_1. The optimal solution is given by

$$N_1 = \bar{F}^{-1}_{\Lambda/\mu}(c_1/c_2) = \lambda/\mu + \bar{F}^{-1}_X(c_1/c_2)(\lambda/\mu)^{\alpha} .$$

Let $\beta^* := \bar{F}^{-1}_X(c_1/c_2)$. We propose the following two-stage newsvendor solution denoted by $u_{2,NV}$.

Definition 1 (two-stage newsvendor solution) *For $\alpha \in (1/2, 1)$, the parameters of two-stage newsvendor solution $u_{2,NV}$ are set as follows:*

1. *At the base stage, the base-stage staffing level is*

$$N_1 := \lambda/\mu + \beta^*(\lambda/\mu)^{\alpha} + o((\lambda/\mu)^{\alpha}).$$

2. *At the surge stage, the surge-stage staffing level is*

$$N_2(N_1,\Lambda) := (X - \beta^*)^+ (\lambda/\mu)^{\alpha} + o_{UI}((\lambda/\mu)^{\alpha}).$$

In the two-stage newsvendor solution, the base-stage capacity is equal to the average offered load, λ/μ, together with a hedging that is the same order as the arrival-rate uncertainty. After the arrival rate realizes at the surge stage, the capacity is brought up to the realized offered load if $X > \beta^*$. Note that the surge staffing is $O(\lambda^\alpha)$, which is of a smaller order than the base staffing. Since X is a continuous random variable, by the definition of β^*, the probability of assigning nonzero surge staffing is equal to c_1/c_2. Moreover, it follows from Assumption 1 that $c_1/(h\mu/\gamma + a\mu) < c_1/c_2$. Thus, in comparison to the single-stage newsvendor solution described in (6), the two-stage newsvendor solution prescribes less capacity at the base stage. This is intuitive, because with the flexibility to respond to surges in demand by raising the staffing level at the surge stage, the two-stage newsvendor solution can be less aggressive in assigning base-stage servers.

Let $\mathcal{C}^\lambda_{2,NV}$ denote the expected total cost defined in (2) under the two-stage newsvendor solution. Recall that $\mathcal{C}^\lambda_{2,*}$ is the optimal total cost for the two-stage optimization problem (2).

Theorem 2 (optimality gap of $\boldsymbol{u_{2,NV}}$) *For $\alpha \in (1/2, 1)$, the two-stage newsvendor solution in Definition 1 has $\mathcal{C}^\lambda_{2,NV} - \mathcal{C}^\lambda_{2,*} = o(\lambda^\alpha)$.*

Since $\alpha > 1/2$, Theorem 1 implies that $\mathcal{C}^\lambda_{1,NV} - \mathcal{C}^\lambda_{2,*} = \Theta(\lambda^\alpha)$. This, together with Theorem 2 and the gap in (7), suggests that $\mathcal{C}^\lambda_{1,NV} - \mathcal{C}^\lambda_{2,NV} = \Theta(\lambda^\alpha)$.

4.2. Refinement for The Two-Stage Newsvendor Solution

We have established in Theorem 2 that the two-stage newsvendor solution achieves an optimality gap of $o(\lambda^\alpha)$ compared to the exact two-stage optimum. In this section, we propose a refinement for the two-stage newsvendor solution which further reduces the optimality gap to $o(\sqrt{\lambda})$. The improvement is achieved by characterizing the $o_{UI}(\lambda^\alpha)$ term in the two-stage newsvendor solution more carefully.

To provide intuition for the refinement, we shall ignore the $o(\lambda^\alpha)$ and $o_{UI}(\lambda^\alpha)$ terms for now, i.e., setting them to zero, in the two-stage newsvendor solution. The key observation is that depending on the realized arrival rate, the two-stage newsvendor solution will result in the system being either underloaded (capacity exceeding offered load), or critically loaded (capacity equal to offered load). In particular, for any realized arrival rate $\ell = \lambda + x\lambda^\alpha\mu^{1-\alpha}$, if $x < \beta^*$, then

$$N_1 + N_2(N_1, \ell) - \ell/\mu = (x - \beta^*)\,(\lambda/\mu)^\alpha = \Theta(\lambda^\alpha).$$

In this case, the stochastic fluctuation of the queue process becomes a secondary effect. More specifically, as we will make clear in Appendix C.1 (see (32) in the proof of Lemma 3), the expected steady-state queue length is $o(\sqrt{\lambda})$. In the case where $x \geq \beta^*$, the total staffing level is equal to ℓ/μ, under which the system operates in the QED regime (Mandelbaum and Zeltyn 2009). We can then add a square-root hedging against the stochastic fluctuation of the queue process. In particular, consider

$$N_1 + N_2(N_1, \ell) = \ell/\mu + \eta\sqrt{\ell/\mu} + o(\sqrt{\ell/\mu}), \quad \text{for some } \eta \in \mathbb{R}. \tag{10}$$

Under the capacity prescription in (10), the expected steady-state queue length is $\Theta(\sqrt{\lambda})$. This fact is well-known and will be made rigorous for our system in the proof of Theorem 3 in Appendix **??**. Thus, to "optimize" queue length of this magnitude, we refine the two-stage newsvendor solution by restricting the $o_{UI}(\lambda^\alpha)$ term to $O(\sqrt{\lambda}) + o_{UI}(\sqrt{\lambda})$, so that it serves as effective safety capacity against system stochasticity.

A few more definitions are needed to formally introduce the refined staffing rule. Let ϕ and Φ be the pdf and cdf of the standard normal distribution, respectively. The hazard rate of the standard normal distribution is given by

$$H(t) = \phi(t)/\Phi(-t), \quad t \in \mathbb{R}.$$

Define

$$\eta^* := \operatorname*{arg\,min}_{\eta \in \mathbb{R}} \; c_2\eta + \left(\frac{h\mu}{\gamma} + a\mu\right) \underbrace{\frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right) - \eta\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta)}}}_{(a)}. \tag{11}$$

η^* is the optimal solution of the square-root staffing problem in (Mandelbaum and Zeltyn 2009). In particular, the term (a) on the right-hand side of (11) is the diffusion approximation (and a bona-fide limit in the QED regime) of the expected steady-state queue length of an $M/M/n+M$ queue with service rate μ, abandonment rate γ, staffing cost c_2, abandonment cost a, and staffing level prescribed in (10) (i.e., with square root staffing parameter η).

We are now ready to introduce the following refinement to the two-stage newsvendor solution. Since the system operates in the QED regime when $X \geq \beta^*$, we refer to this policy as the *two-stage QED staffing rule* and denote it by $u_{2,QED}$.

Definition 2 (two-stage QED staffing rule) *For $\alpha \in (1/2, 1)$, the two-stage QED staffing rule prescribes staffing levels as follows:*

1. *At the base stage, the base-stage staffing level is*

$$N_1 := \lambda/\mu + \beta^*(\lambda/\mu)^\alpha + O(\sqrt{\lambda/\mu}).$$

2. *At the surge stage, the surge-stage staffing level is*

$$N_2(N_1,\Lambda) := (\Lambda/\mu + \eta^* \sqrt{\Lambda/\mu} - N_1)^+ + o_{UI}(\sqrt{\lambda/\mu}).$$

In the two-stage QED staffing rule, the base-stage staffing level is of the same form as in the two-stage newsvendor solution. After the arrival rate is realized at the surge stage, we first compute the optimal staffing level in the QED regime, and then bring up the staffing level to meet the target. Let $\mathcal{C}^{\lambda}_{2,QED}$ denote the expected total cost in (2) under the two-stage QED staffing rule. The two-stage QED staffing rule guarantees a smaller optimality gap than the two-stage newsvendor solution as quantified in the following theorem.

Theorem 3 (optimality gap of $\boldsymbol{u_{2,QED}}$) *For $\alpha \in (1/2,1)$, the two-stage QED staffing rule in Definition 2 has $\mathcal{C}^{\lambda}_{2,QED} - \mathcal{C}^{\lambda}_{2,*} = o(\sqrt{\lambda})$.*

4.3. Effective Translation of The Two-Stage QED Staffing Rule to Small Systems

Theorem 3 establishes that any policy that belongs to the family of the two-stage QED staffing rules in Definition 2 achieves an optimality gap of $o(\sqrt{\lambda})$. The specification of the $o(\lambda^{\alpha})$ term in N_1 and the $o_{UI}(\sqrt{\lambda})$ term in $N_2(N_2,\Lambda)$, though asymptotically indistinguishable in the context of Theorem 3, may have non-negligible impact on system performance for a finite system, especially when λ is small. We next numerically investigate system performance under different specifications of the two-stage QED staffing rule.

To this end, we consider staffing prescriptions of the form

$$N_1 = \lambda/\mu + \beta^*(\lambda/\mu)^{\alpha} + k\sqrt{\lambda/\mu} \quad \text{and} \quad N_2(N_1,\Lambda) = (\Lambda/\mu + \eta^*\sqrt{\Lambda/\mu} - N_1)^+, \quad \text{for } k \in \mathbb{R}. \tag{12}$$

We consider systems with small arrival rates, namely, setting $\lambda = 25, 50, 75, 100$, and vary the value of k in (12) from -3 to 3 in increments of 1. In each experiment, we estimate the steady-state cost by averaging over 1000 realizations of the random variable X. For each mean arrival rate λ, we compare the costs under different values of k, and report the percentage gap between each cost and the minimum one (among the seven) in Tables 2 and 3. For example, in Table 2, when $\lambda = 25$, the system with $k = 1$ achieves a cost of 39.48, which is the smallest among the seven systems corresponding to the different values of k. The system with $k = -3$ achieves a cost of 49.75 and thus has a percentage gap of $(49.75 - 39.48)/39.48 = 26.01\%$. In all experiments, the random variable X is assumed to follow a standard normal distribution. The other system parameters and the resulting value of (β^*, η^*) are listed in the caption of the tables.

Table 2 System performance under different specifications of the two-stage QED staffing rule with $\boldsymbol{\beta^* = 0, \eta^* = 0.610}$
$\boldsymbol{(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 2)}$

λ \ k	-3	-2	-1	0	1	2	3
25	26.01%	15.88%	7.40%	2.10%	0.00%	2.03%	7.93%
50	17.70%	10.63%	5.01%	1.49%	0.00%	1.30%	5.24%
75	14.36%	8.33%	4.15%	1.20%	0.00%	0.99%	4.27%
100	11.66%	6.78%	3.11%	0.88%	0.00%	1.05%	3.90%

Table 3 System performance under different specifications of the two-stage QED staffing rule with $\boldsymbol{\beta^* = 1.282, \eta^* = -0.140}$
$\boldsymbol{(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 10)}$

λ \ k	-3	-2	-1	0	1	2	3
25	74.69%	33.10%	10.11%	0.00%	1.39%	9.25%	19.60%
50	45.96%	20.53%	6.52%	0.00%	0.99%	6.59%	14.33%
75	34.61%	16.61%	5.25%	0.00%	0.91%	5.72%	12.58%
100	26.28%	11.50%	3.48%	0.00%	1.65%	6.02%	11.89%

We first observe from the tables that even though all the staffing prescriptions, i.e., k ranging from -3 to 3, are asymptotically optimal, there are substantial differences in the pre-limit performances. In Table 2, $k = 1$ leads to the best performance across all system scales tested. In Table 3, $k = 0$ leads to the best performance. Second, k has a highly nonlinear effect on the cost. Staffing too few servers tends to result in a larger percentage gap than staffing too many servers at the base stage. In particular, in both tables, $k = -3$ leads to the worst performance. In Table 3, when $\lambda = 25$ and $k = -3$, the percentage gap can be as large as 74.69%. Lastly, we note that as the system scale grows, the performance gap among different policies shrinks. For example, in Table 2, when $\lambda = 25$, the percentage gap between $k = -1$ and $k = 1$ is 7.40%. It reduces to 3.11% when $\lambda = 100$. This is is consistent with our optimality gap quantification.

Besides the experiments reported in Tables 2 and 3, we also summarize a few more sets of simulation results with different surge staffing cost in Appendix H.1. Among all the numerical experiments, we find the following specification of the two-stage QED staffing rule to be effective and robust for small-scale systems:

$$N_1 = \lambda/\mu + \beta^*(\lambda/\mu)^\alpha + \eta^* \sqrt{\lambda/\mu}, \quad \text{and} \quad N_2(N_1, \Lambda) = (\Lambda/\mu + \eta^* \sqrt{\Lambda/\mu} - N_1)^+. \tag{13}$$

The capacity prescription in (13) lends itself to an intuitive explanation. At the base stage, the staffing level consists of the offered load, a hedging against arrival-rate uncertainty, and a hedging against system stochasticity catered to the mean arrival rate λ. At the surge stage, the staffing level is raised to reach the optimal value in the QED regime catered to the realized arrival rate.

Remark 1 *The development so far can be easily generalized to include a "commitment" cost for the surge staff to be "on-call." That is, a compensation $c_2^0 \in \mathbb{R}_+$ per nurse per shift is paid at the base stage to staff a total of $N_2^0 \in \mathbb{N}$ nurses in the on-call pool. Then at the surge stage, the ED manager calls N_2 ($N_2 \le N_2^0$) nurses from the on-call pool to serve as surge staff in the upcoming shift. In this setting, N_1 and N_2^0 are decision variables at the base stage, while N_2 is the decision variable at the surge stage. Analogous results to Proposition 1 and Theorems 1–3 can be derived in this setting. In particular, when surge staffing is beneficial, we find that the orders of cost saving and optimality gap stay the same as those in the original model. Since introducing the commitment cost has minimal impact on the main insights and unnecessarily complicates exposition, we do not present its detailed analysis.*

5. Numerical Experiments

In this section, we perform numerical experiments to demonstrate the cost saving of the two-stage QED staffing rule over other benchmark policies, and examine its sensitivity with respect to the level of arrival-rate uncertainty and cost rates. We compare the following three staffing rules:

(I). Our proposed two-stage QED staffing rule $u_{2,QED}$ prescribes staffing levels

$$N_1 = \lambda/\mu + \beta^*(\lambda/\mu)^\alpha + \eta^*\sqrt{\lambda/\mu}, \quad \text{and} \quad N_2(N_1,\Lambda) = (\Lambda/\mu + \eta^*\sqrt{\Lambda/\mu} - N_1)^+,$$

for $\beta^* = \bar{F}_X^{-1}(c_1/c_2)$, and η^* defined in (11).

(II). The single-stage newsvendor solution $u_{1,NV}$ prescribes staffing levels

$$N_1 = \lambda/\mu + \bar{F}_X^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right)(\lambda/\mu)^\alpha, \quad \text{and} \quad N_2(N_1,\Lambda) = 0.$$

This policy accounts for arrival-rate uncertainty, but only has one scheduling opportunity – the base stage.

(III). The conventional single-stage square-root staffing rule, denoted by $u_{1,QED}$, makes a one-time staffing decision at the base stage, assuming a staffing cost of c_1 and a deterministic arrival rate of λ. In particular, the staffing levels are given by

$$N_1 = \lambda/\mu + \eta^*_{1,QED}\sqrt{\lambda/\mu}, \quad \text{and} \quad N_2(N_1,\Lambda) = 0,$$

where $\eta^*_{1,QED}$ is defined as

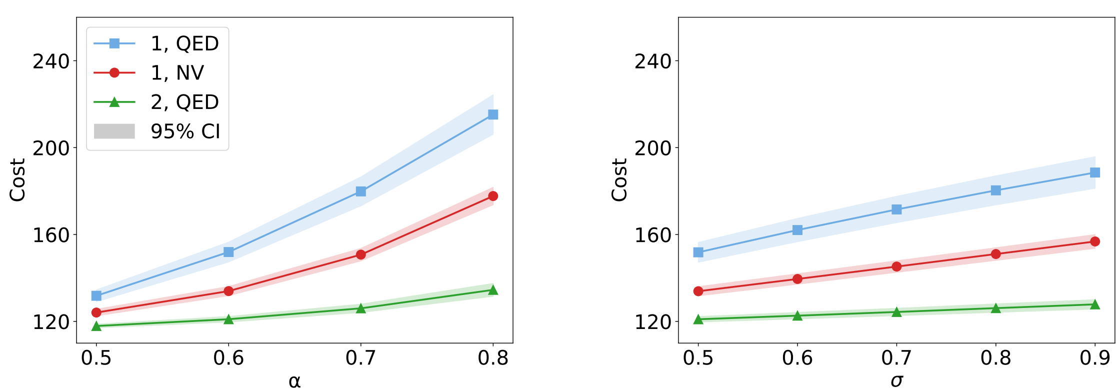
$$\eta^*_{1,QED} := \arg\min_{\eta\in\mathbb{R}} \; c_1\eta + \left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right) - \eta\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta)}}. \tag{14}$$

This policy ignores arrival-rate uncertainty. It is important to distinguish $\eta^*_{1,QED}$ in (14) (used in the single-stage square-root staffing rule) from η^* in (11) (used in the two-stage QED staffing rule). While both serve as coefficients in front of the hedging against system stochasticity, $\eta^*_{1,QED}$ is calculated assuming a staffing cost of c_1 (base-stage cost) and η^* is calculated assuming the a staffing cost of c_2 (surge-stage cost).

5.1. Level of Arrival-Rate Uncertainty

In the first set of experiments, we examine the cost saving of the proposed two-stage QED rule as we vary the magnitude of arrival-rate uncertainty. In particular, we assume that the random variable X is normally distributed with mean 0 and standard deviation σ. We simulate 1000 realizations of X and calculate the expected steady-state cost (where the expectation is taken over the stochastic fluctuations) for each realization. The expected total cost (where the expectation is taken over the random variable X) is then averaged over the expected steady-state costs for all realizations of X. To assess the performance of the two-stage staffing rule with respect to the arrival-rate uncertainty, we vary the order of arrival-rate uncertainty, α, and the standard deviation of X, σ, respectively, with everything else held constant. Figure 1 illustrates the expected total costs under the three policies, with α increasing from 0.5 to 0.8 in Figure 1(a) and σ increasing from 0.5 to 0.9 in Figure 1(b). We observe that among the first three polices, $u_{1,QED}$ performs the worst, while our proposed $u_{2,QED}$ performs the best. More importantly, the cost saving of $u_{2,QED}$ relative to $u_{1,NV}$ or $u_{1,QED}$ grows with the level of arrival-rate uncertainty.

Figure 1 Sensitivity analysis with respect to the order of arrival-rate uncertainty
((a): $\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \sigma = 1$
(b) : $\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \alpha = 0.75$)

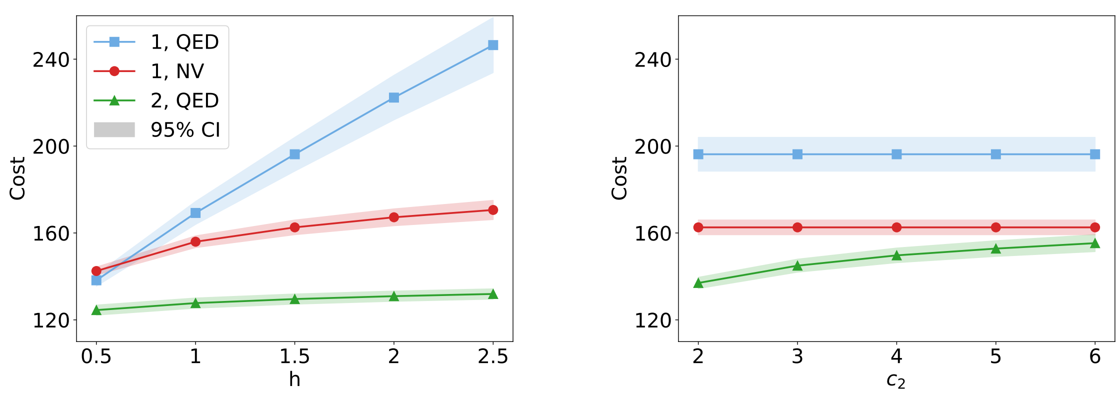


(a) Sensitivity with resect to α (b) Sensitivity with resect to σ

5.2. Cost Rates

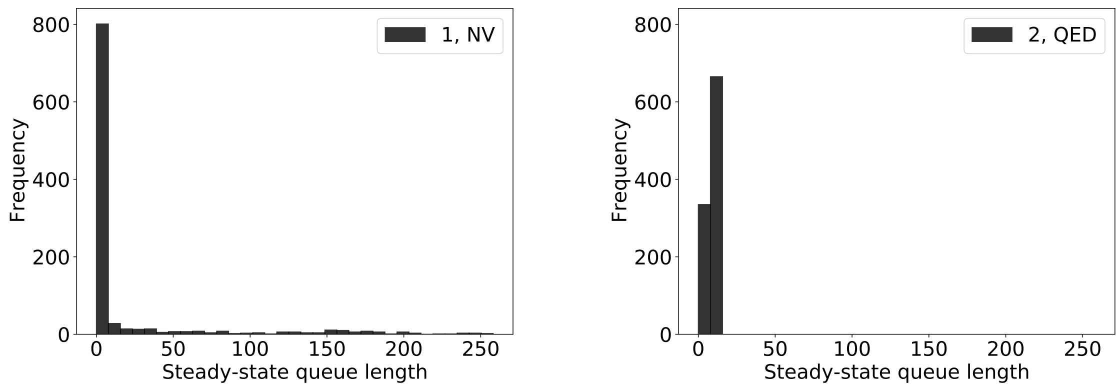
We next investigate the performance of our proposed two-stage policy with respect to the cost parameters. We first compare the costs of the three policies under different holding costs, h, in Figure 2(a). Note $u_{2,QED}$ outperforms $u_{1,QED}$ and $u_{1,NV}$ by a larger magnitude as the holding cost becomes larger. This trend is not surprising, because by making staffing decisions at both the base and surge stages, the two-stage QED staffing rule is able to circumvent understaffing when the realized arrival rate is excessively large. In contrast, due to the inability to adjust the staffing level at the surge stage, the benchmark single-stage policies can result in relatively large queue when the realized arrival rate is large. Figure 3 demonstrates the distribution of the average steady-state queue length for a given value of X over 1000 realizations of X under $u_{1,NV}$ and $u_{2,QED}$. We observe that while the average steady-state queue length for a specific realization of X can be as high as 250 under the single-stage newsvendor solution, it remains below 20 for all realizations of X under the two-stage QED staffing rule. Besides the holding cost, we also vary the surge-stage staffing cost, c_2. Recall from Assumption 1 that the surge staffing cost is larger than the base staffing cost c_1, but smaller than the performance cost $h\mu/\gamma + a\mu$. In the numerical experiment depicted in Figure 2(b), we set $c_1 = 1, h\mu/\gamma + a\mu = 18$, and vary c_2 from 2 to 6. We see that the cost saving of the proposed two-stage policy $u_{2,QED}$ decreases as c_2 increases. In particular, the performance of $u_{2,QED}$ becomes nearly indistinguishable from that of $u_{1,NV}$ when c_2 reaches 6.

Figure 2 **Sensitivity analysis with respect to the cost rates**
((a): $\boldsymbol{\lambda = 100, \mu = 1, \gamma = 0.1, a = 2h, c_1 = 1, c_2 = 1.5, \alpha = 0.75, \sigma = 1}$
(b): $\boldsymbol{\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, \alpha = 0.75, \sigma = 1}$)



(a) Sensitivity with resect to h (b) Sensitivity with resect to c_2

Figure 3 Distribution of the average steady-state queue length
$(\boldsymbol{\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \alpha = 0.75, \sigma = 1})$



(a) Single-stage newsvendor solution (mean = 19.131, std = 49.070)

(b) Two-stage QED staffing rule (mean = 8.029, std = 5.305)

6. Model Extension: Incorporation of Surge-Stage Prediction Error

In the two-stage optimization problem (2), we assume that the realization of the random arrival rate Λ is known exactly at the surge stage. That is, the surge-stage prediction model provides perfect arrival rate information. However, in practice, the surge-stage predictive models may incur some prediction errors. In this section, we investigate a model extension where we allow prediction error in the surge stage.

To incorporate prediction error, we further decompose the random arrival rate into two terms: predictable and unpredictable terms. In particular, we consider random arrival rate of the form

$$\Lambda = \lambda + Y\lambda^{\alpha}\mu^{1-\alpha} + Z\lambda^{\nu}\mu^{1-\nu}, \tag{15}$$

where $\alpha \in (1/2, 1)$, $\nu \in (0, \alpha]$, and Y and Z are continuous random variables independent of each other. We assume that $\mathbb{E}[Y] = \mathbb{E}[Z] = 0$, $\mathbb{E}[|Y|] < \infty$, and $\mathbb{E}[|Z|] < \infty$. In (15), Y and Z can be understood as the *predictable* and *unpredictable* arrival-rate uncertainty, respectively. If there is a prediction model to forecast demand at the surge stage, then Y is the predicted arrival rate and Z is the error (residual) of the prediction model. Then, α captures the scale of the arrival-rate uncertainty and ν captures the scale of the prediction error. It is reasonable to assume that the distributions of Y and Z are known at the base stage. The two-stage staffing problem with prediction error is then formulated as

$$\min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,Y)} \{ c_2 N_2(N_1,Y) + (h + a\gamma)\,\mathbb{E}[Q(N_1 + N_2(N_1,Y), \Lambda)|Y] \} \right] \right\}. \tag{16}$$

To differentiate notation from that of problem (2), we denote the optimal objective value of (16) as $\mathcal{C}_{2,*}^{e,\lambda}$ when there is prediction $\underline{e}$rror at the surge stage.

Similar to problem (2), we again compare to the single-stage optimization problem (4) for Λ in form of (15), and use $\mathcal{C}_{1,*}^{e,\lambda}$ to denote its optimal objective value. To draw connection between the arrival rates in (3) and (15), we can let X be such that

$$X\lambda^{\alpha}\mu^{1-\alpha} = Y\lambda^{\alpha}\mu^{1-\alpha} + Z\lambda^{\nu}\mu^{1-\nu}. \tag{17}$$

In this context, problem (2) can be seen as an "oracle" problem that knows the exact realized arrival rate at the surge stage. We use $\mathcal{C}_{2,*}^{o,\lambda}$ to denote the optimal objective value of the oracle problem (2) for Λ in form of (15). In particular, the oracle problem does not incur any unpredictable arrival-rate uncertainty (prediction error). Intuitively, the impact of the prediction error should depend on how substantial it is. We formalize this for "small" and "moderate/large" prediction errors in the next subsections. The error regime depends on the relationship between the scale of the arrival-rate uncertainty and that of the prediction error.

6.1. Small Prediction Error: $0 < \nu < 1/2$

When $\nu \in (0, 1/2)$, the prediction error is sufficiently small to be "ignored." Doing so does not impact performance. For problem (16), we propose the *two-stage error policy* and denote it by $u_{2,ERR}$.

Definition 3 (two-stage error policy for $\nu < 1/2$) *For $\alpha \in (1/2, 1)$ and $\nu \in (0, 1/2)$, the two-stage error policy prescribes staffing levels as follows:*

1. *At the base stage, the base-stage staffing level is*

$$N_1 := \lambda/\mu + \bar{F}_Y^{-1}(c_1/c_2)(\lambda/\mu)^{\alpha} + O(\sqrt{\lambda/\mu}).$$

2. *At the surge stage, the surge-stage staffing level is*

$$N_2(N_1, Y) := ((\lambda + Y\lambda^{\alpha}\mu^{1-\alpha})/\mu + \eta^* \sqrt{(\lambda + Y\lambda^{\alpha}\mu^{1-\alpha})/\mu} - N_1)^+ + o_{UI}(\sqrt{\lambda/\mu}),$$

   for η^* defined in (11).

When $\nu \in (0, 1/2)$, $u_{2,ERR}$ is similar to $u_{2,QED}$, the latter of which is defined for the case without prediction error. In particular, $u_{2,ERR}$ completely ignores the existence of prediction error Z and makes staffing decisions based on Y only. Let $\mathcal{C}_{2,ERR}^{e,\lambda}$ denote the expected total cost under $u_{2,ERR}$ when the mean arrival rate is λ. Analogous results to Theorems 1 and 3 hold for $u_{2,ERR}$ when prediction error is small, namely, $\nu \in (0, 1/2)$.

Proposition 2 *For $\alpha \in (1/2, 1)$ and $\nu \in (0, 1/2)$, we have*

(I) Cost saving: $\mathcal{C}_{1,*}^{e,\lambda} - \mathcal{C}_{2,*}^{e,\lambda} = \Theta(\lambda^\alpha)$.

(II) Optimality gap: $\mathcal{C}_{2,ERR}^{e,\lambda} - \mathcal{C}_{2,*}^{e,\lambda} = o(\sqrt{\lambda})$.

(III) Cost of prediction error: $\mathcal{C}_{2,*}^{e,\lambda} - \mathcal{C}_{2,*}^{o,\lambda} = o(\sqrt{\lambda})$.

Item (III) in Proposition 2 quantifies the gap between two-stage optimal cost with prediction error and the two-stage optimal cost without prediction error. We observe that when the prediction error is small, i.e., $\nu < 1/2$, there is not much value, from the cost-saving perspective, to further improve the prediction accuracy.

6.2. Moderate to Large Prediction Error: $1/2 \le \nu \le \alpha$

When $\nu \in [1/2, \alpha]$, the prediction error is of a larger order than the system stochasticity and thus can no longer be ignored for staffing. To derive a near-optimal solution to problem (16), we consider the following stochastic-fluid optimization problem

$$\min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,Y)} \left\{ c_2 N_2(N_1,Y) + (h\mu/\gamma + a\mu)\mathbb{E}\left[(\Lambda/\mu - N_1 - N_2(N_1,Y))^+ | Y \right] \right\} \right] \right\}. \quad (18)$$

Let $(\bar{N}_1, \bar{N}_2(\bar{N}_1, Y))$ denote an optimal solution to (18), whose existence is rigorously established in the proof of Proposition 3. When $\nu \in [1/2, \alpha]$, we define the two-stage error policy, $u_{2,ERR}$, to prescribe staffing levels $(\bar{N}_1, \bar{N}_2(\bar{N}_1, Y))$.

When $1/2 < \nu < \alpha$ (moderate prediction error), the prediction error is of a smaller order than the predictable arrival-rate uncertainty. In this case, we still expect that resolving some of the arrival-rate uncertainty at the surge stage can bring a cost saving as large as $O(\lambda^\alpha)$ compared to the optimal single-stage staffing rule. When $\nu = \alpha$ (large prediction error), the prediction error is of the same order as the predictable arrival-rate uncertainty. In this case, the prediction error can be so large that when comparing to the optimal single-stage staffing rule, the cost saving is no longer $O(\lambda^\alpha)$. Thus, we further divide the analysis of this case into two sub-cases as quantified by the following assumption.

Assumption 2 *There exists $p \in (0,1]$ such that*

$$Y + \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) - \bar{F}_{Y+Z}^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right) > 0 \quad \textit{with probability } p.$$

Assumption 2 provides a relationship where the predictable arrival-rate uncertainty is sufficiently large compared to the unpredictable arrival-rate uncertainty. Note that when Y has a bounded support, Assumption 2 may not hold if c_2 is large or Z has a large standard deviation. For a concrete example that violates Assumption 2, consider $Y \sim \text{Uniform}[-1,1]$, $Z \sim \text{Normal}(0, 10^2)$, $h\mu/\gamma + a\mu = 1$, $c_1 = 0.1$, and $c_2 = 0.9$. In this case, $Y + \bar{F}_Z^{-1}(c_2/(h\mu/\gamma + a\mu)) < 0$, $\bar{F}_{Y+Z}^{-1}(c_1/(h\mu/\gamma + a\mu)) > 0$, and Assumption 2 does not hold for all realizations of Y. The intuition is that the predictable arrival-rate uncertainty (Y) is so small compared to the unpredictable arrival-rate uncertainty (Z) that resolving Y only leads to limited cost saving, which is on the order of $o(\lambda^\alpha)$.

Proposition 3 *For $\alpha \in (1/2, 1)$ and $\nu \in [1/2, \alpha]$, we have*

(I) Cost saving: *If $\nu < \alpha$, then $\mathcal{C}^{e,\lambda}_{1,*} - \mathcal{C}^{e,\lambda}_{2,*} = \Theta(\lambda^\alpha)$. If $\nu = \alpha$ and Assumption 2 holds, then $\mathcal{C}^{e,\lambda}_{1,*} - \mathcal{C}^{e,\lambda}_{2,*} = \Theta(\lambda^\alpha)$. If $\nu = \alpha$ and Assumption 2 does not hold, then $\mathcal{C}^{e,\lambda}_{1,*} - \mathcal{C}^{e,\lambda}_{2,*} = o(\lambda^\alpha)$.*

(II) Optimality gap: *$\mathcal{C}^{e,\lambda}_{2,ERR} - \mathcal{C}^{e,\lambda}_{2,*} = O(\sqrt{\lambda})$.*

(III) Cost of prediction error: *$\mathcal{C}^{e,\lambda}_{2,*} - \mathcal{C}^{o,\lambda}_{2,*} = \Theta(\lambda^\nu)$.*

Comparing item (III) in Proposition 3 to item (III) in Proposition 2, we note that when having a large prediction error, there is potentially more cost saving we can gain by improving the prediction accuracy. In particular, when $\nu \geq 1/2$, the cost saving brought by a more accurate prediction model can be as large as $\Theta(\lambda^\nu)$.

6.3. Numerical Experiments for Models with Prediction Error

We conduct numerical experiments in the presence of prediction error, and focus on the case where the magnitude of prediction error is the most salient, namely, $\nu = \alpha$.

We compare the following five staffing rules:

(I) The two-stage error policy $u_{2,ERR}$ introduced in Section 6.2. It has near-optimal performance as established in Proposition 3.

(II) The two-stage QED rule $u_{2,QED}$, which is a straightfoward extension of the two-stage QED rule defined in Definition 2 by ignoring the prediction error: For X defined in (17) (namely, $X := Y + Z$), it assigns

$$N_1 = \lambda/\mu + \bar{F}_X^{-1}(c_1/c_2)(\lambda/\mu)^\alpha + \eta^* \sqrt{\lambda/\mu}$$
$$N_2(N_1, Y) = ((\lambda + Y\lambda^\alpha \mu^{1-\alpha})/\mu + \eta^* \sqrt{(\lambda + Y\lambda^\alpha\mu^{1-\alpha})/\mu} - N_1)^+.$$

The staffing prescription takes into account the distribution of X at the base stage, but uses the realization of Y as a proxy for the realization of X at the surge stage. To simplify notation, we still refer to this policy as $u_{2,QED}$ in the following experiments.

(III) The single-stage newsvendor solution $u_{1,NV}$ as defined in Section 5, assuming we know the distribution of X. Note that for a fixed distribution of X, the single-stage staffing rule and its performance will not be affected by the surge-stage prediction errors.

(IV) The single-stage square-root staffing rule $u_{1,QED}$ as defined in Section 5.

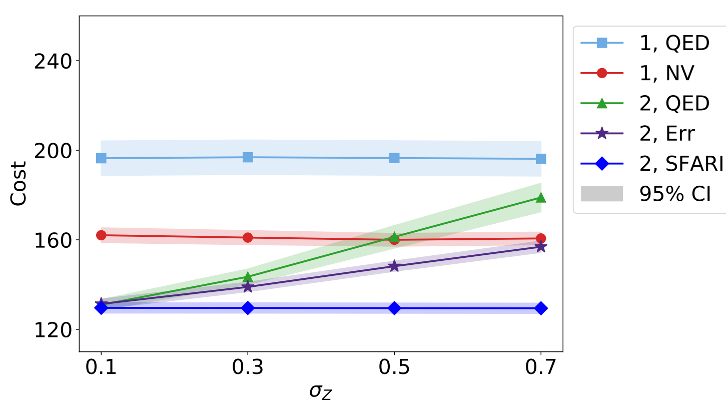
(V) To demonstrate the cost of prediction error, we also consider a policy termed *second-stage full arrival rate information* (SFARI) for the oracle problem, and denote it by $u_{2,SFARI}$. It prescribes staffing levels

$$N_1 = \lambda/\mu + \bar{F}_X^{-1}(c_1/c_2)(\lambda/\mu)^\alpha + \eta^*\sqrt{\lambda/\mu} \quad \text{and} \quad N_2(N_1, \Lambda) = (\Lambda/\mu + \eta^*\sqrt{\Lambda/\mu} - N_1)^+,$$

for η^* defined in (11). Note that $u_{2,SFARI}$ is identical to $u_{2,QED}$ when there is full arrival rate information at the surge stage. It provides a lower bound to the performance under the other policies.

We assume that Y and Z are normally distributed with standard deviation σ_Y and σ_Z, respectively. We then fix the standard deviation of X to be equal to 1, i.e., $\sigma_Y^2 + \sigma_Z^2 = 1$, and vary the σ_Z from 0.1 to 0.7 in increment of 0.2. For each policy and each value of σ_Z, we simulate 1000 independent and identically distributed realizations of the random arrival rate, and use the average to approximate the expected total cost. Figure 4 compare the costs under the six policies with different values of σ_Z. Note that, as expected, the single-stage benchmark policies ($u_{1,NV}$ and $u_{1,QED}$) and the oracle policy ($u_{2,SFARI}$) are unaffected by prediction accuracy. In contrast, the performance of our proposed two-stage policies ($u_{2,ERR}$ and $u_{2,QED}$) degrades as the prediction error increases. When σ_Z is larger than or equal to 0.5, $u_{2,QED}$ yields higher expected total cost than $u_{1,NV}$. On the other hand, $u_{2,ERR}$, which accounts for prediction error in its staffing approach, outperforms the benchmark single-stage policies for all σ_Z. As σ_Z increases from 0.1 to 0.7, the expected total cost under $u_{2,ERR}$ increases from 131.356 to 156.897. This further demonstrates the cost saving we can gain by improving the prediction accuracy. In practice, this can potentially be achieved by employing more sophisticated machine learning models or including more relevant real-time features.

Figure 4 **Sensitivity analysis with respect to prediction error**
$(\boldsymbol{\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \alpha = \nu = 0.75})$



7. Application to the Emergency Department

In this section, we develop a unified framework to guide implementation of the proposed two-stage staffing policy in the actual ED. Our framework consists of three key elements:

1) Estimating the arrival rate distribution, especially the order of arrival-rate uncertainty. This helps us decide whether the ED operates in an environment where surge staffing could be beneficial. In our partner hospital, α is estimated to be 0.769. According to Theorem 1, we can gain substantial cost saving by utilizing the surge staffing in this regime.

2) Building an integrated two-stage prediction model that is synchronized with the staffing decision epochs. At the base stage we can only capture the day-of-the-week and day-versus-night effects, while at the surge stage, we can utilize more real-time information such as the severity profile of patients currently in the ED, the weather condition, etc.

3) Implementing a prediction-driven surge staffing rule. For our partner hospital, we incur significant prediction error at the surge stage. Thus, we employ $u_{2,ERR}$. We also modify $u_{2,ERR}$ to adjust for the transient-shift effect.

7.1. Background and Data

Our partner hospital, NYP CUMC, is an urban academic medical center in New York City. We focus on the Milstein ED at NYP CUMC, which is the main adult ED of the hospital and treats more than 90,000 patients annually. Nurses are scheduled to 12-hour shifts that begin at 7am (day shift) or 7pm (night shift) each day. According to ED management, nursing schedules are typically set 4–8 weeks in advance and the staffing level is difficult to change in real time. If the ED manager anticipates a high patient volume two-to-three hours before the start of a shift, he/she can call in extra nurses to work overtime. Currently, there is a lack of a data-driven approach to determine the appropriate surge staffing levels.

We collect 12 months of data from February 1, 2018 to January 31, 2019. The data contain patient-level records that include 1) patient-flow time stamps (i.e., time stamps for arrival, first evaluation, lab and imaging orders, admission decision, and departure), 2) patient's demographics and severity (i.e., age, gender, arriving source, emergency severity index, chief complaint, comorbidities, and deposition decision), and 3) patient's lab and imaging requests (i.e., different tests and imaging that are ordered for the patient). We also collect data from two other sources: the weather information (i.e., temperature, precipitation, snow, wind, etc.) and Google trend data (i.e., search volume for key words such as "flu," "heart attack," "abuse," etc.). These data allow us to a) estimate arrival-rate uncertainty, b) build a two-stage prediction model where the surge-stage prediction model can utilize rich real-time information, and c) calibrate a high-fidelity simulation model to evaluate different staffing policies.

We first group the shifts into 14 different types based on day of the week and day versus night shift. Table 4 provides some summary statistics for different shifts. We observe that the day shifts see more arrivals than the night shifts, and weekday day shifts see more arrivals than the weekend day shifts. We also note that the coefficient of variation can be as high as 14% for some shifts (e.g., Sunday night shift and Thursday night shift). This suggests that even after we control for day-of-the-week and day-versus-night effects, there can still be quite some uncertainty in demand.

Table 4 **Mean and standard deviation (std) of the shift-level arrival counts**

Day shift							
	Sun	Mon	Tue	Wed	Thur	Fri	Sat
Mean	141.019	207.385	188.769	186.942	185.208	175.173	147.058
Std	15.788	21.503	20.701	23.657	21.004	16.124	12.095

Night shift							
	Sun	Mon	Tue	Wed	Thur	Fri	Sat
Mean	89.462	97.058	97.769	93.711	95.189	96.692	94.115
Std	12.698	12.064	10.547	12.508	13.602	12.199	11.514

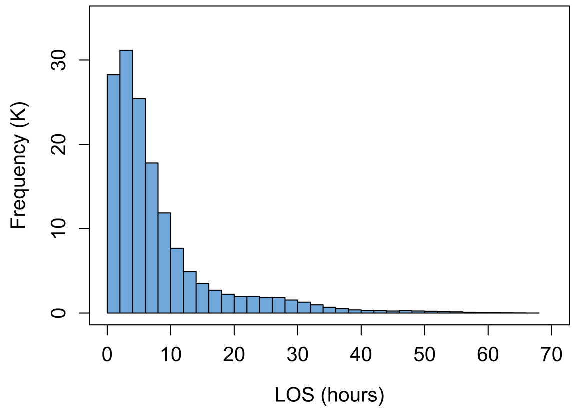
The length of stay (LOS) for each patient is defined as the time difference between the first evaluation time and the departure time. The average LOS in our ED is 8.156 hours due to a long boarding time for patients who need to be admitted into the hospital. As illustrated in Figure 5, patients' LOS is best described by a lognormal distribution whose logarithm has mean equal to 1.597 and standard deviation equal to 1.050. The average waiting time is close to an hour, i.e., 0.975 hours, and the proportion of patients who left without being seen is 4.35%. Properly managing congestion is a key challenge faced by the ED. In what follows, we look into how our data-driven surge planning can help reduce congestion and staffing costs.

7.2. Estimating Arrival-Rate Uncertainty

In this section, we introduce statistical procedures to estimate the arrival-rate uncertainty. Because there are significant day-of-the-week and day-versus-night effects, the shifts are classified into 14 different types as demonstrated in Table 4. Let λ_i denote the mean arrival rate for type $i \in \mathcal{I} := \{1, ..., 14\}$ shift. Since we have one year of data, each shift type i has $n_i = 52$ observations. For each type of shift, we assume that the random arrival rate takes the form

$$\Lambda_i = \lambda_i + \lambda_i^{\alpha} \mu^{1-\alpha} X, \quad i \in \mathcal{I},$$

Figure 5 **Patient LOS distribution**



for μ equal to the inverse of the average LOS. In particular, the uncertainty scaling parameter α and the distribution of X is the same across different types of shifts. We also make the parametric assumption that $X \sim N(0, \sigma^2)$ for some $\sigma \in \mathbb{R}_+$. Then $\Lambda_i \sim N(\lambda_i, \lambda_i^{2\alpha}\mu^{2(1-\alpha)}\sigma^2)$, $i \in \mathcal{I}$.

Let $L_i^{(k)}$ denote the observed arrival count for the kth shift of type i, $1 \le k \le n_i$. We also define $\bar{L}_i := \frac{1}{n_i}\sum_{k=1}^{n_i} L_i^{(k)}$ and $\Sigma_i^2 := \frac{1}{n_i}\sum_{k=1}^{n_i}(L_i^{(k)} - \bar{L}_i)^2$, i.e., the corresponding sample mean and sample variance. Based on the method of moments, we have the following system of equations for the estimators

$$\bar{L}_i = \hat{\lambda}_i, \quad \Sigma_i^2 = \hat{\lambda}_i^{2\hat{\alpha}}\mu^{2(1-\hat{\alpha})}\hat{\sigma}^2, \quad i \in \mathcal{I}. \tag{19}$$

It follows from (19) that

$$\log \Sigma_i = \hat{\alpha} \log \bar{L}_i + \log(\mu^{1-\hat{\alpha}}\hat{\sigma}), \quad i \in \mathcal{I}. \tag{20}$$

Then, we can fit $\hat{\alpha}$ and $\hat{\sigma}$ by solving the following least squares problem

$$\min_{\alpha \in (0,1),\, \gamma \in \mathbb{R}} \sum_{i=1}^{14} \left(\log \Sigma_i - \gamma - \alpha \log \bar{L}_i\right)^2. \tag{21}$$

In particular, let γ^* and α^* denote the optimal solution to the least squares problem (21). Then, $\hat{\alpha} = \alpha^*$ and $\mu^{1-\hat{\alpha}}\hat{\sigma} = \exp(\gamma^*)$.

In Table 5, the first row below header (with $|\mathcal{I}| = 14$) summarizes the point estimates for α and $\mu^{1-\alpha}\sigma$. We also report their corresponding 95% confidence intervals. Based on our estimation, for the Milstein ED, $\alpha = 0.769$ and $\mu^{1-\alpha}X \sim N(0, 0.348^2)$.

To check the robustness of our estimation, we also run a similar analysis by dividing the shifts into 56 different types. In particular, in addition to the day-of-the-week and day-versus-night effects, we also incorporate the season-of-the-year effect. The second row below header (with $|\mathcal{I}| = 56$) summarizes estimation results, which are very close to our original estimation. Lastly, we also consider a non-parametric estimation proposed in Maman (2009), which works for $\alpha > 1/2$ only (see Appendix G). It gives the same results as our original estimation. Since it is a priori unclear for a real-world system whether $\alpha > 1/2$, our parametric estimation method, which allows $\alpha \in (0,1)$, is preferred.

Table 5 Estimated α and standard deviation of X

Number of shift types	$\hat{\alpha}$	95% CI for $\hat{\alpha}$	$\mu^{1-\alpha}\hat{\sigma}$	95% CI for $\mu^{1-\alpha}\hat{\sigma}$
Day-of-week and day/night: $\lvert\mathcal{I}\rvert = 14$	0.769	(0.543, 0.994)	0.344	(0.114, 1.034)
Day-of-week, day/night and seasons: $\lvert\mathcal{I}\rvert = 56$	0.746	(0.558, 0.933)	0.362	(0.135, 0.902)

7.3. Two-Stage Prediction Model

To facilitate base and surge staffing decisions, we need to develop a two-stage prediction model that is synchronized with these decision epochs.

At the base stage, which is several weeks before the start of the shift, there is very limited information we can utilize for demand forecasting. The key features based on our analysis are the day-of-the-week effect and the day-versus-night effect. In particular, the stratified historical averages based on these features are able to capture 88.26% of the variability in shift-level arrival counts.

At the surge stage, which is a few hours before the start of the shift, we have access to more real-time information. We employ a recently developed linear regression model in Hu et al. (2021) to forecast the realized arrival rate. The model utilizes the following five categories of features: (i) *Time-of-the-shift information*, including season of the year, day of the week, day versus night, and whether the shift takes place on, before, or after a national holiday; (ii) *Previous-shift arrival counts*, including the shift-level arrival count 1 day before the shift, the shift-level arrival count 7 days before the shift (a week ago), and a moving average of shift-level arrival counts over the past 30 days; (iii) *Patient severity level*, which is the average of the weighted sum of a total of 17 Charlson comorbidity indices in ICD-10-CM coding for each patient over the past 3 days; (iv) *Google trends*, including the Google search volume for the key words “depression” and “flu” in New York State for the week before the shift; (v) *Weather forecast*, including the minimum temperature, precipitation, snow, wind, and whether the maximum temperature exceeds 86^{o}F on the day of the shift. The estimated coefficients for the covariates in the model are provided in Table 9 in Appendix H.2. This linear regression model is able to capture 90.84% of the variability in shift-level arrival counts.

The root mean-square error (RMSE) of the prediction model is 15.860 at the base stage, and 14.009 at the surge stage. From the prediction accuracy perspective, the real time information is able to reduce the RMSE by 11.67%. That said, what we are more interested in is the value of this gained accuracy in making staffing decisions. We shall investigate this in the next subsection.

The residuals of the surge-stage regression model have mean equal to 0.000 and standard deviation equal to 14.009. Since the standard error of the surge stage prediction is quite high, we use the random arrival rate model with prediction error, i.e., (15), and estimate the distribution of Y and Z next. We assume that $\alpha = \nu$, and Y and Z are both normally distributed. Recall from the random arrival rate model that residuals for shifts of type i at the surge stage are distributed according to $\lambda_i^\alpha \mu^{1-\alpha} Z$, $i \in \mathcal{I}$. Using the point estimates for λ_i and α in Section 7.2, we estimate the standard deviation of Z to be 0.302. Because $X\lambda^\alpha\mu^{1-\alpha} = Y\lambda^\alpha\mu^{1-\alpha} + Z\lambda^\alpha\mu^{1-\alpha}$, based on the estimation for the standard deviation of X in Section 7.2, the standard deviation of Y is 0.111.

7.4. ED-Adapted Two-Stage Staffing Rule

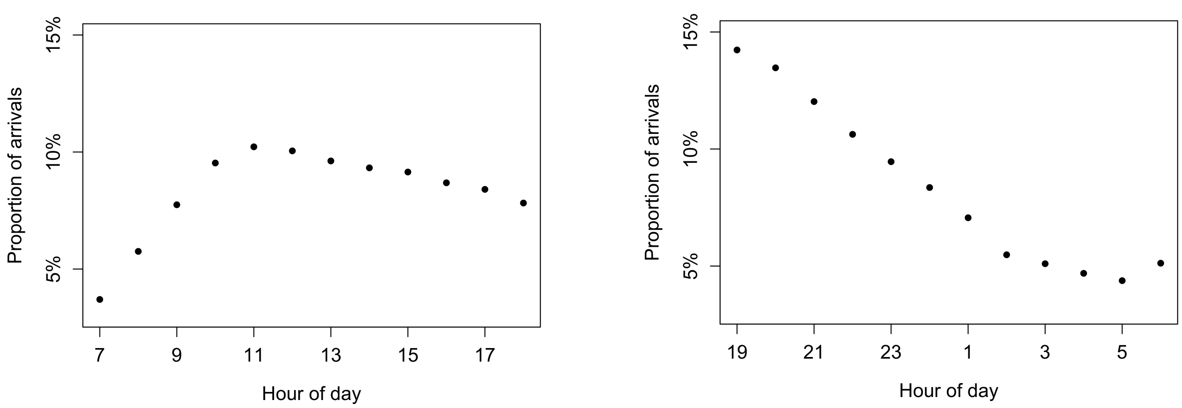
To examine the performance of the proposed two-stage staffing rule, we build a high-fidelity queueing model to the simulate patient flow process in Milstein ED over 52 weeks from February 1, 2018 to January 31, 2019.

7.4.1. Model Calibration The hospital system is modeled as an $M_t/G/N_t + M$ queue, a multi-server queue with time-varying arrival rate at the hourly level, log-normal service time, and time-varying staffing at the shift level, where the servers are the nurses. For shift of type i in the kth week, we assume that the realized arrival rate for that shift is equal to the observed arrival count in data, $L_i^{(k)}$, $1 \leq i \leq 14$, $1 \leq k \leq 52$. The hourly arrival rate for each of the 12 hours in a shift is obtained by scaling $L_i^{(k)}$ proportionally to match the empirical hourly proportion of arrivals as illustrated in Figure 6. In what follows, we shall refer to the $L_i^{(k)}$'s as the realized arrival rates. The LOS for each patients follows a lognormal distribution whose logarithm has mean equal to 1.597 and standard deviation equal to 1.050. Note that this is the best fitted distribution from data. While waiting in queue, patients leave the system without been seen after an exponentially distributed amount of patience time with mean equal to 36 hours. Patients are served in a FCFS manner and once a patient begins service, he/she will not abandon the system. Note that in practice while patients within a severity class (e.g., within the same ESI) are often served FCFS, this is not necessarily the case across different classes. As we are interested in assessing impact of the new staffing approach on system performance (e.g. average waiting time across all patients), rather than on individual patient, FCFS a reasonable simplification. Furthermore, we consider a nurse-to-patient ratio of 1-to-3, which is the number of patients that an ED nurse can treat simultaneously. We scale down the staffing levels suggested by the staffing policies by the nurse-to-patient ratio to get the actual number of nurses needed for the shift.

At the end of each shift, patients who have not finished service are queued up in a FCFS manner (according to their arrival times) for the nurses who are staffed for the upcoming shift to continue treatment. These patients have priority over those who have not started treatment, and do not abandon the system while waiting to resume service. The waiting time for each patient includes the time the he/she waits to be first evaluated by a nurse upon arrival, as well as the period during which his/her treatment is in suspension due to the change of shifts. We remark that while there are different ways to handle patient hand-off at shift transitions (such as having nurses work overtime, or allowing multi-tasking), our assumption on having the patients wait to resume service has minimal—practically insignificant—impact on the performance measures. For the experiments shown in Figure 7 below, when the average waiting time is 32.608 minutes under our proposed policy (alternatively, 34.990 minutes under the single-stage benchmark), the part patients spend waiting to resume service only accounts for 0.113 minutes (alternatively, 6.164 minutes under the single-stage benchmark).

In terms of the costs, we assume that the salary is \$45 per hour for nurses who are staffed at the base stage, and \$67.5 (\$45 $\times$ 1.5) per hour for nurses who are staffed at the surge stage (Weiss et al. 2011). We shall vary the holding and abandonment costs in our numerical experiments to evaluate the performance of different staffing policies.

Figure 6 **Proportion of patient arrivals by hour within day/night shift**



(a) Day shift (b) Night shift

7.4.2. Adjustment to The Staffing Rules The queueing dynamics during each shift in the ED can be quite different from the stylized model considered in Section 2. In particular, based on our model calibration in Section 7.4.1, i) the arrival rate is time-varying, ii) the service-time distribution is lognormal, and iii) each shift is only 12 hours, which may not be long enough for the system to reach stationarity. We single out these deviations and run extensive simulation experiments to check the robustness of the performance of our two-stage error policy. It turns out that our two-stage error policy achieves very robust performance to non-exponential service time distributions and time-varying arrival rates within a shift (Appendix H.3). However, the fact that each shift only lasts for 12 hours and the fact that the arrival rate for the day shift can be twice as large as the arrival rate for the night shift degrades the performance of our proposed policy.

Since the night shift has a much lower arrival rate than the day shift, the day shift usually starts with a lower patient volume than an otherwise stationary system; namely, the number of patients in the system at the beginning of the shift can be much lower than the steady-state average. Similarly, the night shift usually starts with a higher patient volume than an otherwise stationary system. Our proposed policy based on the stylized model is not able to capture these transient shift effects well. We next propose an adjustment to our two-stage error policy that takes these transient shift effects into account. In particular, at the base stage, we increase the base staffing level for night shifts and decrease the base staffing level for day shifts based on the mean arrival rates. Then at the surge stage, we further adjust the surge staffing level based on the current state of the system, i.e., the number of patients in the system towards the end of the current shift, and the predicted service rate of the next shift. Formally, the two-stage error policy is adjusted as follows:

Base Stage: For $1 \leq i \leq 14$, let $N_{1,i}$ denote the base staffing level for shift of type i under $u_{2,ERR}$, which is calculated using $\hat{\lambda}_i$, $\hat{\alpha}$, and the estimated distributions of Y and Z. For shift of type i, calculate the expected steady-state queue length for an $M/M/n+M$ queue with arrival rate $\hat{\lambda}_i$ and number of servers equal to $N_{1,i}$, and denote it by $\bar{Q}_i$. Let Δ_i denote the difference in the expected queue length between two consecutive shifts, i.e., $\Delta_i := \bar{Q}_{i-1} - \bar{Q}_i$, where $\bar{Q}_0 \equiv \bar{Q}_{14}$. The adjusted base-stage staffing level is given by $N_{1,i}^{Adj} := N_{1,i} + \xi_1 \Delta_i$, where $\xi_1 \in \mathbb{R}$ is some base adjustment parameter to be determined.

Surge Stage: For $1 \leq i \leq 14$, $1 \leq k \leq 52$, let $N_{2,i}^{(k)}$ denote the surge staffing level for shift of type i in the kth week under $u_{2,ERR}$, which is calculated using the predicted arrival rate $\hat{\ell}_i^{(k)}$. For each shift, calculate the expected steady-state queue length for an $M/M/n+M$ queue with arrival rate $\hat{\ell}_i^{(k)}$ and number of servers equal to $N_{1,i}^{Adj} + N_{2,i}^{(k)}$, and denote it by $\bar{Q}_i^{(k)}$. Let $Q_i^{(k)}$ be the number of patients in the ED at the end of the previous shift, and let $D_i^{(k)} := \bar{Q}_i^{(k)} - Q_i^{(k)}$. The adjusted surge-stage staffing level is given by $N_{2,i}^{(k),Adj} := N_{2,i}^{(k)} + \xi_2 D_i^{(k)}$, where $\xi_2 \in \mathbb{R}$ is some surge adjustment parameter to be determined.

In the adjustment above, Δ_i can be understood as the difference in the expected patient backlog between two consecutive shifts. Since day shifts have a higher expected queue length than night shifts, the base staffing level is raised for night shifts and reduced for day shifts. Similarly, $D_i^{(k)}$

can be understood as the difference between the actual backlog in the system towards the end of the previous shift and the expected value for the current shift. The surge staffing level is then increased or decreased dynamically depending $D_i^{(k)}$. When determining the base and surge adjustment parameters, we see from extensive numerical experiments that setting $\xi_1 \in [4,8]$ and $\xi_2 \in [1,2]$ gives consistently good system performance. Thus, we set $\xi_1 = 5$ and $\xi_2 = 1$ in the subsequent numerical experiments and suggest using these values in practice.

In what follows, we compare the ED-adapted two-stage error policy to the single-stage newsvendor solution in the hospital setting. To make the comparison fair, similar base adjustment is applied to the single-stage newsvendor solution, i.e., $N_{1,i}^{Adj} = N_{1,i} + 5\Delta_i$. For ease of reference, we keep the same names and acronyms for these ED-adapted policies.

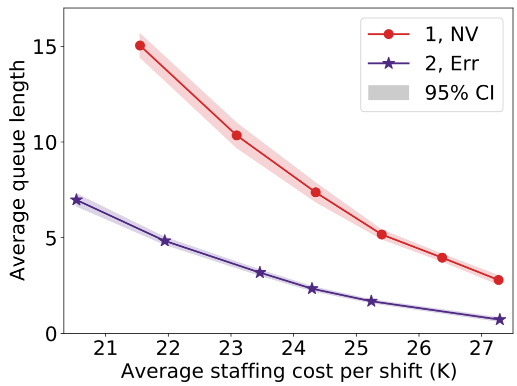
We remark that while it is true that the adjustment parameters, ξ_1 and ξ_2, can be optimized for system, for example, by enumerating of all possible combinations in the simulation. Calculating the exact optimal adjustment can be computationally intensive and the optimal value can be case-dependent. In Appendix H.4, we show through numerical experiments that setting $\xi_1 = 5$ and $\xi_2 = 1$ achieves near-optimal and robust performance.

7.4.3. Performance Evaluation In practice, it can be challenging to calibrate the holding and abandonment costs. To circumvent this difficulty, we fix the ratio between holding and abandonment cost to be 1.5, and calculate the staffing levels for a wide range of holding costs under each policy. In particular, for each holding cost, we calculate the staffing levels under $u_{2,ERR}$ and $u_{1,NV}$, and simulate the ED over 52 weeks to estimate various system performance measures, such as the average waiting time, average queue length, percentage of patients left without been seen, and percentage of patients whose waiting time exceeds 60 minutes. The same experiment is repeated 5 times using different random seeds to construct the 95% confidence intervals for the performance measures. This allows us to construct a tradeoff curve between the staffing costs and the system performances under different staffing rules; see Figure 7. We observe that the tradeoff curve of $u_{2,ERR}$ is strictly below those of $u_{1,NV}$. This suggests that for a fixed system performance target, we are able to achieve it with a much lower staffing cost under the two-stage staffing policy than the single-stage staffing policy.

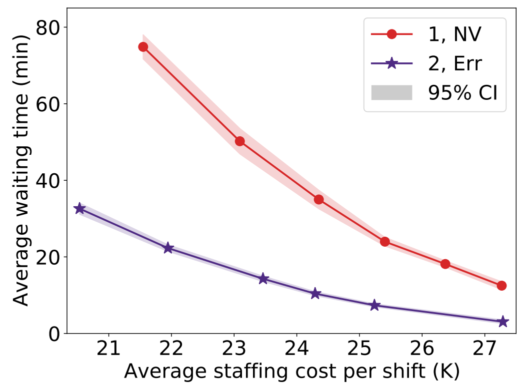
Given some specific performance targets, we calculate the staffing cost needed to achieve the desired service quality under each policy. Table 6 lists the saving in the annual staffing cost of $u_{2,ERR}$ in comparison to $u_{1,NV}$ in order to guarantee that (i) the average queue length is below 5, or (ii) the average waiting time is below 30 minutes, or (iii) the percentage of patients who left without been seen is less than 2%, or (iv) less than 20% of patients wait for more than 60 minutes.

We observe that we are able to achieve 10.67% to 15.86% (\$1.791 M to \$2.875 M) cost savings for different performance requirements. In a setting where many hospitals are operating on thin margins, such savings can have a great impact to the bottom line. Lastly, recall from Section 7.3 that the surge-stage linear regression model is able to improve the prediction accuracy in terms of RMSE at the base stage by 11.16%. Our numerical results suggest that even with this modest gain in prediction accuracy, this information, together with the real-time queue length information, can lead to significant cost savings while ensuring timely access to care.

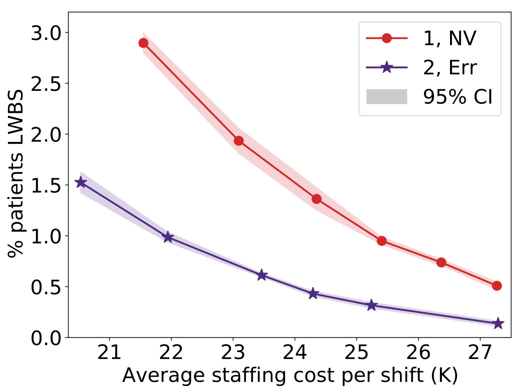
Figure 7 Tradeoff between staffing cost and quality of service



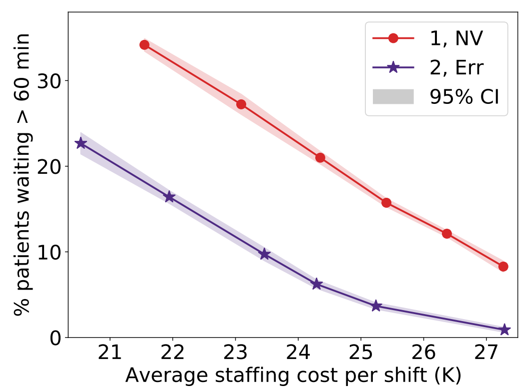
(a) Average queue length



(b) Average waiting time



(c) % patients LWBS



(d) % paitents waiting > 60 min

Table 6 Annual saving in staffing cost to achieve target performance

Policy	Avg queue length < 5	Avg waiting time < 30 min	% patients LWBS $< 2\%$	$< 20\%$ patients wait > 60 min
V.s. $u_{1,NV}$	\$2.704 M (14.51%)	\$2.875 M (15.86%)	\$1.791 M (10.67%)	\$2.493 M (13.91%)

8. Conclusion

In this paper, we study the prediction-driven surge staffing problem in the ED. A key tradeoff in this problem is the base-stage staffing, which is cheaper but faces a higher level of uncertainty versus the surge-stage staffing, which is more expensive but faces a lower level of uncertainty. Our analysis quantifies when surge staffing is beneficial and provides prescriptive staffing rules that are highly interpretable, easy to implement, and achieve near-optimal performance. Our analysis demonstrates that the benefits of surge staffing are substantial when the arrival-rate uncertainty dominates the system stochasticity. To capture this benefit, our proposed policy (for the case with perfect demand forecast) first aims to staff at the base stage by solving a newsvendor problem to serve the expected offered load. It then incorporates a square-root hedging against the system stochasticity at the surge stage. We then extend the analysis to account for prediction error explicitly at the surge stage. To facilitate implementation in the actual ED setting, we develop a unified framework that includes parameter estimation, building a two-scale prediction model that are synchronized with the staffing decision epoches, and developing an effective prediction-driven staffing rule. Using data from the Milstein ED in NYP CUMC, we demonstrate via high-fidelity simulation that our proposed staffing rule achieves significant cost savings for the hospital.

While our theoretical model is unable to capture all features of the real ED (e.g., time-varying arrivals, lognormal service times, etc.), we find that it is able to capture core tradeoffs to provide insights into the management of ED staffing. That said, we also find that transient effects of the large disparity in arrival rates between night and day can have a measurable impact on system performance. As such, it would be interesting as future research to explore a transient (rather than steady-state) analysis of our system. Since closed-form expressions for transient queuing dynamics are limited, new approximation techniques may need to be developed. Moreover, our model considers two discrete staffing epochs with different levels of demand information. Our view of the two-stage decision is informed by the current nurse staffing practice in hospitals. An interesting extension is to examine more granular decision epochs or even a continuous-time model, where both demand information and staffing cost increase as time approaches to the start of the shift. This requires a more granular model of arrival-rate uncertainty, such as those developed in Zhang et al. (2014), Daw and Pender (2018). However, increasing the granularity of decision epochs may also come with certain implementation challenges from the practical perspective.

References

Allon, G., J. A. Van Mieghem. 2010. Global dual sourcing: Tailored base-surge allocation to near-and offshore production. *Management Science* **56**(1) 110–124.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems* **5**(1) 146–194.

Avramidis, A. N., A. Deslauriers, P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* **50**(7) 896–908.

Ban, G.-Y., C. Rudin. 2019. The big data newsvendor: Practical insights from machine learning. *Operations Research* **67**(1) 90–108.

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

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

Batt, R. J., D. S. Kc, B. R. Staats, B. W. Patterson. 2019. The effects of discrete work shifts on a nonterminating service system. *Production and operations management* **28**(6) 1528–1544.

Bernstein, S. L., D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. John McConnell, J. M. Pines, N. Rathlev, et al. 2009. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine* **16**(1) 1–10.

Bertsimas, D., J. Pauphilet, J. Stevens, M. Tandon. 2021. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management* .

Boada-Collado, P., S. Chopra, K. Smilowitz. 2020. The value of information and flexibility with temporal commitments. *Available at SSRN 3452915* .

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

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association* **100**(469) 36–50.

Calegari, R., F. S. Fogliatto, F. R. Lucini, J. Neyeloff, R. S. Kuchenbecker, B. D. Schaan. 2016. Forecasting daily volume and acuity of patients in the emergency department. *Computational and mathematical methods in medicine* **2016** 1–8.

Chan, C. W., M. Huang, V. Sarhangian. 2021. Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. *Operations Research* .

Chang, A. M., D. J. Cohen, A. Lin, J. Augustine, D. A. Handel, E. Howell, H. Kim, J. M. Pines, J. D. Schuur, K. J. McConnell, et al. 2018. Hospital strategies for reducing emergency department crowding: a mixed-methods study. *Annals of emergency medicine* **71**(4) 497–505.

Chen, B. P., S. G. Henderson. 2001. Two issues in setting call centre staffing levels. *Annals of operations research* **108**(1) 175–192.

Chen, X., M. Sim, P. Sun. 2007. A robust optimization perspective on stochastic programming. *Operations research* **55**(6) 1058–1071.

Daw, A., J. Pender. 2018. Queues driven by hawkes processes. *Stochastic Systems* **8**(3) 192–229.

Erlang, A. K. 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal* **10** 189–197.

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

Green, L. V. 2010. Using queueing theory to alleviate emergency department overcrowding. *Wiley Encyclopedia of Operations Research and Management Science* .

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

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

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

Hu, Y., K. D. Cato, C. W. Chan, J. Dong, N. Gavin, S. C. Rossetti, B. P. Chang. 2021. Use of real-time information to predict future arrivals in the emergency department. Working Paper, Columbia Business School.

Ibrahim, R., H. Ye, P. L'Ecuyer, H. Shen. 2016. Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting* **32**(3) 865–874.

Janakiraman, G., S. Seshadri, A. Sheopuri. 2015. Analysis of tailored base-surge policies in dual sourcing inventory systems. *Management Science* **61**(7) 1547–1561.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* **42**(10) 1383–1394.

Kim, S.-H., W. Whitt. 2014. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management* **16**(3) 464–480.

Koçağa, Y. L., M. Armony, A. R. Ward. 2015. Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management* **24**(7) 1101–1117.

Kolesar, P. J., L. V. Green. 1998. Insights on service system design from a normal approximation to erlang's delay formula. *Production and Operations Management* **7**(3) 282–293.

Levi, R., G. Perakis, J. Uichanco. 2015. The data-driven newsvendor problem: new bounds and insights. *Operations Research* **63**(6) 1294–1306.

Liu, Y., W. Whitt. 2012. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations research* **60**(6) 1551–1564.

Maman, S. 2009. Uncertainty in the demand for service: The case of call centers and emergency departments. Ph.D. thesis, Field of Statistics. Technion - Israel Institute of Technology, Haifa, Israel.

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations research* **57**(5) 1189–1205.

Marcilio, I., S. Hajat, N. Gouveia. 2013. Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic emergency medicine* **20**(8) 769–777.

Perakis, G., G. Roels. 2008. Regret in the newsvendor model with partial information. *Operations research* **56**(1) 188–203.

Ramsey, Z., J. S. Palter, J. Hardwick, J. Moskoff, E. L. Christian, J. Bailitz. 2018. Decreased nursing staffing adversely affects emergency department throughput metrics. *Western Journal of Emergency Medicine* **19**(3) 496.

Schweigler, L. M., J. S. Desmond, M. L. McCarthy, K. J. Bukowski, E. L. Ionides, J. G. Younger. 2009. Forecasting models of emergency department crowding. *Academic Emergency Medicine* **16**(4) 301–308.

Smith, D. R., W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* **60**(1) 39–55.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2009. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences* **23**(2) 305.

Sun, X., Y. Liu. 2021. Staffing many-server queues with autoregressive inputs. *Naval Research Logistics (NRL)* **68**(3) 312–326.

Véricourt, F. d., O. B. Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations research* **59**(6) 1320–1331.

Weiss, M. E., O. Yakusheva, K. L. Bobay. 2011. Quality and cost analysis of nurse staffing, discharge preparation, and postdischarge utilization. *Health services research* **46**(5) 1473–1494.

Whitt, W. 1984. Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal* **63**(5) 689–708.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24**(5) 205–212.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Production and operations management* **15**(1) 88–102.

Whitt, W., X. Zhang. 2019. Forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care* **21** 1–18.

Xin, L., D. A. Goldberg. 2018. Asymptotic optimality of tailored base-surge policies in dual-sourcing inventory systems. *Management Science* **64**(1) 437–452.

Yankovic, N., L. V. Green. 2011. Identifying good nursing levels: A queuing approach. *Operations research* **59**(4) 942–955.

Yom-Tov, G. B., A. Mandelbaum. 2014. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16**(2) 283–299.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the m/m/n+ g queue. *Queueing Systems* **51**(3-4) 361–402.

Zhang, X., L. J. Hong, J. Zhang. 2014. Scaling and modeling of call center arrivals. *Proceedings of the Winter Simulation Conference 2014*. IEEE, 476–485.

Zheng, Z., H. Honnappa, P. W. Glynn. 2018. Approximating systems fed by poisson processes with rapidly changing arrival rates. *arXiv preprint arXiv:1807.06805* .

Appendix A: Roadmap for The Main Proofs

In this section, we introduce the notations used throughout the appendices, present a useful lemma, and give a roadmap for the organization of the main proofs.

Let $\alpha \in (0,1)$. Consider an admissible staffing policy $\pi \in \Pi$ with base staffing level N_1 and surge staffing level $N_2(N_1,\Lambda)$. For any realized arrival rate ℓ, the total cost under π is denoted by

$$\mathcal{C}_\pi(\ell) := c_1 N_1 + c_2 N_2(N_1,\ell) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1 + N_2(N_1,\ell),\ell)\right]. \tag{22}$$

We also write

$$\mathcal{C}_\pi(\Lambda) := c_1 N_1 + c_2 N_2(N_1,\Lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1 + N_2(N_1,\Lambda),\Lambda)|\Lambda\right], \quad \text{and} \quad \mathcal{C}_\pi := \mathbb{E}\left[\mathcal{C}(\Lambda)\right].$$

We use the following notations, in addition to the notations introduced in the main paper:

1. For an $M/M/m+M$ queue with m servers and arrival rate λ, we let $\mathbb{P}(AB,m,\lambda)$ denote the steady-state abandonment probability, $W(m,\lambda)$ denote the steady-state waiting time, and $V(m,\lambda)$ denote the steady-state virtual waiting time. $V(m,\lambda)$ is the time that a patient with infinite patience would wait and $W(m,\lambda)$ is the minimum of $V(m,\lambda)$ and the patient's patience time. Let $\mathbb{1}_{(AB,m,\lambda)}$ be the indicator of whether or not a customer arriving to a system in steady-state will abandon, i.e., $\mathbb{P}(AB,m,\lambda) = \mathbb{E}\left[\mathbb{1}_{(AB,m,\lambda)}\right]$. In what follows, we use $\mathbb{P}(AB,m,\Lambda)$ to denote the steady-state abandonment probability conditional on the random arrival rate, i.e., $\mathbb{P}(AB,m,\Lambda) := \mathbb{E}\left[\mathbb{1}_{(AB,m,\Lambda)}|\Lambda\right]$. In particular, $\mathbb{P}(AB,m,\Lambda)$ is a random variable. Similar convention for notation has been used in the literature; see, e.g., Koçağa et al. (2015).

2. For an $M/M/m/m$ queue with m servers and arrival rate λ, we let $\mathbb{P}(BL,m,\lambda)$ denote the steady-state blocking probability, $L(m,\lambda)$ denote the steady-state loss rate, and $\mathbb{1}_{(BL,m,\lambda)}$ be the indicator of whether or not a customer will be blocked in steady state. Note that $L(m,\lambda) = \lambda\mathbb{P}(BL,m,\lambda)$, and $\mathbb{P}(BL,m,\lambda) = \mathbb{E}\left[\mathbb{1}_{(BL,m,\lambda)}\right]$. In what follows, we let $\mathbb{P}(BL,m,\Lambda)$ denote the steady-state blocking probability conditional on the random arrival rate, i.e., $\mathbb{P}(BL,m,\Lambda) := \mathbb{E}\left[\mathbb{1}_{(BL,m,\Lambda)}|\Lambda\right]$. Similar to $\mathbb{P}(AB,m,\Lambda)$, $\mathbb{P}(BL,m,\Lambda)$ is a random variable.

3. For functions $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}$, we use the relation $f\sim k$ to denote that $\lim_{\lambda\to\infty} f(\lambda)/k(\lambda)=1$.

The following lemma will be used in the subsequent development.

Lemma 1 *For the multi-server queue with abandonment,*

$$\mathbb{E}\left[Q(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right]\le \max\{\mu/\gamma,1\}\left((\ell/\mu-N_1-N_2(N_1,\ell))^+ + \sqrt{4\pi/\mu}\sqrt{\ell}+1/\log 2\right). \tag{23}$$

Proof: We conduct the proof in three cases: $\mu=\gamma$, $\mu<\gamma$, and $\mu>\gamma$.

Case 1: $\boldsymbol{\mu=\gamma}$. In this case, Lemma 3 in Bassamboo et al. (2010) directly implies that

$$\mathbb{E}\left[Q(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right]\le (\ell/\mu-N_1-N_2(N_1,\ell))^+ + \sqrt{4\pi/\mu}\sqrt{\ell}+1/\log 2,$$

from which (23) follows.

Case 2: $\boldsymbol{\mu<\gamma}$. In this case, we consider a sequence of auxiliary systems with abandonment rate μ (as opposed to γ), and every other parameter is held the same as in the original system. Comparing the underlying Markov chains of these two sequences of systems, we see that the steady-state queue length in the auxiliary system is stochastically larger than that in the original system. In particular, let $\mathbb{E}\left[\tilde{Q}(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda\right]$ denote the conditional expectation of the steady-state queue in the auxiliary system. It holds that

$$\mathbb{E}\left[Q(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right]\le \mathbb{E}\left[\tilde{Q}(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right].$$

We can apply the same arguments as in Case 1 to the auxiliary system, and infer (23).

Case 3: $\boldsymbol{\mu>\gamma}$. In this case, we consider a sequence of auxiliary systems with abandonment rate μ (as opposed to γ), and every other parameter is held the same as in the original system. Following similar arguments as in the proof of Theorem 3 in Bassamboo et al. (2010), we get that the steady-state abandonment rate in the auxiliary system is larger than that in the original system. In particular, let $\mathbb{P}\left(\tilde{AB},N_1+N_2(N_1,\Lambda),\Lambda\right)$ denote the steady-state abandonment rate in the auxiliary system. It holds that

$$\mathbb{P}\left(AB,N_1+N_2(N_1,\ell),\ell\right)\le \mathbb{P}\left(\tilde{AB},N_1+N_2(N_1,\ell),\ell\right).$$

Since the steady-state abandonment rate must be equal to the steady-state arrival rate of abandoning patients, we have

$$\mu\mathbb{E}\left[\tilde{Q}(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right]=\ell\mathbb{P}\left(\tilde{AB},N_1+N_2(N_1,\ell),\ell\right),$$

and

$$\gamma\mathbb{E}\left[Q(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right]=\ell\mathbb{P}\left(AB,N_1+N_2(N_1,\ell),\ell\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[Q(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right]&=(\ell/\gamma)\mathbb{P}\left(AB,N_1+N_2(N_1,\ell),\ell\right)\\&\le(\ell/\gamma)\mathbb{P}\left(\tilde{AB},N_1+N_2(N_1,\ell),\ell\right)\\&=(\mu/\gamma)\mathbb{E}\left[\tilde{Q}(N_1+N_2(N_1,\Lambda),\Lambda)|\Lambda=\ell\right].\end{aligned}$$

We can apply the same arguments as in Case 1 to the auxiliary system, and (23) follows. Q.E.D.

Appendices B–F contain the proofs of the main results. In Appendix B, we prove Proposition 1 which specifies the nontrivial cost parameter regime for the staffing problem. In Appendix C, we introduce a general family of two-stage staffing policies for all $\alpha \in (0,1)$. We refer to this policy as the two-stage uncertainty hedging rule, and derive its asymptotic performance in Appendices C.1 (for $\alpha > 1/2$) and C.2 (for $\alpha \leq 1/2$). In Appendix C.3, we prove that the two-stage uncertainty hedging rule with properly selected parameters achieves an optimality gap of $o(\lambda^{\max\{1/2,\alpha\}})$ compared to the exact two-stage optimum. As the two-stage newsvendor solution is a special case of the two-stage two-stage uncertainty hedging rule when $\alpha > 1/2$, the optimality gap of the two-stage newsvendor solution (Theorem 2) follows (see Appendix C.4). In Appendix D, we prove Theorem 1 which characterizes the cost saving of the optimal two-stage staffing rule compared to the optimal single-stage policy. This is done by combining the cost quantification under different near-optimal staffing rules and the corresponding optimality gap results. For example, when $\alpha > 1/2$, we first compare the cost under the two-stage newsvendor rule and the single-stage newsvendor rule. We then use the optimality gap of the single-stage newsvendor solution (compared to the single-stage optimal) and the optimality gap of the two-stage newsvendor solution (compared to the two-stage optimal) to quantify the cost saving. In Appendix E, we prove Theorem 3, where we show that the two-stage square-root staffing rule refines the two-stage newsvendor solution and further reduces the optimality gap. Lastly in Appendix F, we analyze the two-stage staffing problem with surge-stage prediction errors. The results for small prediction errors (Proposition 2) are proved in Appendix F.1 and the results for moderate to large prediction errors (Proposition 3) are proved in Appendix F.2

Appendix B: Proof of Proposition 1

Proof: Consider an admissible staffing policy $\pi \in \Pi$ with base staffing level N_1 and surge staffing level $N_2(N_1,\Lambda)$. For any realized arrival rate ℓ, we let $B_1(N_1,N_2(N_1,\ell),\ell)$ denote the steady-state number of busy servers among those that are staffed at the base stage, and let $B_2(N_1,N_2(N_1,\ell),\ell)$ denote the steady-state number of busy servers among those that are staffed at the surge stage. It holds that

$$B_1(N_1,N_2(N_1,\ell),\ell) \leq N_1 \quad \text{and} \quad B_2(N_1,N_2(N_1,\ell),\ell) \leq N_2(N_1,\ell). \tag{24}$$

Note that for $B_1(N_1,N_2(N_1,\ell),\ell)$ and $B_2(N_1,N_2(N_1,\ell),\ell)$ to be well-defined, we need to specify the assignment policy of patients to the base and surge servers. Since the model does not distinguish base and surge servers (i.e., they provide the same quality of service), we assume that patients are randomly assigned to the available servers with equal probability. That said, (24) holds regardless of the assignment policy.

Proof of (I). Following (22), the total cost satisfies

$$\begin{aligned}
&\mathcal{C}_\pi(\ell) = c_1 N_1 + c_2 N_2(N_1,\ell) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\
&\geq c_1 \mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + c_2 \mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \left(\frac{h\mu}{\gamma} + a\mu\right)\frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\
&\geq \min\left\{c_1, c_2, \frac{h\mu}{\gamma} + a\mu\right\}\left(\mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + \mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right]\right) \\
&= \left(\frac{h\mu}{\gamma} + a\mu\right)\frac{\ell}{\mu} \\
&= \left(\frac{h}{\gamma} + a\right)\ell,
\end{aligned} \tag{25}$$

where the second to last equality in (25) follows from the steady-state balance equation:

$$\begin{aligned} \ell &= \mu\mathbb{E}\left[B_1(N_1, N_2(N_1,\ell),\ell)\right] + \mu\mathbb{E}\left[B_2(N_1, N_2(N_1,\ell),\ell)\right] + \gamma\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\ \frac{\ell}{\mu} &= \mathbb{E}\left[B_1(N_1, N_2(N_1,\ell),\ell)\right] + \mathbb{E}\left[B_2(N_1, N_2(N_1,\ell),\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right]. \end{aligned} \tag{26}$$

Moreover, the cost lower bound in (25) can be achieved by staffing zero base and zero surge servers. To see this, let π_0 denote the "zero-staff" policy under which all customers abandon. The long-run average cost for the realized arrival rate ℓ under π_0 is

$$\mathcal{C}_{\pi_0}(\ell) = c_1 0 + c_2 0 + (h + a\gamma)\,\mathbb{E}\left[Q(0,\ell)\right] = (h+a\gamma)\,\mathbb{E}\left[Q(0,\ell)\right].$$

By flow balance, the steady-state rate at which abandoning customers arrive must be equal to the abandonment rate, namely,

$$\ell = \gamma\mathbb{E}\left[Q(0,\ell)\right],$$

which gives that $\mathcal{C}_{\pi_0}(\ell) = (h + a\gamma)\,\ell/\gamma$. Hence, π_0 achieves the cost lower bound, and is optimal to the optimization problem (2).

Proof of (II). Based on π, we construct another admissible policy π' where $\pi' := (0, N_2(N_1,\Lambda) + N_1)$. Namely, if π assigns N_1 base servers and $N_2(N_1,\Lambda)$ surge servers, then π' assigns zero base servers and $N_2(N_1,\Lambda) + N_1$ surge servers. By assumption, either $h\mu/\gamma + a\mu > c_1 > c_2$ or $c_1 > h\mu/\gamma + a\mu > c_2$. It follows from (22) that $\mathcal{C}_{\pi'}(\Lambda) \le \mathcal{C}_{\pi}(\Lambda)$. Thus, it is optimal to set $N_1^* = 0$.

Proof of (III). Based on π, we construct another admissible policy π' where $\pi' := (N_1, 0)$. Namely, π' assigns the same number of base servers as π but zero surge servers for any realized arrival rate. Following (22), the total cost satisfies

$$\begin{aligned} \mathcal{C}_{\pi}(\ell) &= c_1 N_1 + c_2 N_2(N_1,\ell) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\ &\ge c_1 N_1 + c_2\mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \left(\frac{h\mu}{\gamma} + a\mu\right)\frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\ &\ge c_1 N_1 + \left(\frac{h\mu}{\gamma} + a\mu\right)\left(\mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right]\right) \\ &\ge c_1 N_1 + \left(\frac{h\mu}{\gamma} + a\mu\right)\left(\mathbb{E}\left[B_2(N_1,0,\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1,\ell)\right]\right) \\ &= c_1 N_1 + \left(\frac{h\mu}{\gamma} + a\mu\right)\left(0 + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1,\ell)\right]\right) \\ &= \mathcal{C}_{\pi'}(\ell), \end{aligned}$$

where the last inequality follows by observing from (26) that

$$\begin{aligned} &\mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + \mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\ &= \mathbb{E}\left[B_1(N_1,0,\ell)\right] + \mathbb{E}\left[B_2(N_1,0,\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1,\ell)\right] \\ &= \frac{\ell}{\mu}, \end{aligned}$$

and that

$$\mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] \le \mathbb{E}\left[B_1(N_1,0,\ell)\right].$$

Thus, it is optimal to set $N_2^*(N_1,\Lambda) = 0$. Q.E.D.

Appendix C: Two-Stage Uncertainty Hedging Rule

For most of the theoretical development starting from this section, we consider the asymptotic behavior of the system as the mean arrival rate λ grows without bound. Thus, throughout Appendices C–E, we add superscript λ to all the quantities that scale with λ. For example, we add the superscript λ in N_1^λ and $N_2^\lambda(N_1^\lambda, \Lambda^\lambda)$ to denote the dependence of the staffing levels on the mean arrival rate. We use U to denote the set of all sequences of admissible staffing polices. The set U contains policies in form of $u = \{\pi^\lambda : \pi^\lambda \in \Pi^\lambda\}$, where u is a sequence of policies that specifies a two-stage staffing decision for each system along the sequence. Whenever needed, we add the subscript u to the costs (e.g., $\mathcal{C}_u^\lambda$) to mark the dependence of the cost on the staffing policy explicitly.

To facilitate the asymptotic analysis, we re-center and scale the total cost by defining

$$\hat{\mathcal{C}}_u^\lambda(\Lambda) := \frac{\mathcal{C}_u^\lambda(\Lambda) - c_1\lambda/\mu}{(\lambda/\mu)^{\max\{\alpha,1/2\}}}, \quad \text{and} \quad \hat{\mathcal{C}}_u^\lambda := \mathbb{E}\left[\hat{\mathcal{C}}_u^\lambda(\Lambda)\right]. \tag{27}$$

To simplify notation, we denote the sum of the surge-stage staffing and queueing-related cost by

$$\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) := c_2 N_2^\lambda(N_1^\lambda, \ell^\lambda) + (h + a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right]. \tag{28}$$

Replacing the realized arrival rate ℓ^λ with Λ^λ in (28), we define

$$\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda) := c_2 N_2^\lambda(N_1^\lambda, \Lambda^\lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)|\Lambda^\lambda\right],$$

where the expectation operator on the right-hand side is with respect to the queue process. Note that $\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)$ is a constant while $\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)$ is a random variable.

The proofs of the main theorems require analyzing near-optimal staffing polices. In this section, we propose the *two-stage uncertainty hedging rules* and denote it by $u_{2,UH}$. We characterize the system performance under $u_{2,UH}$ as the mean arrival rate λ increases to infinity. We also show that the two-stage newsvendor solution is a special case of the two-stage uncertainty hedging rule. The proof of Theorem 2 follows.

Consider the following staffing policy, which we denote as $u_2(\beta_1, \beta_2(\beta_1, X))$. At the base stage, the base staffing level is set as

$$N_1^\lambda = \lambda/\mu + \beta_1 (\lambda/\mu)^{\max\{\alpha,1/2\}} + o((\lambda/\mu)^{\max\{\alpha,1/2\}}),$$

for $\beta_1 \in \mathbb{R}$. Note that the base staffing level is set to meet the mean demand, together with a hedging that is of the same order as the arrival-rate uncertainty or system stochasticity, whichever is larger. At the surge stage, after the random arrival rate realizes, the surge staffing level is set to

$$N_2^\lambda(N_1^\lambda, \Lambda^\lambda) = \beta_2(\beta_1, X)\,(\lambda/\mu)^{\max\{\alpha,1/2\}} + o_{UI}((\lambda/\mu)^{\max\{\alpha,1/2\}}),$$

where the coefficient $\beta_2(\beta_1, X) \in \mathbb{R}_+$ depends on both the base staffing level and the realized arrival rate. Note that the surge staffing level serves as another hedging against the larger part of arrival-rate uncertainty and system stochasticity. Importantly, the parameter $(\beta_1, \beta_2(\beta_1, X))$ does not scale with λ.

We also denote

$$D_1^\lambda := N_1^\lambda - \lambda/\mu - \beta_1 (\lambda/\mu)^{\max\{\alpha,1/2\}} = o((\lambda/\mu)^{\max\{\alpha,1/2\}})$$

and

$$D_2^\lambda(N_1^\lambda,\Lambda^\lambda) := N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - \beta_2(\beta_1,X)\,(\lambda/\mu)^{\max\{\alpha,1/2\}} = o_{UI}((\lambda/\mu)^{\max\{\alpha,1/2\}}).$$

Note that D_1^λ is a constant. On the other hand, $D_2^\lambda(N_1^\lambda,\Lambda^\lambda)$ may depend on the realization of Λ^λ and is thus a random variable. Recall from Section 1.3 that by $D_2^\lambda(N_1^\lambda,\Lambda^\lambda) = o_{UI}((\lambda/\mu)^{\max\{\alpha,1/2\}})$, we mean that $D_2^\lambda(N_1^\lambda,\Lambda^\lambda)/\,(\lambda/\mu)^{\max\{\alpha,1/2\}} \to 0$ as $\lambda\to\infty$ with probability 1, and there exists some random variable Y with $\mathbb{E}[Y]<\infty$ such that

$$|D_2^\lambda(N_1^\lambda,\Lambda^\lambda)|/\,(\lambda/\mu)^{\max\{\alpha,1/2\}} < Y \quad \text{for all } \lambda>0. \tag{29}$$

We remark that (29) is not restrictive and allows for a wide range of capacity prescriptions. Examples for $D_2^\lambda(N_1^\lambda,\Lambda^\lambda)$ include $(\lambda/\mu)^\tau$ and $(\lambda/\mu)^\tau X$ for $\tau\in(0,\max\{\alpha,1/2\})$, etc.

The two-stage uncertainty hedging rule is defined by properly optimizing the staffing parameter $(\beta_1,\beta_2(\beta_1,X))$ in $u_2(\beta_1,\beta_2(\beta_1,X))$. In particular, we first derive a proper limit for the scaled total cost under $u_2(\beta_1,\beta_2(\beta_1,X))$. Then, $(\beta_1^*,\beta_2^*(\beta_1^*,X))$ is defined as the optimal solution to the limiting cost function.

C.1. Two-Stage Uncertainty Hedging Rule for $\alpha>1/2$

For any realized arrival rate $\ell^\lambda=\lambda+x\lambda^\alpha\mu^{1-\alpha}$, under $u_2(\beta_1,\beta_2(\beta_1,x))$ with parameters β_1 and $\beta_2(\beta_1,x)$, the total staffing level can be written as

$$\begin{aligned} N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda) &= \lambda/\mu + (\beta_1+\beta_2(\beta_1,x))\,(\lambda/\mu)^\alpha + o((\lambda/\mu)^\alpha)\\ &= \frac{\lambda + x\lambda^\alpha\mu^{1-\alpha}}{\mu} + \frac{(\lambda/\mu)^\alpha\,(\beta_1+\beta_2(\beta_1,x)-x)}{(\lambda/\mu+(\lambda/\mu)^\alpha\,x)^\alpha}\left(\frac{\lambda+\lambda^\alpha\mu^{1-\alpha}x}{\mu}\right)^\alpha + o((\lambda/\mu)^\alpha)\\ &= \ell^\lambda/\mu + (\beta_1+\beta_2(\beta_1,x)-x)\left(\ell^\lambda/\mu\right)^\alpha + o((\ell^\lambda/\mu)^\alpha). \end{aligned} \tag{30}$$

Let $\tilde\beta := \beta_1+\beta_2(\beta_1,x)-x$. We first prove an auxiliary lemma on the asymptotic behavior of the steady-state probability of waiting and steady-state probability of abandonment, which facilitates our subsequent analysis on the asymptotic behavior of $\mathcal{R}^\lambda$. The lemma is adapted from Theorem 4.1 and Theorem 4.2 in Maman (2009).

Lemma 2 *Assume that $\alpha>1/2$. For any sequence of realized arrival rate $\ell^\lambda=\lambda+x\lambda^\alpha\mu^{1-\alpha}$, under $u_2(\beta_1,\beta_2(\beta_1,x))$ with parameters β_1 and $\beta_2(\beta_1,x)$, the multi-server queue with abandonment satisfies:*

(i) If $\beta_1+\beta_2(\beta_1,x)>x$, then the delay probability converges to zero exponentially fast as $\lambda\to\infty$. Specifically, for λ large enough,

$$\begin{aligned}&\mathbb{P}\left(W(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)>0\right)\\ &< \frac{1}{\tilde\beta\sqrt{2\pi}}\frac{1}{(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda))^{\alpha-1/2}}\exp\left\{-\frac{(\ell^\lambda/\mu-(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda))+1)^2}{2((N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda))-1)}\right\}.\end{aligned}$$

The probability to abandon of delayed patients decreases at rate $1/(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda))^\alpha$, i.e.,

$$\mathbb{P}\left(AB, N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda | V(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)>0\right) \sim \frac{1}{(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda))^\alpha}\frac{\gamma}{\mu\tilde\beta}.$$

(ii) If $\beta_1+\beta_2(\beta_1,x)<x$, then the delay probability converges to 1 exponentially fast as $\lambda\to\infty$. Specifically, for λ large enough,

$$\mathbb{P}\left(W(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)=0\right) < \frac{1}{|\tilde\beta|\mu^{1-\alpha}(\ell^\lambda)^\alpha}\exp\left\{-\frac{\tilde\beta^2}{8\gamma}\mu^{2-2\alpha}(\ell^\lambda)^{2\alpha-1}\right\}.$$

The probability to abandon of delayed patients decreases at rate $1/(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda))^{1-\alpha}$, *i.e.,*

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda | V(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) > 0\right) \sim \frac{|\tilde{\beta}|}{(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda))^{1-\alpha}}.$$

Proof: Following (30), for total staffing level of the form

$$\ell^\lambda/\mu + (\beta_1 + \beta_2(\beta_1, x) - x)\left(\ell^\lambda/\mu\right)^\alpha + f(\ell),$$

where $f(\ell^\lambda) = o(\sqrt{\ell^\lambda})$, the statement of Lemma 2 follows directly from Theorem 4.1 and Theorem 4.2 from Maman (2009). The work left is to generalize the result to staffing level of the form in (30), where $f(\ell^\lambda) = o((\ell^\lambda)^\alpha)$.

To this end, we show that the proofs of Theorem 4.1 and Theorem 4.2 in Maman (2009) can be generalized to the case where $f(\ell^\lambda) = o((\ell^\lambda)^\alpha)$. Indeed, exactly the same lines of derivation go through when $f(\ell^\lambda) = o((\ell^\lambda)^\alpha)$ (as opposed to $f(\ell^\lambda) = o(\sqrt{\ell^\lambda})$). Just as in Maman (2009), the results follow from Lemmas 4.2 and 4.3 which need to be adapted to this more generalized setting. We next illustrate the generalization of Lemma 4.2 to the general case where $f(\ell^\lambda) = o((\ell^\lambda)^\alpha)$. The other proofs are generalized similarly.

In the proof of Lemma 4.2 in Maman (2009), four places utilize the fact that $f(\ell^\lambda) = o(\sqrt{\ell^\lambda})$. We discuss them one by one. For the rest of this proof, we assume that $\tilde{\beta} > 0$, as in the proof of Lemma 4.2. All numbering of the equations refers to those in Section 4 of Maman (2009).

First, let $\bar{G}(u) := e^{-\gamma u}$ denote the ccdf of the patience time distribution. Following (4.44) and using the definition of δ in (4.40), take

$$\tilde{\gamma} := \frac{1 - \bar{G}(\delta/2)}{2} > 0.$$

Since $\bar{G}(u) < 1$ for all $u > 0$, and $\bar{G}(u) - 1 < -2\tilde{\gamma}$ for all $u > \delta/2$, we get that for λ large enough,

$$\ell^\lambda(\bar{G}(u) - 1) - \tilde{\beta}(\ell^\lambda)^\alpha \mu^{1-\alpha} - f(\ell^\lambda)\mu \le -\tilde{\beta}(\ell^\lambda)^\alpha \mu^{1-\alpha}, \quad \text{for all } u > 0,$$

and

$$\ell^\lambda(\bar{G}(u) - 1) - \tilde{\beta}(\ell^\lambda)^\alpha \mu^{1-\alpha} - f(\ell^\lambda)\mu \le -\tilde{\gamma}\ell^\lambda, \quad \text{for all } u > \delta/2.$$

Therefore, (4.45) and (4.46) hold for the case where $f(\ell^\lambda) = o((\ell^\lambda)^\alpha)$.

Second, in (4.51), define the function

$$r(\ell^\lambda) := \frac{-\tilde{\beta}(\ell^\lambda)^\alpha \mu^{1-\alpha} x - f(\ell^\lambda)\mu x}{\tilde{\beta}\mu^{1-\alpha} x}.$$

Note that for $f(\ell^\lambda) = o((\ell^\lambda)^\alpha)$, we still have $r(\ell^\lambda) \sim (\ell^\lambda)^\alpha$. Therefore, (4.51) still holds by applying Lemma 2.1 in Maman (2009) with $m = 0, k_1 = \alpha, l_1 = 1, k_2 = 1, l_2 = 2$.

Third, utilizing the same fact that $r(\ell^\lambda) \sim (\ell^\lambda)^\alpha$, (4.55) goes through by applying Lemma 2.1 in Maman (2009) with $m = 1, k_1 = \alpha, l_1 = 1, k_2 = 1, l_2 = 2$.

Lastly, for

$$n := N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda) = \ell^\lambda/\mu + \tilde{\beta}\left(\ell^\lambda/\mu\right)^\alpha + o((\ell^\lambda/\mu)^\alpha),$$

it holds that

$$\frac{(\ell^\lambda/\mu - n + 1)^2}{2(n-1)} \sim \frac{\tilde{\beta}^2}{\mu^{2\alpha-1}}(\ell^\lambda)^{2\alpha-1},$$

so the last line in the proof of Lemma 4.2 goes through. Q.E.D.

Lemma 3 *Assume that $\alpha > 1/2$. For any sequence of realized arrival rates $\ell^\lambda = \lambda + x\lambda^\alpha\mu^{1-\alpha}$, under $u_2(\beta_1, \beta_2(\beta_1, x))$ with parameters β_1 and $\beta_2(\beta_1, x)$, we have*

$$\frac{1}{(\lambda/\mu)^\alpha}\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) \to \hat{r}(\beta_1, \beta_2(\beta_1, x), x) \quad \text{as } \lambda \to \infty,$$

where the function $\hat{z}: \mathbb{R}\times\mathbb{R}_+\times\mathbb{R}\to\mathbb{R}_+$ is defined as

$$\hat{r}(\beta_1, \beta_2(\beta_1, x), x) := \begin{cases} c_2\beta_2(\beta_1, x) & \text{if } \beta_1 + \beta_2(\beta_1, x) \geq x \\ c_2\beta_2(\beta_1, x) + (h\mu/\gamma + a\mu)\,(x - \beta_1 - \beta_2(\beta_1, x)) & \text{if } \beta_1 + \beta_2(\beta_1, x) < x. \end{cases} \tag{31}$$

Proof: It follows from (2.8)–(2.11) in Maman (2009) that when the patience time is exponentially distributed, we have

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda\right) = \mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda | V(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) > 0\right)$$
$$\mathbb{P}\left(W(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) > 0\right).$$

By Lemma 2 and the flow balance equation that

$$\ell^\lambda\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda\right) = \gamma\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right],$$

the following cases hold:

(i) If $\beta_1 + \beta_2(\beta_1, x) > x$, then for λ large enough,

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda\right)$$
$$< \frac{\gamma}{\mu\tilde{\beta}^2\sqrt{2\pi}}\frac{1}{(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda))^{2\alpha-1/2}}\exp\left\{-\frac{(\ell^\lambda/\mu - (N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda)) + 1)^2}{2((N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda)) - 1)}\right\}.$$

Therefore,

$$\lim_{\lambda\to\infty}\frac{1}{\sqrt{\lambda/\mu}}\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] = 0. \tag{32}$$

(ii) If $\beta_1 + \beta_2(\beta_1, x) < x$, then for λ large enough,

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda\right) \sim \frac{|\tilde{\beta}|}{(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda))^{1-\alpha}}.$$

Therefore,

$$\lim_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^\alpha}\mathbb{E}\left[Q(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] = \frac{\mu}{\gamma}(x - \beta_1 - \beta_2(\beta_1, x)). \tag{33}$$

Lastly, when $\beta_1 + \beta_2(\beta_1, x) = x$, we get from Lemma 1 that

$$\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] \leq \max\{\mu/\gamma, 1\}\left(\left(\ell^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, \ell^\lambda)\right)^+ + \sqrt{4\pi/\mu}\sqrt{\ell^\lambda} + 1/\log 2\right)$$
$$= o((\lambda/\mu)^\alpha) + \max\{\mu/\gamma, 1\}\sqrt{4\pi/\mu}\sqrt{\ell^\lambda} + \max\{\mu/\gamma, 1\}/\log 2.$$

Then,

$$\lim_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^\alpha}\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] = 0. \tag{34}$$

The statement of the lemma then follows from (32), (33), and (34). Q.E.D.

Based on Lemma 3, let β_1^* and $\beta_2^*(\beta_1, X)$ be the optimal solution to

$$\min_{\beta_1\in\mathbb{R}}\left\{c_1\beta_1 + \mathbb{E}\left[\min_{\beta_2(\beta_1, X)\in\mathbb{R}_+}\hat{r}(\beta_1, \beta_2(\beta_1, X), X)\right]\right\}, \quad \text{for } \hat{z} \text{ defined in (31)}.$$

It is straightforward to derive that

$$\beta_1^* = \arg\min_{\beta\in\mathbb{R}} c_1\beta + c_2\mathbb{E}\left[(X-\beta)^+\right] = \bar{F}_X^{-1}(c_1/c_2), \quad \text{and} \quad \beta_2^*(\beta_1, X) = (X - \beta_1)^+. \tag{35}$$

Then, the two-stage uncertainty hedging rule is defined as $u_2(\beta_1, \beta_2(\beta_1, X))$ with parameters β_1^* and $\beta_2^*(\beta_1^*, X)$ in (35). Note that $u_{2,UH}$ is exactly the two-stage newsvendor solution in Definition 1.

The next lemma establishes the asymptotic performance of $u_{2,UH}$.

Lemma 4 *Assume that* $\alpha > 1/2$*. Under the two-stage uncertainty hedging rule defined in* (35) *(equivalently, the two-stage newsvendor solution), we have*

$$\hat{\mathcal{C}}^\lambda \to c_1\beta_1^* + \mathbb{E}\left[\hat{r}(\beta_1^*, \beta_2^*(\beta_1^*, X), X)\right] \quad \text{as } \lambda \to \infty,$$

for $\hat{z}$ *defined in* (31).

Proof: It follows from Lemma 3 that

$$\hat{\mathcal{C}}^\lambda(\Lambda^\lambda) \to c_1\beta_1^* + \hat{r}\left(\beta_1^*, \beta_2^*(\beta_1^*, X), X\right) \quad w.p.1 \quad \text{as } \lambda \to \infty.$$

Hence, to prove the claim, it is sufficient to show that

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\hat{\mathcal{C}}^\lambda(\Lambda^\lambda)\right] = \mathbb{E}\left[\lim_{\lambda\to\infty} \hat{\mathcal{C}}^\lambda(\Lambda^\lambda)\right] \tag{36}$$

To this end, we utilize the dominated convergence theorem.

Note that

$$\begin{aligned}\hat{\mathcal{C}}^\lambda(\Lambda^\lambda) =& c_1\beta_1^* + c_2\beta_2^*(\beta_1^*, X) + \frac{1}{(\lambda/\mu)^\alpha}\left(D_1^\lambda + D_2^\lambda(N_1^\lambda, \Lambda^\lambda)\right) \\ &+ \frac{1}{(\lambda/\mu)^\alpha}(h + a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)|\Lambda^\lambda\right].\end{aligned} \tag{37}$$

For the first two terms on the right-hand side of (37), it follows from the definition of $\beta_2^*(\beta_1^*, X)$ that

$$|c_1\beta_1^*| + |c_2\beta_2^*(\beta_1^*, X)| \le c_2\left(|\beta_1^*| + |X|\right),$$

where recall that $\mathbb{E}\left[|X|\right] < \infty$.

For the third term on the right-hand side of (37), note that D_1^λ is a constant that is $o((\lambda/\mu)^\alpha)$. This, together with (29), implies that there exists some random variable $\tilde{Y}$ with $\mathbb{E}[\tilde{Y}] < \infty$ such that

$$\frac{1}{(\lambda/\mu)^\alpha}\left(|D_1^\lambda| + |D_2^\lambda(N_1^\lambda, \Lambda^\lambda)|\right) < \tilde{Y}.$$

For the last term on the right-hand side of (37), we utilize Lemma 1 to get that

$$\begin{aligned}&\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)|\Lambda^\lambda\right] \\ &\le \max\{\mu/\gamma, 1\}\left(\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, \Lambda^\lambda)\right)^+ + \sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right) \\ &\le \max\{\mu/\gamma, 1\}\left(\left(\Lambda^\lambda/\mu - N_1^\lambda\right)^+ + \sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right) \\ &= \max\{\mu/\gamma, 1\}\left(\left((X - \beta_1^*)(\lambda/\mu)^\alpha - D_1^\lambda\right)^+ + \sqrt{4\pi/\mu}\sqrt{\lambda/\mu + X\lambda^\alpha\mu^{1-\alpha}} + 1/\log 2\right) \\ &\le \max\{\mu/\gamma, 1\}\left((|X| + |\beta_1^*|)(\lambda/\mu)^\alpha + |D_1^\lambda| + \sqrt{4\pi/\mu}\sqrt{\lambda/\mu} + \sqrt{4\pi/\mu}\sqrt{|X|\lambda^\alpha\mu^{1-\alpha}} + 1/\log 2\right).\end{aligned} \tag{38}$$

In (38), $D_1^\lambda = o((\lambda/\mu)^\alpha)$ is a constant. In addition, for λ large enough, we have

$$\frac{1}{(\lambda/\mu)^\alpha}\sqrt{4\pi/\mu}\sqrt{|X|\lambda^\alpha\mu^{1-\alpha}} \le \sqrt{4\pi/\mu}\sqrt{|X|}.$$

By Jensen's inequality, $\mathbb{E}\left[\sqrt{|X|}\right] \le \sqrt{\mathbb{E}\left[|X|\right]} < \infty$. Therefore, there exists some random variable Y with $\mathbb{E}\left[Y\right] < \infty$, such that

$$\frac{1}{(\lambda/\mu)^\alpha}(h + a\gamma)\mathbb{E}\left[Q^\lambda(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)|\Lambda^\lambda\right] \le Y.$$

Therefore, $|\hat{\mathcal{C}}^\lambda(\Lambda^\lambda)|$ in (37) is uniformly bounded by an integrable random variable, and (36) is justified. Q.E.D.

C.2. Two-Stage Uncertainty Hedging Rule for $\alpha \le 1/2$

Recall that ϕ and Φ are the pdf and cdf of the standard normal random distribution, respectively. The hazard rate of the standard normal distribution is $H(t) = \phi(t)/\Phi(-t)$, for $t \in \mathbb{R}$.

Lemma 5 *Assume that $\alpha \le 1/2$. For any sequence of realized arrival rate $\ell^\lambda = \lambda + x\lambda^\alpha\mu^{1-\alpha}$, under $u_2(\beta_1, \beta_2(\beta_1, x))$ with parameters β_1 and $\beta_2(\beta_1, x)$, we have*

$$\frac{1}{\sqrt{\lambda/\mu}}\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) \to \hat{r}\left(\beta_1, \beta_2(\beta_1, x), x\right) \quad \text{as } \lambda \to \infty,$$

where the function $\hat{z} : \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ is defined as

$$\hat{r}\left(\beta_1, \beta_2(\beta_1, x), x\right) := c_2\beta_2(\beta_1, x) + \left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right) - \left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)}{H\left(-\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\right)}}. \tag{39}$$

PROOF: For any realized arrival rate $\ell^\lambda = \lambda + \lambda^\alpha\mu^{1-\alpha}x$, the total staffing level satisfies

$$\sqrt{N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda)}(1 - \rho^\lambda) \to \beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}} \quad \text{as } \lambda \to \infty.$$

By Theorem 4.1 in Zeltyn and Mandelbaum (2005), we have

$$\begin{aligned}
&\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda\right) \\
&= \left[1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)}{H(-(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}))}\right]^{-1} \frac{1}{\sqrt{N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda)}}\sqrt{\frac{\gamma}{\mu}} \\
&\quad \left[H\left((\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}})\sqrt{\frac{\mu}{\gamma}}\right) - (\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}})\sqrt{\frac{\mu}{\gamma}}\right] + o\left(\frac{1}{\sqrt{N_1^\lambda + N_2^\lambda}}\right) \\
&= \sqrt{\frac{\mu}{\lambda}} \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right) - \left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)}{H\left(-\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\right)}} + o\left(\frac{1}{\sqrt{\lambda/\mu}}\right).
\end{aligned}$$

From the steady-state flow balance equation

$$\gamma\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] = \left(\lambda + \lambda^\alpha\mu^{1-\alpha}x\right)\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda\right),$$

we get that

$$\begin{aligned}
&\frac{1}{\sqrt{\lambda/\mu}}\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] \\
&\to \frac{\mu}{\gamma} \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\left(\beta_1 + \beta_2 - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right) - \left(\beta_1 + \beta_2 - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\left(\beta_1 + \beta_2 - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)}{H\left(-\left(\beta_1 + \beta_2 - x\mathbb{1}_{\{\alpha=1/2\}}\right)\right)}}, \quad \text{as } \lambda \to \infty,
\end{aligned}$$

and the statement follows. Q.E.D.

Based on Lemma 5, let β_1^* and $\beta_2^*(\beta_1, X)$ be the optimal solution to

$$\min_{\beta_1 \in \mathbb{R}}\left\{c_1\beta_1 + \mathbb{E}\left[\min_{\beta_2(\beta_1, X) \in \mathbb{R}_+} \hat{r}(\beta_1, \beta_2(\beta_1, X), X)\right]\right\}, \quad \text{for } \hat{z} \text{ defined in (39).} \tag{40}$$

Then, the two-stage uncertainty hedging rule, $u_{2,UH}$, is defined as $u_2(\beta_1, \beta_2(\beta_1, X))$ with parameters β_1^* and $\beta_2^*(\beta_1^*, X)$, i.e.,

$$N_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^{1/2} + o((\lambda/\mu)^{1/2}), \quad \text{and} \quad N_2^\lambda(N_1^\lambda, \Lambda^\lambda) = \beta_2^*(\beta_1^*, X)(\lambda/\mu)^{1/2} + o_{UI}((\lambda/\mu)^{1/2}).$$

Remark 2 *The existence of β_1^* and $\beta_2^*(\beta_1^*, X)$ follows from the same lines of analysis as those for the conventional single-stage square-root staffing rule considered in the literature (see, e.g., Garnett et al. (2002), Zeltyn and Mandelbaum (2005), Mandelbaum and Zeltyn (2009)). For completeness, we outline the key steps and omit the lengthy algebraic derivation. Given β_1 and $X = x$, it can be seen from* (39) *that $\hat{r}(\beta_1, \beta_2, x)$ is continuous in β_2. In addition, it can be checked that $\hat{r}(\beta_1, \beta_2, x) \to \infty$ as $\beta_2 \to \infty$. Thus, an optimal solution $\beta_2^*(\beta_1, x)$ exists for the inner minimization problem in* (40)*. The existence of β_1^* can be argued similarly. Let $g(\beta_1) := c_1\beta_1 + \mathbb{E}\left[\hat{r}\left(\beta_1, \beta_2^*(\beta_1, X), X\right)\right]$. It can be checked that $g(\beta_1) \to \infty$ as $\beta_1 \to \infty$. In addition, under the condition that $\mu > \gamma$ or $(h + a\gamma)\mu > c_1\gamma$ (this latter condition is implied by Assumption 1), we have $g(\beta_1) \to \infty$ as $\beta_1 \to -\infty$. The existence of an optimal solution β_1^* then follows from the continuity of $g(\beta_1)$ in β_1.*

Before we establish the asymptotic performance of $u_{2,UH}$, we first prove an auxiliary lemma.

Lemma 6 *Assume that $\alpha \le 1/2$. Under the two-stage uncertainty hedging rule defined in* (40)*, there exists a random variable $\tilde{X}$ such that $\beta_2^*(\beta_1, X) \le \tilde{X}$ and $\mathbb{E}[\tilde{X}] < \infty$.*

Proof: For any realized arrival rate $\ell^\lambda = \lambda + x\lambda^\alpha\mu^{1-\alpha}$, we start by rewriting (39) as

$$\begin{aligned}
&\hat{r}\left(\beta_1, \beta_2(\beta_1, x), x\right)\\
&:= c_2\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right) - c_2\left(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}\right) +\\
&\left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right) - \left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)}{H\left(-\left(\beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}\right)\right)}}.
\end{aligned}$$

Let $\tilde{\beta} := \beta_1 + \beta_2(\beta_1, x) - x\mathbb{1}_{\{\alpha=1/2\}}$, and denote

$$g(\tilde{\beta}) := \left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\tilde{\beta}\sqrt{\frac{\mu}{\gamma}}\right) - \tilde{\beta}\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\tilde{\beta}\sqrt{\frac{\mu}{\gamma}}\right)}{H\left(-\tilde{\beta}\right)}}.$$

It follows from Section 2.1 in the Online Appendix of Mandelbaum and Zeltyn (2009) that the function g monotonically decreases from infinity to 0.

Define

$$\tilde{\beta}^* := \underset{\tilde{\beta} \ge \beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}}{\arg\min} \quad c_2\tilde{\beta} + g(\tilde{\beta}). \tag{41}$$

Note that by construction, we have

$$\beta_2^*(\beta_1, x) = \tilde{\beta}^* - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}.$$

Corresponding to (41), let

$$\tilde{\beta}^\dagger := \underset{\tilde{\beta} \in \mathbb{R}}{\arg\min} \quad c_2\tilde{\beta} + g(\tilde{\beta}),$$

where unlike $\tilde{\beta}^*$, $\tilde{\beta}^\dagger$ is a global minimizer of the objective function over the real line. The existence of $\tilde{\beta}^\dagger$ follows from the same lines of arguments as in Remark 2.

We discuss the following cases:

Case 1: If $\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}} \le \beta^\dagger$, then $\tilde{\beta}^* = \beta^\dagger$, and

$$\beta_2^*(\beta_1, x) = \beta^\dagger - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}. \tag{42}$$

Case 2: If $\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}} > \beta^\dagger$, then let $\epsilon > 0$, and let $M \in \mathbb{R}$ be such that (i) $M > \epsilon/c_2$, and (ii) for all $x > M$, we have $0 \le g(x) < \epsilon$. There are two subcases:

Case 2(i): If $\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}} \le M$, then exactly one of the following two scenarios holds:

Case 2(i.a): $\tilde{\beta}^* \le M$, so that

$$\beta_2^*(\beta_1, x) \le M - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}. \tag{43}$$

Case 2(i.b): $\tilde{\beta}^* > M$. In this case, (41) can be rewritten as

$$\tilde{\beta}^* = \underset{\tilde{\beta} \ge M}{\arg\min} \quad c_2\tilde{\beta} + g(\tilde{\beta}).$$

Note that for all $y > 2M$, it follows from the definition of M that

$$c_2 M + g(M) < c_2 y + g(y). \tag{44}$$

Therefore, $\tilde{\beta}^* \le 2M$, and

$$\beta_2^*(\beta_1, x) \le 2M - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}. \tag{45}$$

Case 2(ii): If $\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}} > M$, then by definition of M, (44) holds for all $y > 2(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}})$. Hence, $\tilde{\beta}^* \le 2(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}})$, and

$$\beta_2^*(\beta_1, x) \le 2(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}) - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}} = \beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}. \tag{46}$$

In summary, by (42), (43), (45), and (46), we get that

$$\beta_2^*(\beta_1, x) \le |\beta^\dagger| + 2M + |\beta_1| + |x|. \tag{47}$$

Let $\tilde{X} := |\beta^\dagger| + 2M + |\beta_1| + |X|$. The statement follows from (47) and $\mathbb{E}[|X|] < \infty$. Q.E.D.

The following lemma establishes the asymptotic performance of $u_{2,UL}$.

Lemma 7 *Assume that $\alpha \le 1/2$. Under the two-stage uncertainty hedging rule defined in (40), we have*

$$\hat{\mathcal{C}}^\lambda \to c_1\beta_1^* + \mathbb{E}\left[\hat{r}\left(\beta_1^*, \beta_2^*(\beta_1^*, X), X\right)\right] \quad \text{as } \lambda \to \infty,$$

for $\hat{z}$ defined in (39).

Proof: It follows from Lemma 5 that

$$\hat{\mathcal{C}}^\lambda(\Lambda^\lambda) \to c_1\beta_1^* + \hat{r}\left(\beta_1^*, \beta_2^*(\beta_1^*, X), X\right) \quad w.p.1 \quad \text{as } \lambda \to \infty.$$

Hence, to prove the claim, it is sufficient to show

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\hat{\mathcal{C}}^\lambda(\Lambda^\lambda)\right] = \mathbb{E}\left[\lim_{\lambda\to\infty} \hat{\mathcal{C}}^\lambda(\Lambda^\lambda)\right] \tag{48}$$

To this end, we utilize the dominated convergence theorem.

We start by writing

$$
\begin{aligned}
\hat{\mathcal{C}}^{\lambda}(\Lambda^{\lambda}) =& c_1\beta_1^* + c_2\beta_2^*(\beta_1^*, X) + \frac{1}{\sqrt{\lambda/\mu}}\left(D_1^{\lambda} + D_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda})\right) \\
&+ \frac{1}{\sqrt{\lambda/\mu}}(h + a\gamma)\mathbb{E}\left[Q(N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda})|\Lambda^{\lambda}\right] \\
=& c_1\beta_1^* + c_2\beta_2^*(\beta_1^*, X) + \frac{1}{\sqrt{\lambda/\mu}}\left(D_1^{\lambda} + D_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda})\right) \\
&+ \frac{\Lambda^{\lambda}}{\sqrt{\lambda/\mu}}\,(h/\gamma + a)\,\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right),
\end{aligned}
\tag{49}
$$

where the last equality follows from

$$
\gamma\mathbb{E}\left[Q(N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda})|\Lambda^{\lambda}\right] = \Lambda^{\lambda}\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right).
$$

Recall that $\mathbb{P}(BL, m, \lambda)$ is the steady-state blocking probability for an $M/M/m/m$ queue with number of servers equal to m and arrival rate equal to λ. It follows from a simple coupling argument that

$$
\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) \le \mathbb{P}\left(BL, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right). \tag{50}
$$

Since the Erlang blocking probability is increasing in the offered load and $N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}) \ge 0$, we further have

$$
\mathbb{P}\left(BL, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) \le \mathbb{P}\left(BL, N_1^{\lambda}, \lambda + |X|\lambda^{\alpha}\mu^{1-\alpha}\right). \tag{51}
$$

In addition, recall that $L(m, \lambda)$ is the steady-state loss rate in an $M/M/m/m$ queue with number of servers equal to m and arrival rate equal to λ. In particular, $L(m, \lambda)$ satisfies $L(m, \lambda) = \lambda\mathbb{P}(BL, m, \lambda)$, and by Theorem 1 in Smith and Whitt (1981),

$$
L(N_1^{\lambda}, \lambda + |X|\lambda^{\alpha}\mu^{1-\alpha}) \le L(N_1^{\lambda} - 1, \lambda) + L(1, |X|\lambda^{\alpha}\mu^{1-\alpha}). \tag{52}
$$

Combining (50)–(52), we have

$$
\begin{aligned}
\Lambda^{\lambda}\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) &\le \Lambda^{\lambda}\mathbb{P}\left(BL, N_1^{\lambda}, \lambda + |X|\lambda^{\alpha}\mu^{1-\alpha}\right) \\
&\le \lambda\mathbb{P}\left(BL, N_1^{\lambda} - 1, \lambda\right) + |X|\lambda^{\alpha}\mu^{1-\alpha}\mathbb{P}\left(BL, 1, |X|\lambda^{\alpha}\mu^{1-\alpha}\right) \\
&\le \lambda\mathbb{P}\left(BL, N_1^{\lambda} - 1, \lambda\right) + |X|\lambda^{\alpha}\mu^{1-\alpha}.
\end{aligned}
\tag{53}
$$

Dividing both sides of (53) by $\sqrt{\lambda/\mu}$, we get that

$$
\frac{\Lambda^{\lambda}}{\sqrt{\lambda/\mu}}\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) \le \frac{\lambda}{\sqrt{\lambda/\mu}}\mathbb{P}\left(BL, N_1^{\lambda} - 1, \lambda\right) + |X|\frac{\lambda^{\alpha}\mu^{1-\alpha}}{\sqrt{\lambda/\mu}}, \tag{54}
$$

where the first term on the right-hand side of (54) is a constant. By equation (17) in Whitt (1984),

$$
\lim_{\lambda\to\infty} \frac{\lambda}{\sqrt{\lambda/\mu}}\,\mathbb{P}\left(BL, N_1^{\lambda} - 1, \lambda\right) = \mu\frac{\phi(\beta_1^*)}{\Phi(\beta_1^*)}. \tag{55}
$$

Furthermore,

$$
\lim_{\lambda\to\infty} |X|\frac{\lambda^{\alpha}\mu^{1-\alpha}}{\sqrt{\lambda/\mu}} = \begin{cases} \mu|X| & \text{if } \alpha = 1/2 \\ 0 & \text{if } \alpha < 1/2. \end{cases} \tag{56}
$$

By (54)–(56), we have for λ large enough,

$$
\frac{\Lambda^{\lambda}}{\sqrt{\lambda/\mu}}\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) \le \mu\frac{\phi(\beta_1^*)}{\Phi(\beta_1^*)} + \mu|X|.
$$

This, together with Lemma 6, the assumption that $\mathbb{E}[|X|] < \infty$, and the requirement on D_1^{λ} and $D_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda})$, implies that $|\hat{\mathcal{C}}^{\lambda}(\Lambda^{\lambda})|$ in (49) is uniformly bounded by an integrable random variable, and the interchange of limit and expectation in (48) is justified. Q.E.D.

C.3. Optimality Gap of $u_{2,UH}$

In Appendices C.1 and C.2, we propose the two-stage uncertainty hedging rule, which prescribes staffing levels

$$N_1^\lambda = \lambda/\mu + \beta_1^* (\lambda/\mu)^{\max\{\alpha,1/2\}} + o((\lambda/\mu)^{\max\{\alpha,1/2\}})$$
$$N_2^\lambda(N_1^\lambda, \Lambda^\lambda) = \beta_2^*(\beta_1^*, X)(\lambda/\mu)^{\max\{\alpha,1/2\}} + o_{UI}((\lambda/\mu)^{\max\{\alpha,1/2\}}).$$

When $\alpha > 1/2$, β_1^* and $\beta_2^*(\beta_1^*, X)$ are defined in (35), so that the capacity prescription is identical to that under the two-stage newsvendor solution. When $\alpha \leq 1/2$, β_1^* and $\beta_2^*(\beta_1^*, X)$ are defined in (40). Let $\mathcal{C}_{2,UH}^\lambda$ be the expected total cost defined under the two-stage uncertainty hedging rules. Recall that $\mathcal{C}_{2,*}^\lambda$ is the optimal total cost for the two-stage optimization problem (2). The next lemma quantifies the optimality gap of the proposed policy to the exact two-stage optimum.

Lemma 8 *For $\alpha \in (0,1)$, we have $\mathcal{C}_{2,UH}^\lambda - \mathcal{C}_{2,*}^\lambda = o(\lambda^{\max\{1/2,\alpha\}})$.*

PROOF: The key of the proof is to show that for any sequence of policies $u \in U$,

$$\liminf_{\lambda\to\infty} \hat{\mathcal{C}}_u^\lambda \geq \lim_{\lambda\to\infty} \hat{\mathcal{C}}_{2,UH}^\lambda. \tag{57}$$

Note that the limit on the right-hand side of (57) is well defined because of Lemma 4 and Lemma 7.

First, it is without loss of generality to consider a sequence of policies $u \in U$ under which

$$\liminf_{\lambda\to\infty} \frac{N_1^\lambda - \lambda/\mu}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} > -\infty. \tag{58}$$

To see this, for any sequence realized arrival rate ℓ^λ, recall from the proof of Proposition 1 that $B_1(N_1^\lambda, N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)$ and $B_2(N_1^\lambda, N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)$ denote the steady-state number of busy servers among the base and surge staff, respectively. It follows that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] &= c_2 N_2^\lambda(N_1^\lambda,\ell^\lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] \\
&\geq \frac{c_2}{\mu}\mu\mathbb{E}\left[B_2(N_1^\lambda, N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] + \left(\frac{h}{\gamma} + a\right)\gamma\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] \\
&\geq \min\left\{\frac{c_2}{\mu}, \frac{h}{\gamma} + a\right\}\left(\mu\mathbb{E}\left[B_2(N_1^\lambda, N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] + \gamma\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \\
&= \min\left\{\frac{c_2}{\mu}, \frac{h}{\gamma} + a\right\}\left(\ell^\lambda - \mu\mathbb{E}\left[B_1(N_1^\lambda, N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \\
&\geq \min\left\{\frac{c_2}{\mu}, \frac{h}{\gamma} + a\right\}\left(\ell^\lambda - \mu N_1^\lambda\right) \\
&= c_2\left(\frac{\ell^\lambda}{\mu} - N_1^\lambda\right).
\end{aligned}$$

Replacing ℓ^λ with Λ^λ, taking expectation, and recalling that $\mathbb{E}[X] = 0$ give

$$\mathbb{E}\left[\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right] \geq c_2\left(\frac{\lambda}{\mu} - N_1^\lambda\right).$$

Then, the scaled cost $\hat{\mathcal{C}}_u^\lambda$ satisfies

$$\begin{aligned}
\hat{\mathcal{C}}_u^\lambda &= c_1 \frac{N_1^\lambda - \lambda/\mu}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} + \frac{\mathbb{E}\left[\mathcal{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right]}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} \\
&\geq c_1 \frac{N_1^\lambda - \lambda/\mu}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} + c_2 \frac{\lambda/\mu - N_1^\lambda}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} \\
&= (c_2 - c_1)\frac{\lambda/\mu - N_1^\lambda}{(\lambda/\mu)^{\max\{1/2,\alpha\}}}.
\end{aligned} \tag{59}$$

If (58) does not hold, then it follows from (59) and Assumption 1 that $\liminf_{\lambda\to\infty}\hat{\mathcal{C}}_u^\lambda=\infty$. For the purpose of characterizing (near-)optimal staffing rules, we assume without loss of generality that $\liminf_{\lambda\to\infty}\hat{\mathcal{C}}_u^\lambda<\infty$.

Now, consider a subsequence of systems indexed by λ_i on which the lim inf in (57) is obtained, namely,

$$\lim_{\lambda_i\to\infty}\hat{\mathcal{C}}_u^{\lambda_i}=\liminf_{\lambda\to\infty}\hat{\mathcal{C}}_u^{\lambda}.$$

Along this subsequence,

$$\hat{\mathcal{C}}_u^{\lambda_i}=\frac{c_1\left(N_1^{\lambda_i}-\lambda_i/\mu\right)}{(\lambda_i/\mu)^{\max\{1/2,\alpha\}}}+\frac{\mathbb{E}\left[\mathcal{R}^{\lambda_i}(N_1^{\lambda_i},N_2^{\lambda_i}(N_1^{\lambda_i},\Lambda^{\lambda_i}),\Lambda^{\lambda_i})\right]}{(\lambda_i/\mu)^{\max\{1/2,\alpha\}}}.$$

Since the second term is non-negative, it must be the case that

$$\limsup_{\lambda_i\to\infty}\frac{c_1\left(N_1^{\lambda_i}-\lambda_i/\mu\right)}{(\lambda_i/\mu)^{\max\{1/2,\alpha\}}}<\infty.$$

Hence,

$$-\infty<\liminf_{\lambda_i\to\infty}\frac{N_1^{\lambda_i}-\lambda_i/\mu}{(\lambda_i/\mu)^{\max\{1/2,\alpha\}}}\le\limsup_{\lambda_i\to\infty}\frac{N_1^{\lambda_i}-\lambda_i/\mu}{(\lambda_i/\mu)^{\max\{1/2,\alpha\}}}<\infty.$$

Then, Bolzano-Weierstrass theorem indicates that any subsequence has a further convergent sub-subsequence indexed by λ_{i_j} along which

$$\frac{N_1^{\lambda_{i_j}}-\lambda_{i_j}/\mu}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\to\beta_1\in\mathbb{R}\quad\text{as }\lambda_{i_j}\to\infty. \tag{60}$$

It follows from (60) that

$$\begin{aligned}
\lim_{\lambda_{i_j}\to\infty}\hat{\mathcal{C}}_u^{\lambda_{i_j}}&\ge\lim_{\lambda_{i_j}\to\infty}\frac{c_1\left(N_1^{\lambda_{i_j}}-\lambda_{i_j}/\mu\right)}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}+\liminf_{\lambda_{i_j}\to\infty}\frac{\mathbb{E}\left[\mathcal{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\Lambda^{\lambda_{i_j}}),\Lambda^{\lambda_{i_j}})\right]}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\\
&=c_1\beta_1+\liminf_{\lambda_{i_j}\to\infty}\frac{\mathbb{E}\left[\mathcal{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\Lambda^{\lambda_{i_j}}),\Lambda^{\lambda_{i_j}})\right]}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\\
&\ge c_1\beta_1+\mathbb{E}\left[\liminf_{\lambda_{i_j}\to\infty}\frac{\mathcal{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\Lambda^{\lambda_{i_j}}),\Lambda^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\right],
\end{aligned}\tag{61}$$

where the last inequality follows from Fatou's lemma.

Next, we are going to establish that for any realized arrival rate $\ell^{\lambda_{i_j}}$,

$$\liminf_{\lambda_{i_j}\to\infty}\frac{\mathcal{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}}),\ell^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\ge\hat{r}\left(\beta_1,\beta_2^*(\beta_1,x),x\right). \tag{62}$$

In (62), when $\alpha>1/2$, $\hat{z}$ is defined in (31) and $\beta_2^*(\beta_1,X)$ is defined in (35). In the other case where $\alpha\le 1/2$, $\hat{z}$ is defined in (39) and $\beta_2^*(\beta_1,X)$ is defined in (40). To see that (62) holds, define

$$\hat{N}_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}}):=N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}})/\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}.$$

Observe that the sequence $\{\hat{N}_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}}):\lambda_{i_j}>0\}$ satisfies exactly one of the following three cases:

(i) $\hat{N}_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}})\to\beta_2\in\mathbb{R}_+$ as $\lambda_{i_j}\to\infty$.

(ii) $\hat{N}_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}})\to\infty$ as $\lambda_{i_j}\to\infty$.

(iii) $\hat{N}_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}})$ does not converge.

For case (i), (62) follows from Lemma 3, Lemma 5, and the definition of $\beta_2^*(\beta_1, x)$.

For case (ii), we have

$$\begin{aligned}
\frac{\mathcal{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, \ell^{\lambda_{i_j}}), \ell^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}} &= \frac{c_2 N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, \ell^{\lambda_{i_j}}) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^{\lambda_{i_j}} + N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, \ell^{\lambda_{i_j}}), \ell^{\lambda_{i_j}})\right]}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}} \\
&= c_2 \frac{N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, \ell^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}} + \frac{(h+a\gamma)\,\mathbb{E}\left[Q(N_1^{\lambda_{i_j}} + N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, \ell^{\lambda_{i_j}}), \ell^{\lambda_{i_j}})\right]}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}} \\
&\to \infty \quad \text{as } \lambda_{i_j} \to \infty,
\end{aligned}$$

and (62) holds.

For case (iii), we can further consider a further subsequence indexed by $\lambda_{i_{j_k}}$ along which $\hat{N}_2^{\lambda_{i_{j_k}}}(N_1^{\lambda_{i_{j_k}}}, \ell^{\lambda_{i_{j_k}}})$ converges. Such subsequence exists because a sequence has no convergent subsequence if and only if it approaches infinity. The same arguments for case (i) can be applied to establish (62).

Now, it follows from (61) and (62) that

$$\begin{aligned}
\lim_{\lambda_{i_j} \to \infty} \hat{\mathcal{C}}_u^{\lambda_{i_j}} &\geq c_1 \beta_1 + \mathbb{E}\left[\liminf_{\lambda_{i_j} \to \infty} \frac{\mathcal{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}}, \Lambda^{\lambda_{i_j}}), \Lambda^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\right] \\
&\geq c_1 \beta_1 + \mathbb{E}\left[\hat{r}\left(\beta_1, \beta_2^*(\beta_1, X), X\right)\right].
\end{aligned}$$

Furthermore, since β_1^* is constructed such that

$$c_1 \beta_1 + \mathbb{E}\left[\hat{r}\left(\beta_1, \beta_2^*(\beta_1, X), X\right)\right] \geq c_1 \beta_1^* + \mathbb{E}\left[\hat{r}\left(\beta_1^*, \beta_2^*(\beta_1^*, X), X\right)\right],$$

it follows that

$$\lim_{\lambda_{i_j} \to \infty} \hat{\mathcal{C}}_u^{\lambda_{i_j}} \geq c_1 \beta_1^* + \mathbb{E}\left[\hat{r}\left(\beta_1^*, \beta_2^*(\beta_1^*, X), X\right)\right] = \lim_{\lambda_{i_j} \to \infty} \hat{\mathcal{C}}_{2,UH}^{\lambda_{i_j}},$$

where the last equality follows from Lemma 4 and Lemma 7. Since the subsequence indexed by λ_{i_j} is arbitrary, we have established (57).

Next, we apply (57) to the sequence of exact optimal two-stage staffing rules, i.e., $u_{2,*}$, and get that

$$\liminf_{\lambda \to \infty} \hat{\mathcal{C}}_{2,*}^{\lambda} \geq \lim_{\lambda \to \infty} \hat{\mathcal{C}}_{2,UH}^{\lambda}.$$

By the optimality of $u_{2,*}$, we also have

$$\limsup_{\lambda \to \infty} \hat{\mathcal{C}}_{2,*}^{\lambda} \leq \lim_{\lambda \to \infty} \hat{\mathcal{C}}_{2,UH}^{\lambda}.$$

Thus,

$$\lim_{\lambda \to \infty} \hat{\mathcal{C}}_{2,*}^{\lambda} = \lim_{\lambda \to \infty} \hat{\mathcal{C}}_{2,UH}^{\lambda}. \tag{63}$$

The statement follows from (63). Q.E.D.

The following corollary is a direct consequence from the proof of Lemma 8.

Corollary 1 *For $\alpha \in (0,1)$, let β_1^* and $\beta_2^*(\beta_1^*, X)$ be defined in (35) when $\alpha > 1/2$, and defined in (40) when $\alpha \leq 1/2$. Consider a sequence of staffing policies $u = \{\pi^\lambda : \lambda > 0\} = \{N_1^\lambda, N_2^\lambda(N_1^\lambda, \Lambda^\lambda) : \lambda > 0\}$. If there does not exist a subsequence indexed by λ_i along which $\{N_1^{\lambda_i}, N_2^{\lambda_i}(N_1^{\lambda_i}, \Lambda^{\lambda_i}) : \lambda_i > 0\}$ is prescribed as*

$$N_1^{\lambda_i} = \lambda_i/\mu + \beta_1^* \left(\lambda_i/\mu\right)^{\max\{\alpha,1/2\}} + o((\lambda_i/\mu)^{\max\{\alpha,1/2\}})$$

$$N_2^{\lambda_i}(N_1^{\lambda_i}, \Lambda^{\lambda_i}) = \beta_2^*(\beta_1^*, X)\left(\lambda_i/\mu\right)^{\max\{\alpha,1/2\}} + o_{UI}((\lambda_i/\mu)^{\max\{\alpha,1/2\}}),$$

then $\mathcal{C}_u^\lambda - \mathcal{C}_{2,UH}^\lambda \geq \Theta(\lambda^{\max\{\alpha,1/2\}})$.

Corollary 1 indicates that it is without loss of optimality to consider the family of two-stage uncertainty hedging rule. To improve upon the $o(\lambda^{\max\{\alpha,1/2\}})$ optimality gap established in Lemma 8, we need to consider refinement which puts further restrictions on the $o((\lambda_i/\mu)^{\max\{\alpha,1/2\}})$ term in N_1^λ and the $o_{UI}((\lambda_i/\mu)^{\max\{\alpha,1/2\}})$ term in $N_2^\lambda(N_1^\lambda,\Lambda^\lambda)$. In the special case when $\alpha>1/2$, it is without loss of optimality to consider the family of two-stage newsvendor solutions. The two-stage QED rule is a refinement of the two-stage newsvendor solution that reduces the optimality gap from $o(\lambda^\alpha)$ to $o(\sqrt{\lambda})$.

C.4. Proof of Theorem 2

Proof: Note that the two-stage uncertainty hedging rule when $\alpha>1/2$ is equivalent to the two-stage newsvendor solution. The statement follows from Lemma 8. Q.E.D.

Appendix D: Proof of Theorem 1

The proof of Theorem 1 builds on the performance quantification of $u_2(\beta_1,\beta_2(\beta_1,X))$ and $u_{2,UH}$ introduced in Appendix C. For the sequence of systems indexed by λ, recall that $\mathcal{C}_{1,*}^\lambda$ is the optimal total cost for the single-stage optimization problem (4), and $\mathcal{C}_{2,*}^\lambda$ is the optimal total cost for the two-stage optimization problem (2). We establish Theorem 1 for different values of α.

D.1. Benefit of Surge Staffing When $\alpha<1/2$

Lemma 9 *If $\alpha<1/2$, then $\mathcal{C}_{1,*}^\lambda-\mathcal{C}_{2,*}^\lambda=o(\sqrt{\lambda})$.*

Proof: We start by determining the parameters β_1^* and $\beta_2^*(\beta_1^*,X)$ defined in (40) for the two-stage uncertainty hedging rule when $\alpha<1/2$. In particular, for any realization x of the random variable X, the function $\hat{z}$ in (39) becomes

$$\hat{r}\left(\beta_1,\beta_2(\beta_1,x),x\right)=c_2\beta_2(\beta_1,x)+\left(\frac{h\mu}{\gamma}+a\mu\right)\frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left((\beta_1+\beta_2(\beta_1,x))\sqrt{\frac{\mu}{\gamma}}\right)-(\beta_1+\beta_2(\beta_1,x))\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left((\beta_1+\beta_2(\beta_1,x))\sqrt{\frac{\mu}{\gamma}}\right)}{H(-(\beta_1+\beta_2(\beta_1,x)))}}.$$

Note that $\hat{r}\left(\beta_1,\beta_2(\beta_1,x),x\right)$ does not depend on the realization x. Hence, given β_1, we have that $\beta_2^*(\beta_1,x)=\arg\min_{\beta_2\in\mathbb{R}_+}\hat{r}(\beta_1,\beta_2(\beta_1,x),x)$ does not depend on x either. Then β_1^* and $\beta_2^*(\beta_1^*,x)$ jointly solve

$$\min_{\beta_1\in\mathbb{R},\,\beta_2(\beta_1,x)\in\mathbb{R}_+}\quad c_1\beta_1+\hat{r}\left(\beta_1,\beta_2(\beta_1,x),x\right).$$

By the assumption that $c_1<c_2$. Thus, it is optimal to set

$$\beta_1^*:=\arg\min_{\beta_1\in\mathbb{R}}\ c_1\beta_1+\left(\frac{h\mu}{\gamma}+a\mu\right)\frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\beta_1\sqrt{\frac{\mu}{\gamma}}\right)-\beta_1\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left(\beta_1\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\beta_1)}},$$

and $\beta_2^*\left(\beta_1^*,x\right):=0$ for all realizations x of the random variable X.

In this case, the two-stage uncertainty hedging rule is equivalent to the conventional single-stage square-root staffing rule (with staffing cost c_1, holding cost h, and abandonment cost a). Then,

$$\mathcal{C}_{2,UH}^\lambda-\mathcal{C}_{1,*}^\lambda\geq 0\quad\text{for all }\lambda>0. \tag{64}$$

In addition, we establish in Lemma 8 that

$$\mathcal{C}_{2,UH}^\lambda-\mathcal{C}_{2,*}^\lambda=o(\sqrt{\lambda}). \tag{65}$$

The statement follows from (64) and (65). Q.E.D.

D.2. Benefit of Surge Staffing When $\alpha = 1/2$

Lemma 10 *If $\alpha = 1/2$, then $\mathcal{C}^\lambda_{1,*} - \mathcal{C}^\lambda_{2,*} = O(\sqrt{\lambda})$.*

Proof: Consider $\beta_2^\dagger(\beta_1, X) := 0$ for all β_1, and $\beta_1^\dagger := \arg\min_{\beta_1 \in \mathbb{R}} \; c_1\beta_1 + \mathbb{E}\left[\hat{r}\left(\beta_1, \beta_2^\dagger(\beta_1, X), X\right)\right]$. Note that $\beta_1^\dagger$ and $\beta_2^\dagger(\beta_1, X)$ provide a feasible pair of parameters for $u_2(\beta_1, \beta_2(\beta_1, X))$. Let $\mathcal{C}^\lambda_{2,\dagger}$ denote the expected total cost under $u_2(\beta_1^\dagger, \beta_2^\dagger(\beta_1^\dagger, X))$. It follows from similar derivation as in the proof of Lemma 7 that

$$\lim_{\lambda\to\infty} \hat{\mathcal{C}}^\lambda_{2,\dagger} = c_1\beta_1^\dagger + \mathbb{E}\left[\hat{r}\left(\beta_1^\dagger, \beta_2^\dagger(\beta_1^\dagger, X), X\right)\right].$$

Since $(\beta_1^\dagger, \beta_2^\dagger(\beta_1^\dagger, x))$ is not necessarily optimal for the optimization problem in (40), we have

$$c_1\beta_1^\dagger + \mathbb{E}\left[\hat{r}\left(\beta_1^\dagger, \beta_2^\dagger(\beta_1^\dagger, X), X\right)\right] \geq c_1\beta_1^* + \mathbb{E}\left[\hat{r}\left(\beta_1^*, \beta_2^*(\beta_1^*, X), X\right)\right].$$

It then follows from Lemma 7 that

$$\mathcal{C}^\lambda_{2,\dagger} - \mathcal{C}^\lambda_{2,UH} = O(\sqrt{\lambda}). \tag{66}$$

Moreover, since $\beta_2^\dagger(\beta_1^\dagger, X) = 0$, this policy is equivalent to a single-stage staffing rule. By Proposition 3 in Bassamboo et al. (2010), we get that

$$\mathcal{C}^\lambda_{2,\dagger} - \mathcal{C}^\lambda_{1,*} = O(\sqrt{\lambda}). \tag{67}$$

Lastly, by Lemma 8, we have

$$\mathcal{C}^\lambda_{2,UH} - \mathcal{C}^\lambda_{2,*} = o(\sqrt{\lambda}). \tag{68}$$

The statement follows from (66)–(68). Q.E.D.

Figure 8 below illustrates the performance gap between the employed policies in the proof of Lemma 10.

Figure 8 **Cost saving for $\alpha = 1/2$**

$$\begin{tikzcd}
\mathcal{C}^\lambda_{2,\dagger} \arrow[r, "O(\sqrt{\lambda})"] \arrow[d, "O(\sqrt{\lambda})"'] & \mathcal{C}^\lambda_{2,UH} \arrow[d, "o(\sqrt{\lambda})"] \\
\mathcal{C}^\lambda_{1,*} \arrow[r, "O(\sqrt{\lambda})"'] & \mathcal{C}^\lambda_{2,*}
\end{tikzcd}$$

D.3. Benefit of Surge Staffing When $\alpha > 1/2$

Lemma 11 *If $\alpha > 1/2$, then $\mathcal{C}^\lambda_{1,*} - \mathcal{C}^\lambda_{2,*} = \Theta(\lambda^\alpha)$.*

Proof: Under the two-stage newsvendor solution, the base-stage staffing level is $\lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + o\left((\lambda/\mu)^\alpha\right)$, where β_1^* is given by

$$\beta_1^* = \arg\min_{\beta_1 \in \mathbb{R}} \; c_1\beta_1 + c_2\mathbb{E}\left[(X - \beta_1)^+\right].$$

Moreover, Lemma 4 establishes that

$$\hat{\mathcal{C}}^\lambda_{2,NV} \to c_1\beta_1^* + c_2\mathbb{E}\left[(X - \beta_1^*)^+\right] \quad \text{as } \lambda \to \infty.$$

In comparison, under the single-stage newsvendor solution, the base-stage staffing level is $\lambda/\mu+\beta_{NV}(\lambda/\mu)^{\alpha}$, where β_{NV} is given by

$$\beta_{NV}=\operatorname*{arg\,min}_{\beta\in\mathbb{R}} c_1\beta+\left(\frac{h\mu}{\gamma}+a\mu\right)\mathbb{E}\left[(X-\beta)^{+}\right].$$

Similar lines of arguments as in the proof of Lemma 4 show that

$$\hat{\mathcal{C}}^{\lambda}_{1,NV}\to c_1\beta_{NV}+\left(\frac{h\mu}{\gamma}+a\mu\right)\mathbb{E}\left[(X-\beta_{NV})^{+}\right]\quad \text{as } \lambda\to\infty.$$

Therefore, if

$$\operatorname*{arg\,min}_{\beta\in\mathbb{R}} c_1\beta+\left(\frac{h\mu}{\gamma}+a\mu\right)\mathbb{E}\left[(X-\beta)^{+}\right]>\operatorname*{arg\,min}_{\beta\in\mathbb{R}} c_1\beta+c_2\mathbb{E}\left[(X-\beta)^{+}\right], \tag{69}$$

then

$$\lim_{\lambda\to\infty}\hat{\mathcal{C}}^{\lambda}_{1,NV}>\lim_{\lambda\to\infty}\hat{\mathcal{C}}^{\lambda}_{2,NV},$$

so that

$$\mathcal{C}^{\lambda}_{1,NV}-\mathcal{C}^{\lambda}_{2,NV}=\Theta(\lambda^{\alpha}). \tag{70}$$

Note that a sufficient condition for (69) to hold is that X is a continuous random variable, i.e., with a proper density function.

Moreover, by Theorem 1 in Bassamboo et al. (2010), we get that

$$\mathcal{C}^{\lambda}_{1,NV}-\mathcal{C}^{\lambda}_{1,*}=O(\lambda^{1-\alpha})=o(\sqrt{\lambda}). \tag{71}$$

By Lemma 8, we also have

$$\mathcal{C}^{\lambda}_{2,NV}-\mathcal{C}^{\lambda}_{2,*}=o(\lambda^{\alpha}). \tag{72}$$

The statement follows from (70)–(72). Q.E.D.

Figure 9 below illustrates the performance gap between the employed policies in the proof of Lemma 11.

Figure 9 Cost saving for $\alpha>1/2$

$$\begin{tikzcd}
\mathcal{C}^{\lambda}_{1,NV} \arrow[r, "\Theta(\lambda^{\alpha})"] \arrow[d, "o(\sqrt{\lambda})"'] & \mathcal{C}^{\lambda}_{2,NV} \arrow[d, "o(\lambda^{\alpha})"] \\
\mathcal{C}^{\lambda}_{1,*} \arrow[r, "\Theta(\lambda^{\alpha})"'] & \mathcal{C}^{\lambda}_{2,*}
\end{tikzcd}$$

Theorem 1 follows from Lemmas 9–11.

Appendix E: Proof of Theorem 3

Before we prove Theorem 3, we first prove an important auxiliary result on the asymptotic equivalence of the family of two-stage newsvendor solutions, and then establish the asymptotic performance of the family of two-stage QED rules. We assume throughout this section that $\alpha > 1/2$.

Recall that the two-stage newsvendor policy takes the form

$$N_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + D_1^\lambda, \quad N_2^\lambda(N_1^\lambda, \Lambda^\lambda) = \beta_2^*(\beta_1^*, X)(\lambda/\mu)^\alpha + D_2^\lambda(N_1^\lambda, \Lambda^\lambda), \tag{73}$$

for $D_1^\lambda = o\left((\lambda/\mu)^\alpha\right)$, and $D_2^\lambda(N_1^\lambda, \Lambda^\lambda) = o_{UI}\left((\lambda/\mu)^\alpha\right)$. Let u be a policy of the form (73). Based on u, we can construct another policy $\tilde{u}$, where

$$\tilde{N}_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + \tilde{D}_1^\lambda, \quad \text{and} \quad \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda) = \beta_2^*(\beta_1^*, X)(\lambda/\mu)^\alpha + \tilde{D}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda),$$

for

$$\tilde{D}_1^\lambda := 0, \quad \text{and} \quad \tilde{D}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda) := \begin{cases} D_2^\lambda(N_1^\lambda, \Lambda^\lambda) & \text{if } X < \beta_1^* \\ D_1^\lambda + D_2^\lambda(N_1^\lambda, \Lambda^\lambda) & \text{if } X \ge \beta_1^*. \end{cases}$$

Let $\mathcal{C}_u^\lambda$ and $\mathcal{C}_{\tilde{u}}^\lambda$ denote the expected total cost under u and $\tilde{u}$, respectively.

Lemma 12 *If $\mathcal{C}_u^\lambda < \mathcal{C}_{\tilde{u}}^\lambda$, then $\mathcal{C}_{\tilde{u}}^\lambda - \mathcal{C}_u^\lambda = o(\sqrt{\lambda})$.*

Proof: Let $\mathcal{S}_u^\lambda$ and $\mathcal{S}_{\tilde{u}}^\lambda$ denote the expected staffing cost under u and $\tilde{u}$, respectively. By construction, u and $\tilde{u}$ have the same expected staffing cost, namely,

$$\begin{aligned}
\mathcal{S}_u^\lambda &= c_1(\lambda/\mu) + c_1\beta_1^*(\lambda/\mu)^\alpha + c_1 D_1^\lambda + \mathbb{E}\left[c_2\beta_2^*(\beta_1^*, X) + c_2 D_2^\lambda(N_1^\lambda, \Lambda^\lambda)\right] \\
&= c_1(\lambda/\mu) + c_1\beta_1^*(\lambda/\mu)^\alpha + c_2\frac{c_1}{c_2} D_1^\lambda + \mathbb{E}\left[c_2\beta_2^*(\beta_1^*, X) + c_2 D_2^\lambda(N_1^\lambda, \Lambda^\lambda)\right] \\
&= c_1(\lambda/\mu) + c_1\beta_1^*(\lambda/\mu)^\alpha + c_2 D_1^\lambda \mathbb{P}\left(X \ge \beta_1^*\right) + \mathbb{E}\left[c_2\beta_2^*(\beta_1^*, X) + c_2 D_2^\lambda(N_1^\lambda, \Lambda^\lambda)\right] \\
&= \mathcal{S}_{\tilde{u}}^\lambda,
\end{aligned}$$

where the second to last equality follows from $\beta_1^* = \bar{F}_X^{-1}\left(c_1/c_2\right)$ and the assumption that X is a continuous random variable.

We next consider queue length. If $D_1^\lambda < 0$, then by construction of $\tilde{u}$, $\tilde{u}$ prescribes a higher staffing level than u when $X < \beta_1^*$, and prescribes the same staffing level as u when $X \ge \beta_1^*$. Thus,

$$\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\right] \ge \mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\right],$$

and $\mathcal{C}_u^\lambda \ge \mathcal{C}_{\tilde{u}}^\lambda$.

Therefore, it is without loss of generality to assume that $D_1^\lambda \ge 0$ for all $\lambda > 0$. We again divide the discussion into two cases: $X \ge \beta_1^*$ and $X < \beta_1^*$. If the realized random variable satisfies $x \ge \beta_1^*$, then

$$\tilde{D}_1^\lambda + \tilde{D}_2^\lambda(\tilde{N}_1^\lambda, \ell^\lambda) = D_1^\lambda + D_2^\lambda(N_1^\lambda, \ell^\lambda),$$

where $\ell^\lambda = \lambda + x\lambda^\alpha\mu^{1-\alpha}$. This implies that

$$\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X \ge \beta_1^*\}}\right] = \mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X \ge \beta_1^*\}}\right]. \tag{74}$$

In the other case where $X < \beta_1^*$, it follows from (32) in the proof of Lemma 3 that

$$\begin{aligned}&\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}|\Lambda^\lambda\right] \\ &= \lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}|\Lambda^\lambda\right] = 0.\end{aligned} \tag{75}$$

The above equality and subsequent inequalities involving random variables hold in a path-by-path sense. Furthermore, recall from Lemma 1 that

$$\begin{aligned}\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)|\Lambda^\lambda\right] &\le \max\{\mu/\gamma, 1\}\left(\left(\Lambda^\lambda/\mu - N_1^\lambda\right)^+ + \sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right) \\ &\le \max\{\mu/\gamma, 1\}\left(\sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right),\end{aligned}$$

where the second inequality follows because $D_1^\lambda \ge 0$. Thus, there exists a random variable Y with $\mathbb{E}[Y] < \infty$ such that

$$\frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}|\Lambda^\lambda\right] \le Y, \quad \text{for all } \lambda > 0.$$

Moreover, the same derivation applies to $\tilde{u}$. Thus, we can apply the dominated convergence theorem to (75) and get that

$$\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] = \frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] = 0. \tag{76}$$

Now, we write $\mathcal{C}_u^\lambda$ as

$$\begin{aligned}\mathcal{C}_u^\lambda &= \mathcal{S}_u^\lambda + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\right] \\ &= \mathcal{S}_u^\lambda + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X\ge\beta_1^*\}}\right].\end{aligned} \tag{77}$$

In addition, we write $\mathcal{C}_{\tilde{u}}^\lambda$ as

$$\begin{aligned}\mathcal{C}_{\tilde{u}}^\lambda &= \mathcal{S}_{\tilde{u}}^\lambda + (h + a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\right] \\ &= \mathcal{S}_{\tilde{u}}^\lambda + (h + a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] + (h + a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X\ge\beta_1^*\}}\right].\end{aligned} \tag{78}$$

Then,

$$\begin{aligned}\mathcal{C}_u^\lambda - \mathcal{C}_{\tilde{u}}^\lambda &= (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] - (h + a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] \\ &= o(\sqrt{\lambda}),\end{aligned}$$

where the first equality follows from (74), (77) and (78), and the second equality follows from (76). Q.E.D.

Recall from Section 4.2 that $u_{2,QED}$ takes the form

$$N_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + O(\sqrt{\lambda/\mu}), \quad \text{and} \quad N_2^\lambda(N_1^\lambda, \Lambda^\lambda) = (\Lambda^\lambda/\mu + \eta^*\sqrt{\Lambda^\lambda/\mu} - N_1^\lambda)^+ + o_{UI}(\sqrt{\lambda/\mu}).$$

For a sequence of policies $u \in U$, let

$$\bar{\mathcal{C}}_u^\lambda := \frac{1}{(\lambda/\mu)^{1/2}}\left(\mathcal{C}_u^\lambda - c_1\frac{\lambda}{\mu} - c_1\beta_1^*\left(\frac{\lambda}{\mu}\right)^\alpha - c_2\mathbb{E}\left[(X - \beta_1^*)^+\right]\left(\frac{\lambda}{\mu}\right)^\alpha\right). \tag{79}$$

In addition, define the mapping $\psi : \mathbb{R} \to \mathbb{R}$ as

$$\psi(x) := \begin{cases} 0 & \text{if } x < \beta_1^* \\ c_2\eta^* + \left(\frac{h\mu}{\gamma} + a\mu\right)\dfrac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta^*\sqrt{\frac{\mu}{\gamma}}\right) - \eta^*\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta^*\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta^*)}} & \text{if } x \ge \beta_1^*. \end{cases} \tag{80}$$

Lemma 13 *We have*

$$\lim_{\lambda\to\infty} \bar{\mathcal{C}}^{\lambda}_{2,QED} = \mathbb{E}\left[\psi(X)\right].$$

Proof: Consider an arbitrary two-stage QED policy u of the form

$$N_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + D_1^\lambda, \quad \text{and} \quad N_2^\lambda(N_1^\lambda,\Lambda^\lambda) = (\Lambda^\lambda/\mu + \eta^*\sqrt{\Lambda^\lambda/\mu} - N_1^\lambda)^+ + J(N_1^\lambda,\Lambda^\lambda),$$

for $D_1^\lambda \in \mathbb{R}$, $D_1^\lambda = O(\sqrt{\lambda/\mu})$, and $J(N_1^\lambda,\Lambda^\lambda) = o_{UI}(\sqrt{\lambda/\mu})$.

For base staffing level, it holds that

$$c_1\left(N_1^\lambda - \lambda/\mu - \beta_1^*(\lambda/\mu)^\alpha - D_1^\lambda\right) = 0.$$

For surge staffing level, we have

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{\lambda/\mu}} c_2\left(N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha + D_1^\lambda \mathbb{1}_{\{X>\beta_1^*\}}\right) = \bar{n}(X), \tag{81}$$

where

$$\bar{n}(X) := \begin{cases} 0 & \text{if } X<\beta_1^* \\ c_2\eta^* & \text{if } X>\beta_1^*. \end{cases}$$

We next show that

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\frac{1}{\sqrt{\lambda/\mu}} c_2\left(N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha + D_1^\lambda \mathbb{1}_{\{X>\beta_1^*\}}\right)\right] = \mathbb{E}\left[\bar{n}(X)\right]. \tag{82}$$

To see (82), note that when $X<\beta_1^*$,

$$\begin{aligned} |N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+(\lambda/\mu)^\alpha + D_1^\lambda\mathbb{1}_{\{X>\beta_1^*\}}| &= |(\Lambda^\lambda/\mu + \eta^*\sqrt{\Lambda^\lambda/\mu} - N_1^\lambda)^+ + J(N_1^\lambda,\Lambda^\lambda)| \\ &= \left|\left((X-\beta_1^*)(\lambda/\mu)^\alpha + \eta^*\sqrt{\Lambda^\lambda/\mu} - D_1^\lambda\right)^+ + J(N_1^\lambda,\Lambda^\lambda)\right| \\ &\le |\eta^*|\sqrt{\Lambda^\lambda/\mu} + |D_1^\lambda| + |J(N_1^\lambda,\Lambda^\lambda)|. \end{aligned}$$

When $X>\beta_1^*$,

$$\begin{aligned} &|N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+(\lambda/\mu)^\alpha + D_1^\lambda\mathbb{1}_{\{X>\beta_1^*\}}| \\ &= |(\Lambda^\lambda/\mu + \eta^*\sqrt{\Lambda^\lambda/\mu} - N_1^\lambda)^+ + J(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+(\lambda/\mu)^\alpha + D_1^\lambda| \\ &= \left|\left((X-\beta_1^*)(\lambda/\mu)^\alpha + \eta^*\sqrt{\Lambda^\lambda/\mu} - D_1^\lambda\right)^+ + J(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+(\lambda/\mu)^\alpha + D_1^\lambda\right| \\ &= \begin{cases} |\eta^*\sqrt{\Lambda^\lambda/\mu} - D_1^\lambda + J(N_1^\lambda,\Lambda^\lambda) + D_1^\lambda| & \text{if } (X-\beta_1^*)(\lambda/\mu)^\alpha \ge -\eta^*\sqrt{\Lambda^\lambda/\mu} + D_1^\lambda \\ |J(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+(\lambda/\mu)^\alpha + D_1^\lambda| & \text{if } (X-\beta_1^*)(\lambda/\mu)^\alpha < -\eta^*\sqrt{\Lambda^\lambda/\mu} + D_1^\lambda \end{cases} \\ &\le |\eta^*|\sqrt{\Lambda^\lambda/\mu} + 2|D_1^\lambda| + |J(N_1^\lambda,\Lambda^\lambda)|. \end{aligned}$$

Thus, in both cases, there exists some random variable Y with $\mathbb{E}[Y]<\infty$ such that

$$\left|\frac{1}{\sqrt{\lambda/\mu}}\left(N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha\right) + D_1^\lambda\mathbb{1}_{\{X>\beta_1^*\}}\right| < Y, \quad \text{for all } \lambda>0.$$

The first equality in (82) can then be justified by (81) and the dominated convergence theorem.

For queue length, it follows from (32) in the proof of Lemma 3 (for the case where $X<\beta_1^*$), and the same analysis as in the proof of Lemma 6 (for the case where $X>\beta_1^*$) that

$$\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}}(h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right] = \bar{q}(X), \tag{83}$$

where

$$\bar{q}(X) := \begin{cases} 0 & \text{if } X < \beta_1^* \\ \left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta^*\sqrt{\frac{\mu}{\gamma}}\right) - \eta^*\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta^*\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta^*)}} & \text{if } X > \beta_1^*. \end{cases}$$

We next show that

$$\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right] = \mathbb{E}\left[\bar{q}(X)\right]. \tag{84}$$

To see (84), it follows from Lemma 1 that

$$\begin{aligned} \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right] &\le \max\{\mu/\gamma,1\}\left(\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda,\Lambda^\lambda)\right)^+ + \sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right) \\ &\le \begin{cases} \max\{\mu/\gamma,1\}\left(|D_1^\lambda| + \sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right) & \text{if } X < \beta_1^* \\ \max\{\mu/\gamma,1\}\left(|J(N_1^\lambda,\Lambda^\lambda)| + \sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda} + 1/\log 2\right) & \text{if } X > \beta_1^*. \end{cases} \end{aligned}$$

Thus, there exists some random variable Y with $\mathbb{E}[Y] < \infty$ such that

$$\frac{1}{(\lambda/\mu)^{1/2}} (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right] < Y, \quad \text{for all } \lambda > 0.$$

The first equality in (84) is justified by (83) and the dominated convergence theorem.

Then, for $\bar{\mathcal{C}}_u^\lambda$ defined in (79) and ψ defined in (80),

$$\begin{aligned} \bar{\mathcal{C}}_u^\lambda &= \frac{1}{(\lambda/\mu)^{1/2}}\Bigg(c_1 N_1^\lambda + c_2\mathbb{E}\left[N_2^\lambda(N_1^\lambda,\Lambda^\lambda)\right] + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right] \\ &\quad - c_1\frac{\lambda}{\mu} - c_1\beta_1^*\left(\frac{\lambda}{\mu}\right)^\alpha - c_2\mathbb{E}\left[(X-\beta_1^*)^+\right]\left(\frac{\lambda}{\mu}\right)^\alpha\Bigg) \\ &= \frac{1}{(\lambda/\mu)^{1/2}}\Bigg(c_1\left(N_1^\lambda - \frac{\lambda}{\mu} - \beta_1^*\left(\frac{\lambda}{\mu}\right)^\alpha - D_1^\lambda\right) + c_2\mathbb{E}\left[N_2^\lambda(N_1^\lambda,\Lambda^\lambda) - (X-\beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha + D_1^\lambda \mathbb{1}_{\{X>\beta_1^*\}}\right] \\ &\quad + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right]\Bigg) \\ &= \mathbb{E}\left[\psi(X)\right], \end{aligned}$$

from which the statement follows. Q.E.D.

We now present the proof of Theorem 3.

Proof: [Proof of Theorem 3] It follows from (57) in the proof of Lemma 8 that for all $u \in U$,

$$\liminf_{\lambda\to\infty} \hat{\mathcal{C}}_u^\lambda \ge \lim_{\lambda\to\infty} \hat{\mathcal{C}}_{2,NV}^\lambda = c_1\beta_1^* + c_2\mathbb{E}\left[(X-\beta_1^*)^+\right],$$

where $\beta_1^* = \bar{F}_X^{-1}(c_1/c_2)$. Thus, for a sequence of policies $u \in U$, we consider $\bar{\mathcal{C}}_u^\lambda$ defined in (79). We next show that for all $u \in U$,

$$\liminf_{\lambda\to\infty} \bar{\mathcal{C}}_u^\lambda \ge \lim_{\lambda\to\infty} \bar{\mathcal{C}}_{2,QED}^\lambda, \tag{85}$$

where the limit on the right-hand side of (85) is rigorously established in Lemma 13. Similar to the proof of Lemma 8, for the purpose of characterizing (near-)optimal staffing rules, we assume without loss of generality that $\limsup_{\lambda\to\infty} \bar{\mathcal{C}}_u^\lambda < \infty$.

First, by Corollary 1, it is without loss of optimality to consider a sequence of policies u of the form

$$N_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + D_1^\lambda, \quad N_2^\lambda = (X-\beta_1^*)^+(\lambda/\mu)^\alpha + D_2^\lambda(N_1^\lambda,\Lambda^\lambda),$$

for $D_1^\lambda = o\left((\lambda/\mu)^\alpha\right)$ and $D_2^\lambda(N_1^\lambda,\Lambda^\lambda) = o_{UI}\left((\lambda/\mu)^\alpha\right)$, i.e., the two-stage newsvendor solutions.

In addition, Lemma 12 implies that it is without loss of generality to consider a sequence of policies where $D_1^\lambda = 0$ for all $\lambda > 0$. Thus, we can write

$$\begin{aligned}\bar{\mathcal{C}}_u^\lambda &= \frac{1}{(\lambda/\mu)^{1/2}}\mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right] - c_2(X-\beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha\right] \\ &= \frac{1}{(\lambda/\mu)^{1/2}}\mathbb{E}\left[c_2 D_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right]\right].\end{aligned}$$

By Fatou's lemma,

$$\liminf_{\lambda\to\infty}\bar{\mathcal{C}}_u^\lambda \ge \mathbb{E}\left[\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2 D_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right]\right)\right]. \tag{86}$$

We are going to establish that for any realized arrival rate $\ell^\lambda = \lambda + x\lambda^\alpha\mu^{1-\alpha}$,

$$\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2 D_2^\lambda(N_1^\lambda,\ell^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \ge \psi(x), \tag{87}$$

where ψ is defined in (80). To this end, define

$$\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda) := \frac{1}{(\lambda/\mu)^{1/2}} D_2^\lambda(N_1^\lambda,\ell^\lambda).$$

Observe that the sequence $\left\{\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda) : \lambda > 0\right\}$ satisfies exactly one of the following four cases:

(i) $\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda) \to \infty$ as $\lambda\to\infty$.

(ii) $\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda) \to -\infty$ as $\lambda\to\infty$.

(iii) $\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda) \to \eta \in \mathbb{R}$ as $\lambda\to\infty$.

(iv) $\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda)$ does not converge.

For case (i), since $\mathbb{E}\left[Q^\lambda(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] \ge 0$,

$$\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2 D_2^\lambda(N_1^\lambda,\ell^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \ge \liminf_{\lambda\to\infty} c_2\bar{D}_2^\lambda(N_1^\lambda,\ell^\lambda) = \infty.$$

For case (ii), this case is only possible when $x > \beta_1^*$. This is because otherwise, $\beta_2^* = 0$, so that $D_2^\lambda \ge 0$ for all $\lambda > 0$. Now since $x > \beta_1^*$, we have

$$\begin{aligned}&(h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] \\ =&\left(\frac{h}{\gamma}+a\right)\gamma\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] \\ =&\left(\frac{h}{\gamma}+a\right)\left(\ell^\lambda - \mu\mathbb{E}\left[B_2(N_1^\lambda,N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right] - \mu\mathbb{E}\left[B_1(N_1^\lambda,N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \\ \ge&\left(\frac{h}{\gamma}+a\right)\left(\ell^\lambda - \mu N_2^\lambda(N_1^\lambda,\ell^\lambda) - \mu N_1^\lambda\right) \\ =&\left(\frac{h\mu}{\gamma}+a\mu\right)\left(-D_2^\lambda(N_1^\lambda,\ell^\lambda)\right),\end{aligned}$$

where recall from the proof of Proposition 1 that $B_1(N_1^\lambda,N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)$ is the steady-state number of busy servers among those that are staffed at the base stage, and $B_2(N_1^\lambda,N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)$ is the steady-state number of busy servers among those that are staffed at the surge stage. Therefore,

$$\begin{aligned}&\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2 D_2^\lambda(N_1^\lambda,\ell^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \\ \ge&\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2 D_2^\lambda(N_1^\lambda,\ell^\lambda) + \left(\frac{h\mu}{\gamma}+a\mu\right)\left(-D_2^\lambda(N_1^\lambda,\ell^\lambda)\right)\right) \\ =&\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2 - \frac{h\mu}{\gamma} - a\mu\right)D_2^\lambda(N_1^\lambda,\ell^\lambda) \\ =&\infty.\end{aligned}$$

For case (iii), it follows from (32) in the proof of Lemma 3 (for the case where $x < \beta_1^*$), and the same analysis as in the proof of Lemma 6 (for the case where $x > \beta_1^*$) that

$$
\begin{aligned}
&\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} \left(c_2 D_2^\lambda(N_1^\lambda, \ell^\lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right]\right) \\
=& c_2\eta + \lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] \\
=& \begin{cases} c_2\eta & \text{if } x < \beta_1^* \\ c_2\eta + \left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right) - \eta\sqrt{\frac{\mu}{\gamma}}\right]}{1 + \sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta)}} & \text{if } x > \beta_1^*. \end{cases}
\end{aligned}
$$

Moreover, in the scenario where $x < \beta_1^*$, we have $\beta_2^*(\beta_1^*, x) = 0$, so it must be that $D_2^\lambda \geq 0$ and $\eta \geq 0$. Therefore, (87) follows from the definition of η^* in (11).

For case (iv), we can further consider a subsequence indexed by λ_i along which $\bar{D}_2^{\lambda_i}(N_1^{\lambda_i}, \ell^{\lambda_i})$ converges. Such subsequence exists because a sequence has no convergent subsequence if and only if it approaches infinity. The same arguments for case (iii) can be applied to establish (87).

So far we have established (87). This, together with (86) and Lemma 13, gives that

$$
\begin{aligned}
\liminf_{\lambda\to\infty} \bar{\mathcal{C}}_u^\lambda &\geq \mathbb{E}\left[\liminf_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}} \left(c_2 D_2^\lambda(N_1^\lambda, \Lambda^\lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)|\Lambda^\lambda\right]\right)\right] \\
&\geq \mathbb{E}\left[\psi(X)\right] \\
&= \lim_{\lambda\to\infty} \bar{\mathcal{C}}_{2,QED}^\lambda,
\end{aligned}
$$

which establishes (85).

In this last step, note that by (85), we have

$$\liminf_{\lambda\to\infty} \bar{\mathcal{C}}_{2,*}^\lambda \geq \lim_{\lambda\to\infty} \bar{\mathcal{C}}_{2,QED}^\lambda.$$

Moreover, by the optimality of $u_{2,*}$, it holds that

$$\limsup_{\lambda\to\infty} \bar{\mathcal{C}}_{2,*}^\lambda \leq \lim_{\lambda\to\infty} \bar{\mathcal{C}}_{2,QED}^\lambda.$$

Therefore,

$$\lim_{\lambda\to\infty} \bar{\mathcal{C}}_{2,*}^\lambda = \lim_{\lambda\to\infty} \bar{\mathcal{C}}_{2,QED}^\lambda,$$

which implies that $\mathcal{C}_{2,QED}^\lambda - \mathcal{C}_{2,*}^\lambda = o(\sqrt{\lambda})$. Q.E.D.

Appendix F: Model with Surge-Stage Prediction Error

Recall that we use F_Y (alternatively, f_Y) and F_Z (alternatively, f_Z) to denote the cdf (alternatively, probability density function) of Y and Z, respectively.

F.1. Small Prediction Error: Proof of Proposition 2

Proof: Statement (I) follows exactly the same lines of analysis as the proof of Theorem 1 for $\alpha > 1/2$. Statement (II) follows exactly the same lines of analysis as the proof of Theorem 3. Lastly, following the same lines of analysis as the proof of Theorem 3, we can show that $\mathcal{C}_{2,ERR}^{e,\lambda} - \mathcal{C}_{2,*}^{o,\lambda} = o(\sqrt{\lambda})$. This, together with statement (II), implies statement (III). To elaborate on the generalization, we explain why the proof of Proposition 2 follows directly from the analysis of the case with perfect surge-stage prediction. In particular, when $\nu < 1/2$, the two-stage error policy takes the same form as the two-stage QED rule, with random variable X (alternatively, its realization x) replaced by random variable Y (alternatively, its realization y). For $\ell^\lambda = \lambda + y\lambda^\alpha \mu^{1-\alpha} + z\lambda^\nu \mu^{1-\nu}$, it still holds that if $y < F_Y^{-1}(c_1/c_2)$, then

$$N_1^\lambda + N_2^\lambda(N_1^\lambda, y) = \ell^\lambda/\mu + F_Y^{-1}(c_1/c_2)\left(\ell^\lambda/\mu\right)^\alpha + O(\sqrt{\ell^\lambda/\mu}).$$

In the other case where $y \geq F_Y^{-1}(c_1/c_2)$, we have

$$N_1^\lambda + N_2^\lambda(N_1^\lambda, y) = \ell^\lambda/\mu + \eta^* \left(\ell^\lambda/\mu\right)^\alpha + o(\sqrt{\ell^\lambda/\mu}),$$

for η^* defined in (11). The rest of the analysis is generalized similarly. Q.E.D.

F.2. Moderate to Large Prediction Error: Proof of Proposition 3

PROOF: We first show that there exists an optimal solution to (18). In particular, consider the inner-problem in (18):

$$\min_{N_2^\lambda(N_1^\lambda, Y)} \left\{ c_2 N_2^\lambda(N_1^\lambda, Y) + (h\mu/\gamma + a\mu)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ |Y\right]\right\}. \tag{88}$$

Note that (88) is a newsvendor problem with unit capacity cost c_2, unit sales price $h\mu/\gamma + a\mu$, random demand $\Lambda^\lambda/\mu - N_1^\lambda|Y$ (where the randomness lies in random variable Z), and capacity decision $N_2^\lambda(N_1^\lambda, Y)$. The optimal solution is given by

$$\bar{N}_2^\lambda(N_1^\lambda, Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + Y\left(\frac{\lambda}{\mu}\right)^\alpha - N_1^\lambda\right)^+.$$

Given $\bar{N}_2^\lambda(N_1^\lambda, Y)$, the outer-problem is given by $\min_{N_1^\lambda} h(N_1^\lambda)$, where

$$h(N_1^\lambda) := c_1 N_1^\lambda + \mathbb{E}\left[c_2 \bar{N}_2^\lambda(N_1^\lambda, Y) + (h\mu/\gamma + a\mu)\left(\Lambda^\lambda/\mu - N_1^\lambda - \bar{N}_2^\lambda(N_1^\lambda, Y)\right)^+\right].$$

Differentiating $h(N_1^\lambda)$ with respect to N_1^λ gives

$$\begin{aligned}\frac{\partial}{\partial N_1^\lambda} h(N_1^\lambda) = c_1 - c_2 \mathbb{P}&\left(\left(\frac{\lambda}{\mu}\right)^\alpha Y > \left(N_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu\right)\right) \\ &- \left(\frac{h\mu}{\gamma} + a\mu\right)\mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^\alpha Y \leq \left(N_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu\right),\right. \\ &\left.\left(\frac{\lambda}{\mu}\right)^\alpha Y + \left(\frac{\lambda}{\mu}\right)^\nu Z > N_1^\lambda - \frac{\lambda}{\mu}\right).\end{aligned}$$

By observation, $\frac{\partial}{\partial N_1^\lambda} h(N_1^\lambda)$ is continuous in N_1^λ, and there exist $N_1^{\lambda,L}$ and $N_1^{\lambda,U}$ such that $\frac{\partial}{\partial N_1^\lambda} h(N_1^{\lambda,L}) < 0$ and $\frac{\partial}{\partial N_1^\lambda} h(N_1^{\lambda,H}) > 0$. Thus, the intermediate value theorem implies that there exists critical point $\bar{N}_1^\lambda$ such that $\frac{\partial}{\partial N_1^\lambda} h(\bar{N}_1^\lambda) = 0$. In addition, $h(N_1^\lambda)$ is convex in N_1^λ, because

$$\begin{aligned}&\frac{\partial^2}{\partial (N_1^\lambda)^2} h(N_1^\lambda) \\ &= \left(\frac{h\mu}{\gamma} + a\mu\right)\left(\frac{\lambda}{\mu}\right)^{-\nu} \int_{-\infty}^{\left(\frac{\lambda}{\mu}\right)^{-\alpha}\left(N_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu\right)} f_Y(y) f_Z\left(\left(\frac{\lambda}{\mu}\right)^{-\nu}\left(N_1^\lambda - \frac{\lambda}{\mu} - y\left(\frac{\lambda}{\mu}\right)^\alpha\right)\right) dy \geq 0.\end{aligned}$$

Hence, $\bar{N}_1^\lambda$ is a global minimum of $h(N_1^\lambda)$, and $(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y))$ is optimal to (18).

Proof of (I). We discuss the following two cases: $\nu < \alpha$ and $\nu = \alpha$.

Case 1: $\nu < \alpha$. When $\nu < \alpha$, similar lines of analysis as the proof of Theorem 1 for $\alpha < 1/2$ go through. Due to the similarity in the steps, we shall present the key structure of the proof and omit the details.

Consider the two-stage staffing rule denoted by u, where the staffing levels are given by

$$N_1^\lambda := \lambda/\mu + \bar{F}_Y^{-1}(c_1/c_2)(\lambda/\mu)^\alpha, \quad \text{and} \quad N_2^\lambda(N_1^\lambda, Y) := \left(Y - \bar{F}_Y^{-1}(c_1/c_2)\right)^+ (\lambda/\mu)^\alpha.$$

Following the definition of $\hat{\mathcal{C}}_u^\lambda$ in (27), we define

$$\hat{\mathcal{C}}_u^{e,\lambda} := \frac{\mathcal{C}_u^{e,\lambda} - c_1\lambda/\mu}{(\lambda/\mu)^{\max\{\alpha,1/2\}}}.$$

Similar lines of arguments as in the proof of Lemma 4 establish that

$$\hat{\mathcal{C}}_u^{e,\lambda} \to c_1\bar{F}_Y^{-1}(c_1/c_2) + c_2\mathbb{E}\left[(Y - \bar{F}_Y^{-1}(c_1/c_2))^+\right] \quad \text{as } \lambda \to \infty.$$

In comparison, consider the single-stage staffing rule denoted by $\tilde{u}$, where the base-stage staffing level is

$$N_1^\lambda := \frac{\lambda}{\mu} + \bar{F}_Y^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right)(\lambda/\mu)^\alpha.$$

Similar lines of arguments as in the proof of Lemma 4 show that

$$\hat{\mathcal{C}}_{\tilde{u}}^{e,\lambda} \to c_1\bar{F}_Y^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right) + \left(\frac{h\mu}{\gamma} + a\mu\right)\mathbb{E}\left[\left(Y - \bar{F}_Y^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right)\right)^+\right] \quad \text{as } \lambda \to \infty,$$

where $\hat{\mathcal{C}}_{\tilde{u}}^{e,\lambda}$ is defined the same way as $\hat{\mathcal{C}}_u^{e,\lambda}$ but for policy $\tilde{u}$ instead.

By Assumption 1 and the continuity of Y, it can be verified that $\lim_{\lambda\to\infty}\hat{\mathcal{C}}_{\tilde{u}}^{e,\lambda} > \lim_{\lambda\to\infty}\hat{\mathcal{C}}_u^{e,\lambda}$. Thus,

$$\mathcal{C}_{\tilde{u}}^{e,\lambda} - \mathcal{C}_u^{e,\lambda} = \Theta(\lambda^\alpha).$$

Moreover, similar derivation as in the proof of Lemma 8 gives that

$$\mathcal{C}_{\tilde{u}}^{e,\lambda} - \mathcal{C}_{1,*}^{e,\lambda} = o(\lambda^\alpha) \quad \text{and} \quad \mathcal{C}_u^{e,\lambda} - \mathcal{C}_{2,*}^{e,\lambda} = o(\lambda^\alpha).$$

The statement follows.

Case 2: $\boldsymbol{\nu = \alpha}$. Consider the two-stage staffing rule denoted by u, where the staffing levels are given by

$$N_1^\lambda := \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha, \quad \text{and} \quad N_2^\lambda(N_1^\lambda, Y) := \beta_2^*(\beta_1^*, Y)(\lambda/\mu)^\alpha,$$

where β_1^* and $\beta_2^*(\beta_1^*, Y)$ jointly solve

$$\min_{\beta_1}\left\{c_1\beta_1 + \mathbb{E}\left[\min_{\beta_2(\beta_1,Y)\in\mathbb{R}_+}\left\{c_2\beta_2(\beta_1,Y) + (h\mu/\gamma + a\mu)\,\mathbb{E}\left[(Y + Z - \beta_1 - \beta_2(\beta_1,Y))^+ \,\middle|\, Y\right]\right\}\right]\right\}. \tag{89}$$

We first show that an optimal solution to (89) exists. Consider the inner-problem in (89):

$$\min_{\beta_2(\beta_1,Y)\in\mathbb{R}_+} c_2\beta_2(\beta_1,Y) + (h\mu/\gamma + a\mu)\,\mathbb{E}\left[(Y + Z - \beta_1 - \beta_2(\beta_1,Y))^+ \,\middle|\, Y\right]. \tag{90}$$

Note that (90) is a newsvendor problem with unit capacity cost c_2, unit sales price $h\mu/\gamma + a\mu$, random demand $Y + Z - \beta_1|Y$ (where the randomness lies in random variable Z), and capacity decision $\beta_2(\beta_1, Y)$. The optimal solution is given by

$$\beta_2^*(\beta_1, Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + Y - \beta_1\right)^+. \tag{91}$$

Given $\beta_2^*(\beta_1, Y)$, the outer-problem is given by $\min_{\beta_1\in\mathbb{R}} h(\beta_1)$, where

$$h(\beta_1) := \left\{c_1\beta_1 + \mathbb{E}\left[c_2\beta_2^*(\beta_1,Y) + (h\mu/\gamma + a\mu)\,(Y + Z - \beta_1 - \beta_2^*(\beta_1,Y))^+\right]\right\}.$$

Differentiating $h(\beta_1)$ with respect to β_1 gives

$$\begin{aligned}\frac{\partial}{\partial \beta_1} h(\beta_1) = c_1 - c_2 \mathbb{P}\left(Y > \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + \beta_1\right) \\ - \left(\frac{h\mu}{\gamma} + a\mu\right) \mathbb{P}\left(Y \leq \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + \beta_1, Y + Z > \beta_1\right).\end{aligned} \tag{92}$$

By observation, $\frac{\partial}{\partial \beta_1} h(\beta_1)$ is continuous in β_1, and there exist β_1^L and β_1^U such that $\frac{\partial}{\partial \beta_1} h(\beta_1^L) < 0$ and $\frac{\partial}{\partial \beta_1} h(\beta_1^H) > 0$. Thus, the intermediate value theorem implies that there exists critical point β_1^* such that $\frac{\partial}{\partial \beta_1} h(\beta_1^*) = 0$. In addition, $h(\beta_1)$ is convex in β_1, because

$$\frac{\partial^2}{\partial \beta_1^2} h(\beta_1) = \left(\frac{h\mu}{\gamma} + a\mu\right) \int_{-\infty}^{\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + \beta_1} f_Y(y) f_Z(-y + \beta_1) dy \geq 0.$$

Hence, β_1^* is a global minimum of $h(\beta_1)$.

Following similar lines of arguments as in the proof of Lemma 4 and Lemma 8, we get that

$$\lim_{\lambda \to \infty} \hat{\mathcal{C}}_u^{e,\lambda} = c_1 \beta_1^* + \mathbb{E}\left[c_2 \beta_2^*(\beta_1^*, Y) + (h\mu/\gamma + a\mu)\left(Y + Z - \beta_1^* - \beta_2^*(\beta_1^*, Y)\right)^+\right],$$

and

$$\mathcal{C}_u^{e,\lambda} - \mathcal{C}_{2,*}^{e,\lambda} = o(\lambda^\alpha). \tag{93}$$

Next, consider the single-stage policy denoted by $\tilde{u}$, where the base-stage staffing level is given by $N_1^\lambda := \lambda/\mu + \tilde{\beta}(\lambda/\mu)^\alpha$, for

$$\tilde{\beta} := \underset{\beta \in \mathbb{R}}{\arg\min}\; c_1 \beta + \left(\frac{h\mu}{\gamma} + a\mu\right) \mathbb{E}\left[(Y + Z - \beta)^+\right] = \bar{F}_{Y+Z}^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right). \tag{94}$$

Similar derivation as in the proof of Lemma 4 gives that

$$\lim_{\lambda \to \infty} \hat{\mathcal{C}}_{\tilde{u}}^{e,\lambda} = c_1 \tilde{\beta} + (h\mu/\gamma + a\mu)\, \mathbb{E}\left[\left(Y + Z - \tilde{\beta}\right)^+\right].$$

Theorem 1 in Bassamboo et al. (2010) establishes that

$$\mathcal{C}_{\tilde{u}}^{e,\lambda} - \mathcal{C}_{1,*}^{e,\lambda} = O(\lambda^{1-\alpha}). \tag{95}$$

If Assumption 2 holds, then

$$\beta_2^*(\beta_1^*, Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + Y - \beta_1^*\right)^+ > 0 \quad \text{with probability } p > 0. \tag{96}$$

To see (96), suppose for the sake of contradiction that $\beta_2^*(\beta_1^*, Y) = 0$ with probability 1. It follows by solving $\frac{\partial}{\partial \beta_1} h(\beta_1^*) = 0$ in (92) that $\beta_1^* = \tilde{\beta}$, for $\tilde{\beta}$ defined in (94). However, plugging in the value of $\tilde{\beta}$ in (91) gives that

$$\beta_2^*(\beta_1^*, Y) = \beta_2^*(\tilde{\beta}_1, Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + Y - \bar{F}_{Y+Z}^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right)\right)^+.$$

This, together with Assumption 2, implies that $\beta_2^*(\beta_1^*, Y) > 0$ with probability $p > 0$, a contradiction. Thus, (96) holds. It follows from (96) that $\lim_{\lambda \to \infty} \hat{\mathcal{C}}_{\tilde{u}}^{e,\lambda} > \lim_{\lambda \to \infty} \hat{\mathcal{C}}_u^{e,\lambda}$, so that

$$\mathcal{C}_{\tilde{u}}^{e,\lambda} - \mathcal{C}_u^{e,\lambda} = \Theta(\lambda^\alpha). \tag{97}$$

In the other case where Assumption 2 does not hold, similar derivation shows that $\beta_1^* = \tilde{\beta}$ and $\beta_2^*(\beta_1^*, Y) = \beta_2^*(\tilde{\beta}, Y) = 0$ is optimal to (89), and

$$\mathcal{C}_{\tilde{u}}^{e,\lambda} - \mathcal{C}_{u}^{e,\lambda} = o(\lambda^{\alpha}). \tag{98}$$

The statement follows from (93), (95), (97), and (98).

Proof of (II). We discuss the following three cases: $\mu = \gamma$, $\mu > \gamma$, and $\mu < \gamma$.

Case 1: $\boldsymbol{\mu = \gamma}$. It follows from Lemma 3 in Bassamboo et al. (2010) that for any staffing prescriptions N_1^λ and $N_2^\lambda(N_1^\lambda, Y)$, we have

$$\begin{aligned}
&\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \\
\le& \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y, Z\right] \\
\le& \left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ + \sqrt{\frac{4\pi}{\mu}}\sqrt{\Lambda^\lambda}\exp\left(-\frac{\mu}{4\Lambda^\lambda}\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^2\right) + \frac{1}{\log 2}.
\end{aligned} \tag{99}$$

Taking expectation of (99) conditional on Y gives

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right] &\le \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right] \\
&\le \mathbb{E}\left[\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right] + \mathbb{E}\left[\sqrt{\frac{4\pi}{\mu}}\sqrt{\Lambda^\lambda}\middle| Y\right] + \frac{1}{\log 2}.
\end{aligned} \tag{100}$$

It follows from (100) that

$$\begin{aligned}
&c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] \\
\le& c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
\le& c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] + \mathbb{E}\left[\sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda}\right] + 1/\log 2 \\
\le& c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] + \sqrt{4\pi/\mu}\sqrt{\lambda} \\
&+ \sqrt{4\pi/\mu}\sqrt{\lambda^\alpha \mu^{1-\alpha}\mathbb{E}\left[|Y|\right]} + \sqrt{4\pi/\mu}\sqrt{\lambda^\nu \mu^{1-\nu}\mathbb{E}\left[|Z|\right]} + 1/\log 2,
\end{aligned} \tag{101}$$

where the last inequality follows from the reverse Jensen's inequality, and the fact that Y and Z are independent.

Let $(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*}, Y))$ denotes the optimal solution to problem (16). We have

$$\begin{aligned}
\mathcal{C}_{2,Err}^{e,\lambda} &= c_1 \bar{N}_1^\lambda + \mathbb{E}\left[c_2 \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[Q(\bar{N}_1^\lambda + \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
&\overset{(a)}{\le} c_1 \bar{N}_1^\lambda + \mathbb{E}\left[c_2 \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(\bar{N}_1^\lambda + \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma\right] + O(\sqrt{\lambda}) \\
&\overset{(b)}{\le} c_1 N_1^{\lambda,*} + \mathbb{E}\left[c_2 N_2^{\lambda,*}(N_1^{\lambda,*}, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(N_1^{\lambda,*} + N_2^{\lambda,*}(N_1^{\lambda,*}, Y)\right)\right)^+ |Y\right]/\gamma\right] + O(\sqrt{\lambda}) \\
&\overset{(c)}{\le} c_1 N_1^{\lambda,*} + \mathbb{E}\left[c_2 N_2^{\lambda,*}(N_1^{\lambda,*}, Y) + (h + a\gamma)\mathbb{E}\left[Q(N_1^{\lambda,*} + N_2^{\lambda,*}(N_1^{\lambda,*}, Y), \Lambda^\lambda)|Y\right]\right] + O(\sqrt{\lambda}) \\
&= \mathcal{C}_{2,*}^{e,\lambda} + O(\sqrt{\lambda}),
\end{aligned}$$

where (a) follows from (101), (b) follows from the optimality of $(\bar{N}_1, \bar{N}_2(\bar{N}_1, Y))$ to problem (18), and (c) follows from (101) again.

Case 2: $\boldsymbol{\mu > \gamma}$. To simply notation, define

$$\begin{aligned}\mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) &:= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\ &= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h/\gamma + a)\,\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right)\right],\end{aligned} \tag{102}$$

where $\mathbb{P}(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y)$ denotes the steady-state abandonment probability conditional on Y, i.e., $\mathbb{P}(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y) := \mathbb{E}\left[\mathbb{1}_{(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)}|Y\right]$. In addition, define

$$\bar{\mathcal{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) := c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma\right]. \tag{103}$$

Note that $(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) = \arg\min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \bar{\mathcal{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y))$.

Consider an auxiliary sequence of systems with the same parameters as the original sequence of systems except that its abandonment rate is equal to μ; that is, systems in this sequence have a higher abandonment rate compared to the original sequence. We refer to this sequence as Sequence II and add the superscript II to all quantities associated with it, e.g., $\mu^{II} = \mu, \gamma^{II} = \mu$. Quantities associated with the original sequence of system are denoted without superscripts. For systems in Sequence II, we choose the cost parameters to be the following: $c_1^{II} = c_1, c_2^{II} = c_2, a^{II} = a$, and $h^{II} = h\mu/\gamma$. The analogues of (102) and (103) for Sequence II are

$$\begin{aligned}\mathcal{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) &:= c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + \left(h^{II}/\gamma^{II} + a^{II}\right)\mathbb{P}\left(AB^{II}, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right)\right] \\ &= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h/\gamma + a)\,\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right)\right] \\ &= \mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)),\end{aligned}$$

and

$$\begin{aligned}\bar{\mathcal{C}}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) &:= c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[\left(\Lambda^\lambda - \mu^{II}\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma^{II}\right] \\ &= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma\right] \\ &= \bar{\mathcal{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)).\end{aligned} \tag{104}$$

From the proof of Theorem 3 in Bassamboo et al. (2010), we have

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right) \le \mathbb{P}\left(AB^{II}, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right),$$

which implies that

$$\mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \le \mathcal{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)). \tag{105}$$

Applying (101) to Sequence II, we get that

$$\begin{aligned}&\mathcal{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\ =&c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[Q^{II}(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\ \le&c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[\left(\Lambda^\lambda/\mu^{II} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] + O(\sqrt{\lambda}) \\ =&\bar{\mathcal{C}}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) + O(\sqrt{\lambda})\end{aligned} \tag{106}$$

Next, consider another auxiliary sequence of systems with the same parameters as the original sequence of systems except that its service rate is equal to γ; that is, systems in this sequence have a lower service rate compared to the original sequence. We refer to this sequence as Sequence III and add the superscript III to all quantities associated with Sequence III, e.g., $\mu^{III}=\gamma, \gamma^{III}=\gamma$. For systems in Sequence III, we choose the cost parameters to be the following: $c_1^{III}=c_1\gamma/\mu, c_2^{III}=c_2\gamma/\mu, a^{III}=a$, and $h^{III}=h$. The analogues of (102) and (103) for Sequence III are

$$\begin{aligned}\mathcal{C}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y)) :=& c_1^{III}N_1^\lambda + \mathbb{E}\left[c_2^{III}N_2^\lambda(N_1^\lambda,Y) + \left(h^{III}/\gamma^{III}+a^{III}\right)\mathbb{P}\left(AB^{III}, N_1^\lambda+N_2^\lambda(N_1^\lambda,Y),Y\right)\right]\\ =& c_1\gamma/\mu N_1^\lambda + \mathbb{E}\left[c_2\gamma/\mu N_2^\lambda(N_1^\lambda,Y) + (h/\gamma+a)\,\mathbb{P}\left(AB^{III}, N_1^\lambda+N_2^\lambda(N_1^\lambda,Y),Y\right)\right],\end{aligned}$$

and

$$\begin{aligned}&\bar{\mathcal{C}}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y))\\ :=& c_1^{III}N_1^\lambda + \mathbb{E}\left[c_2^{III}N_2^\lambda(N_1^\lambda,Y) + (h^{III}+a^{III}\gamma^{III})\mathbb{E}\left[\left(\Lambda^\lambda-\mu^{III}\left(N_1^\lambda+N_2^\lambda(N_1^\lambda,Y)\right)\right)^+|Y\right]/\gamma^{III}\right]\\ =& c_1\gamma/\mu N_1^\lambda + \mathbb{E}\left[c_2\gamma/\mu N_2^\lambda(N_1^\lambda,Y) + (h+a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda-\gamma\left(N_1^\lambda+N_2^\lambda(N_1^\lambda,Y)\right)\right)^+|Y\right]/\gamma\right]\\ =& \bar{\mathcal{C}}^{e,\lambda}(\gamma/\mu N_1^\lambda, \gamma/\mu N_2^\lambda(N_1^\lambda,Y)).\end{aligned} \tag{107}$$

From the proof of Theorem 3 in Bassamboo et al. (2010), we have

$$\mathbb{P}\left(AB, N_1^\lambda+N_2^\lambda(N_1^\lambda,Y),Y\right) \geq \mathbb{P}\left(AB^{III}, \mu/\gamma\left(N_1^\lambda+N_2^\lambda(N_1^\lambda,Y)\right),Y\right),$$

which implies that

$$\mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y)) \geq \mathcal{C}^{e,\lambda,III}(\mu/\gamma N_1^\lambda, \mu/\gamma N_2^\lambda(N_1^\lambda,Y)). \tag{108}$$

Applying (101) to Sequence III, we get that

$$\begin{aligned}&\mathcal{C}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y))\\ =& c_1^{III}N_1^\lambda + \mathbb{E}\left[c_2^{III}N_2^\lambda(N_1^\lambda,Y) + (h^{III}+a^{III}\gamma^{III})\mathbb{E}\left[Q^{III}(N_1^\lambda+N_2^\lambda(N_1^\lambda,Y),\Lambda^\lambda)|Y\right]\right]\\ \geq& c_1^{III}N_1^\lambda + \mathbb{E}\left[c_2^{III}N_2^\lambda(N_1^\lambda,Y) + (h^{III}+a^{III}\gamma^{III})\mathbb{E}\left[\left(\Lambda^\lambda/\mu^{III}-N_1^\lambda-N_2^\lambda(N_1^\lambda,Y)\right)^+\middle|Y\right]\right]\\ =& \bar{\mathcal{C}}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y)),\end{aligned} \tag{109}$$

which implies that

$$\begin{aligned}\mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*},Y)) =& \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda,Y)} \mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y))\\ \overset{(d)}{\geq}& \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda,Y)} \mathcal{C}^{e,\lambda,III}(\mu/\gamma N_1^\lambda, \mu/\gamma N_2^\lambda(N_1^\lambda,Y))\\ \overset{(e)}{\geq}& \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda,Y)} \bar{\mathcal{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda,Y))\\ =& \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)),\end{aligned} \tag{110}$$

where (d) follows from (108), and (e) follows from (107) and (109).

Lastly, we can write

$$\begin{aligned}&\mathcal{C}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*},Y))\\ =& \mathcal{C}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) + \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*},Y))\\ \overset{(f)}{\leq}& \mathcal{C}^{e,\lambda,II}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) + \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*},Y))\\ \overset{(g)}{=}& \mathcal{C}^{e,\lambda,II}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \bar{\mathcal{C}}^{e,\lambda,II}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) + \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda,Y)) - \mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*},Y))\\ \overset{(h)}{=}& O(\sqrt{\lambda}),\end{aligned}$$

where (f) follows from (105), (g) follows from (104), and (h) follows from (106) and (110).

Case 3: $\mu < \gamma$. The analysis for Case 3 is similar to that for Case 2. In particular, we again consider Sequence II and Sequence III as constructed in Case 2.

For Sequence II, it follows by construction that

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right) \geq \mathbb{P}\left(AB^{II}, \left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right), Y\right),$$

which implies that

$$\mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \geq \mathcal{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)).$$

Applying (101) to Sequence II, we get that

$$\begin{aligned}
&\mathcal{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
=& c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[Q^{II}(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
\geq& c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[\left(\Lambda/\mu^{II} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] \\
=& c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h/\gamma + a)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]\right] \\
=& \bar{\mathcal{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)),
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*}, Y)) &= \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&\geq \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \mathcal{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&\geq \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \bar{\mathcal{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&= \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)).
\end{aligned} \tag{111}$$

For Sequence III, it follows by construction that

$$\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right) \leq \mathbb{P}\left(AB^{III}, \mu/\gamma\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right), Y\right),$$

which implies that

$$\mathcal{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \leq \mathcal{C}^{e,\lambda,III}(\mu/\gamma N_1^\lambda, \mu/\gamma N_2^\lambda(N_1^\lambda, Y)). \tag{112}$$

Applying (101) to Sequence III, we get that

$$\begin{aligned}
&\mathcal{C}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
=& c_1^{III} N_1^\lambda + \mathbb{E}\left[c_2^{III} N_2^\lambda(N_1^\lambda, Y) + (h^{III} + a^{III}\gamma^{III})\mathbb{E}\left[Q^{III}(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
\leq& c_1^{III} N_1^\lambda + \mathbb{E}\left[c_2^{III} N_2^\lambda(N_1^\lambda, Y) + (h^{III} + a^{III}\gamma^{III})\mathbb{E}\left[\left(\Lambda^\lambda/\mu^{III} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] + O(\sqrt{\lambda}) \\
=& \bar{\mathcal{C}}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) + O(\sqrt{\lambda})
\end{aligned} \tag{113}$$

Lastly, we can write

$$\begin{aligned}
&\mathcal{C}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) - \mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*}, Y)) \\
=& \mathcal{C}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) - \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) + \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) - \mathcal{C}^{e,\lambda}(N_1^*, N_2^*(N_1^*, Y)) \\
\overset{(i)}{\leq}& \mathcal{C}^{e,\lambda,III}(\mu/\gamma \bar{N}_1^\lambda, \mu/\gamma \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) - \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) + \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) - \mathcal{C}^{e,\lambda}(N_1^*, N_2^*(N_1^*, Y)) \\
\overset{(j)}{=}& \mathcal{C}^{e,\lambda,III}(\mu/\gamma \bar{N}_1^\lambda, \mu/\gamma \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) - \bar{\mathcal{C}}^{e,\lambda,III}(\mu/\gamma \bar{N}_1^\lambda, \mu/\gamma \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) + \bar{\mathcal{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)) \\
&- \mathcal{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*}, Y)) \\
\overset{(k)}{=}& O(\sqrt{\lambda}),
\end{aligned}$$

where (i) follows from (112), (j) follows from (107), and (k) follows from (111) and (113).

Proof of (III). For the oracle problem, we consider the following stochastic-fluid optimization problem

$$\min_{N_1^\lambda}\left\{c_1 N_1^\lambda + \mathbb{E}\left[\min_{N_2^\lambda(N_1^\lambda,\Lambda^\lambda)}\left\{c_2 N_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h\mu/\gamma + a\mu)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda,\Lambda^\lambda)\right)^+ |\Lambda^\lambda\right]\right\}\right]\right\}. \quad (114)$$

whose optimal solution is given by

$$\hat{N}_1^\lambda = \bar{F}^{-1}_{\Lambda^\lambda/\mu}(c_1/c_2)\,(\lambda/\mu)^\alpha, \quad \text{and} \quad \hat{N}_2^\lambda(\hat{N}_1^\lambda,\Lambda^\lambda) = (\Lambda^\lambda/\mu - \hat{N}_1^\lambda)^+.$$

We denote the staffing rule that prescribes $(\hat{N}_1^\lambda, \hat{N}_2^\lambda(\hat{N}_1^\lambda,\Lambda^\lambda))$ as $\hat{u}$. The same lines of analysis used to show statement (II) can be applied to establish that

$$\mathcal{C}^{o,\lambda}_{\hat{u}} - \mathcal{C}^{o,\lambda}_{2,*} = O(\sqrt{\lambda}). \quad (115)$$

Recall from the proof of Proposition 3 that $u_{2,ERR}$ prescribes staffing levels $(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y))$ where

$$\bar{N}_2^\lambda(N_1^\lambda, Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + Y\left(\frac{\lambda}{\mu}\right)^\alpha - N_1^\lambda\right)^+,$$

and $\bar{N}_1^\lambda$ solves

$$\begin{aligned} 0 =& c_1 - c_2\mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^\alpha Y > \left(\bar{N}_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu\right)\right) \\ &- \left(\frac{h\mu}{\gamma} + a\mu\right)\mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^\alpha Y \le \left(\bar{N}_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu\right),\right. \\ &\left.\left(\frac{\lambda}{\mu}\right)^\alpha Y + \left(\frac{\lambda}{\mu}\right)^\nu Z > \bar{N}_1^\lambda - \frac{\lambda}{\mu}\right). \end{aligned} \quad (116)$$

Next, we compare the two inner-optimization problems in (18) and (114). It holds that

$$\begin{aligned} &\mathbb{E}\left[\min_{N_2^\lambda(\bar{N}_1^\lambda, Y)}\left\{c_2 N_2^\lambda(\bar{N}_1^\lambda, Y) + \left(\frac{h\mu}{\gamma} + a\mu\right)\mathbb{E}\left[\left(\frac{\Lambda^\lambda}{\mu} - \bar{N}_1^\lambda - N_2^\lambda(\bar{N}_1^\lambda, Y)\right)^+ \middle| Y\right]\right\}\right] \\ &\quad - \mathbb{E}\left[\min_{N_2^\lambda(\bar{N}_1^\lambda,\Lambda^\lambda)}\left\{c_2 N_2^\lambda(\bar{N}_1^\lambda,\Lambda^\lambda) + \left(\frac{h\mu}{\gamma} + a\mu\right)\left(\frac{\Lambda^\lambda}{\mu} - \bar{N}_1^\lambda - N_2^\lambda(\bar{N}_1^\lambda,\Lambda^\lambda)\right)^+\right\}\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[c_2\left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda\right)^+\right. \\ &\quad + \left(\frac{h\mu}{\gamma} + a\mu\right)\left(\frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha + z\left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda - \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda\right)^+\right)^+ \\ &\quad \left. - c_2\left(\frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha + z\left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda\right)\right] f_Y(y) f_Z(z)\,dy\,dz. \end{aligned} \quad (117)$$

Denote part of the integrand in (117) as

$$\begin{aligned} g^\lambda(y,z) :=& c_2\left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda\right)^+ \\ &+ \left(\frac{h\mu}{\gamma} + a\mu\right)\left(\frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha + z\left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda - \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda\right)^+\right)^+ \\ &- c_2\left(\frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^\alpha + z\left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda\right)^+. \end{aligned}$$

By construction of the two optimization problems, it holds that $g^\lambda(y,z) \ge 0$ for all $y, z \in \mathbb{R}$. Moreover, it follows from (116) that at least one of the following two cases holds:

(i) $\mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^{\alpha} Y > \left(\bar{N}_1^{\lambda} - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\nu}\right)\right) > 0$;

(ii) $\mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^{\alpha} Y \leq \left(\bar{N}_1^{\lambda} - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\nu}\right), \left(\frac{\lambda}{\mu}\right)^{\alpha} Y + \left(\frac{\lambda}{\mu}\right)^{\nu} Z > \bar{N}_1^{\lambda} - \frac{\lambda}{\mu}\right) > 0.$

Note that if $\left(\frac{\lambda}{\mu}\right)^{\alpha} y > \left(\bar{N}_1^{\lambda} - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\nu}\right)$ and $z \neq \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)$, then

$$\begin{aligned}
g^{\lambda}(y,z) = {} & c_2\left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\nu} + \frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^{\alpha} - \bar{N}_1^{\lambda}\right) \\
& + \left(\frac{h\mu}{\gamma} + a\mu\right)\left(z\left(\frac{\lambda}{\mu}\right)^{\nu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\nu}\right)^{+} \\
& - c_2\left(\frac{\lambda}{\mu} + y\left(\frac{\lambda}{\mu}\right)^{\alpha} + z\left(\frac{\lambda}{\mu}\right)^{\nu} - \bar{N}_1^{\lambda}\right)^{+} \\
= {} & \Theta(\lambda^{\nu}).
\end{aligned}$$

In addition, if $\left(\frac{\lambda}{\mu}\right)^{\alpha} y \leq \left(\bar{N}_1^{\lambda} - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\nu}\right)$ and $\left(\frac{\lambda}{\mu}\right)^{\alpha} y + \left(\frac{\lambda}{\mu}\right)^{\nu} z > \bar{N}_1^{\lambda} - \frac{\lambda}{\mu}$, then

$$g^{\lambda}(y,z) = (h\mu/\gamma + a\mu - c_2)\left(\lambda/\mu + y\,(\lambda/\mu)^{\alpha} + z\,(\lambda/\mu)^{\nu} - \bar{N}_1^{\lambda}\right)^{+} = \Theta(\lambda^{\nu}).$$

Therefore, we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g^{\lambda}(y,z) f_Y(y) f_Z(z)\, dy dz = \Theta(\lambda^{\nu}). \tag{118}$$

It follows from (117), (118), and the construction of stochastic-fluid problems (18) and (114) that

$$\mathcal{C}_{2,ERR}^{e,\lambda} - \mathcal{C}_{\hat{u}}^{o,\lambda} = \Theta(\lambda^{\nu}). \tag{119}$$

The statement follows from (115), (119), and statement (II). Q.E.D.

Appendix G: Non-Parametric Estimation of α and σ

In this section, we provide more details of the non-parametric estimation proposed in Maman (2009) to approximate the relationship between α and σ in the random arrival rate (3). In particular, this method does not impose any distributional assumption on X. However, it requires that $\alpha > 1/2$.

Let L_i be a generic random variable denoting the arrival count during a type-i shift, $i \in \mathcal{I}$. Since $L_i|\Lambda_i \sim$Poisson(Λ_i), we have

$$\begin{aligned}
\mathbb{E}[L_i] &= \mathbb{E}[\mathbb{E}[L_i|\Lambda_i]] = \lambda_i \\
\mathrm{Var}(L_i) &= \mathrm{Var}(\mathbb{E}[L_i|\Lambda_i]) + \mathbb{E}[\mathrm{Var}(L_i|\Lambda_i)] = \lambda_i^{2\alpha}\sigma^2 + \lambda_i, \quad i \in \mathcal{I}.
\end{aligned}$$

Thus,

$$\frac{\mathrm{Std}(L_i)}{\lambda_i^{\alpha}} = \left(\sigma^2 + \lambda_i^{1-2\alpha}\right)^{1/2}, \quad i \in \mathcal{I}.$$

In addition, since $\alpha > 1/2$,

$$\lim_{\lambda\to\infty}\left(\log \mathrm{Std}(L_i) - \alpha \log \lambda_i\right) = \log \sigma, \quad i \in \mathcal{I}.$$

Hence, it holds for large λ_i that

$$\log \mathrm{Std}(L_i) \approx \alpha \log \lambda_i + \log \sigma, \quad i \in \mathcal{I}.$$

Using sample mean $\bar{L}_i$ to approximate λ_i and sample standard deviation Σ_i to approximate $\mathrm{Std}(L_i)$, we get that

$$\log \Sigma_i \approx \hat{\alpha} \log \bar{L}_i + \log \hat{\sigma}, \quad i \in \mathcal{I},$$

which is equivalent to (20) in our parametric estimation setting.

Appendix H: Supplementary Numerical Experiments

H.1. Translation of The Two-Stage QED Staffing Rule

In this appendix we conduct more numerical experiments to examine system performance under the two-stage QED staffing rule with different specifications of k in (12). In what follows, we repeat the experiments in Tables 2 (with $c_2 = 2$) and 3 (with $c_2 = 10$) for other values of surge staffing costs, i.e., $c_2 = 6, 14$. We remark that for the system parameters under consideration, Assumption 1 requires that $c_2 < 18$. The results of these experiments corroborate the efficacy of the particular form of the two-stage QED staffing rule proposed in (13) for small systems.

Table 7 System performance under different specifications of the two-stage QED staffing rule with $\boldsymbol{\beta^* = 0.967, \eta^* = 0.120}$

$\boldsymbol{(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 6)}$

λ \ k	-3	-2	-1	0	1	2	3
25	60.30%	29.61%	9.85%	1.12%	0.00%	5.64%	15.09%
50	41.01%	20.29%	7.18%	1.19%	0.00%	3.44%	9.99%
75	30.78%	15.83%	5.33%	0.81%	0.00%	2.91%	8.66%
100	23.59%	10.91%	3.29%	0.23%	0.00%	2.93%	7.91%

Table 8 System performance under different specifications of the two-stage QED staffing rule with $\boldsymbol{\beta^* = 1.465, \eta^* = -0.380}$

$\boldsymbol{(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 14)}$

λ \ k	-3	-2	-1	0	1	2	3
25	76.65%	31.24%	7.00%	0.00%	4.10%	13.40%	24.12%
50	49.06%	21.06%	5.20%	0.00%	3.47%	10.51%	18.65%
75	37.26%	15.58%	3.26%	0.00%	2.63%	8.67%	15.11%
100	27.70%	11.56%	2.20%	0.00%	2.59%	7.64%	13.83%

H.2. Surge-Stage Linear Regression Model

Table 9 below provides the estimated coefficients in the surge-stage linear regression model.

Table 9: Surge-stage linear regression results

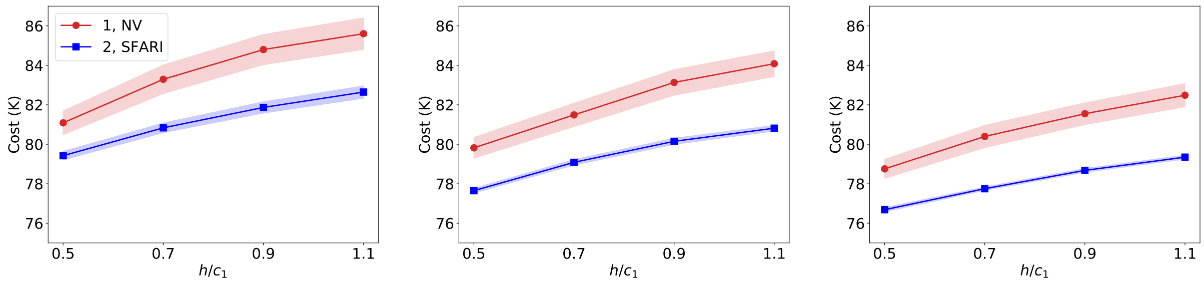
	Dependent variable:
	Observed
Monday day	119.972*** (115.275, 124.668)
Tuesday day	97.307*** (91.680, 102.934)
Wednesday day	96.277*** (91.056, 101.497)
Thursday day	93.560*** (88.420, 98.700)
Friday day	83.007*** (77.792, 88.222)
Saturday day	57.421*** (51.948, 62.894)
Sunday day	53.682*** (48.349, 59.014)
Monday night	9.599*** (4.116, 15.082)

Tuesday night	6.170* (0.915, 11.426)
Wednesday night	2.755 (−2.481, 7.990)
Thursday night	3.963 (−1.235, 9.161)
Friday night	5.650* (0.213, 11.088)
Saturday night	5.496* (0.161, 10.832)
Winter	3.021 (−0.699, 6.741)
Summer	−1.574 (−4.919, 1.772)
Fall	−2.355 (−5.519, 0.808)
Holiday	−22.392*** (−28.168, −16.616)
Holiday − 1 day	−10.137*** (−15.761, −4.513)
Holiday + 1 day	16.840*** (11.174, 22.507)
Min temperature	0.532*** (0.344, 0.719)
Precipitation	−0.160*** (−0.249, −0.071)
Snow	−0.169*** (−0.224, −0.114)
Wind	0.078** (0.018, 0.139)
Max temperature $\geq$ 86°F	−5.761*** (−9.292, −2.231)
1-day lag	0.013 (−0.030, 0.055)
7-day lag	0.038 (−0.001, 0.078)
30-day moving average	0.012 (−0.041, 0.065)
Google trend "depression"	−0.098 (−0.231, 0.034)
Google trend "flu"	0.270** (0.087, 0.452)
Average weighted comorbidity score per patient over the last 3 days	14.848* (0.345, 29.352)
Constant	57.365*** (22.998, 91.733)
Observations	730
R^2	0.908
Adjusted R^2	0.904
Residual Std. Error	14.316 (df = 699)
F Statistic	231.112*** (df = 30; 699)
Note:	*p<0.1; **p<0.05; ***p<0.01

H.3. Robustness of The Proposed Staffing Rule

In this section we conduct numerical experiments to check the robustness of the proposed staffing rules with respect to ED-specific patient-flow dynamics. In particular, we consider the parameters associated with Thursday day shifts, and run simulations incorporating different levels of ED-specific features that are not considered in the theoretical model. To prevent prediction error from confounding the results, we assume prefect demand information at the surge stage. In particular, we compare the oracle policy $u_{2,SFARI}$ with the single-stage newsvendor solution $u_{1,NV}$. Figure 10(a) provides a reference to the theoretical setting, where we assume exponential service times, constant arrival rate during the shift (which is equal to the average shift-level arrival rate shown in Table 4), and initialize Thursday day shift at its expected steady-state queue length conditional on the realized arrival rate. The cost curves are generated by increasing the holding cost so that its ratio to the base-stage staffing cost is from 0.5 to 1.1 in increment of 0.2. The 95% confidence intervals are derived by simulating 1000 realizations of Thursday day shifts for each holding cost and each policy. With everything else held constant to that in Figure 10(a), Figure 10(b) assumes lognormal (as opposed to exponential) service times, and Figure 10(c) considers both lognormal service times and hourly-varying arrival rates. We observe that the cost curves in Figures 10(a)–10(c) are very similar. This implies that lognormal service times and hourly-varying arrival rates do not significantly deviate system performance from that in the theoretical setting.

Figure 10 **Impact of service time distribution and non-stationary arrival rate**



(a) Theoretical setting (b) Lognormal service time (c) Lognormal service time and hourly-varying arrival rate

H.4. ED-Catered Staffing Adjustment

In this section we compare the proposed ED-catered staffing adjustment to the optimized one among the same family of adjustment schemes. Recall from Section 7.4.2 that to account for the end-of-shift effects, we propose an adjustment scheme for the two-stage error policy and heuristically set $\xi_1 = 5$ and $\xi_2 = 1$. To make the comparison fair, we apply the same base-stage adjustment to the single-stage newsvendor solution in the ED setting. In what follows, we optimize the adjustment parameters for the two policies separately. In particular, we simulate the ED over 52 weeks for a wide range of holding costs whose ratio to the base-stage staffing cost range from 0.5 to 1.1. We allow the abandonment cost to grow proportionally to the holding cost by fixing their ratio to be 1.5. For each policy and each holding cost, we enumerate ξ_1 (as well as ξ_2 for the two-stage error policy) from 0 to 10 in increment of 1. (Note that the optimal adjustment parameters are both policy- and cost-dependent.) Figure 11 compares the tradeoff curves of $u_{2,ERR}$ and $u_{1,NV}$ using (i) the heuristic adjustment, (ii) the optimized adjustment, and (iii) no adjustment. We note that when there is no adjustment, $u_{2,ERR}$ outperforms $u_{1,NV}$ for all holding costs. Compared to no adjustment, the heuristic and optimized adjustments further reduce the expected total costs for $u_{2,ERR}$. In addition, the tradeoff curves generated using the heuristic and optimized adjustments are almost indistinguishable. These results demonstrate significant value from applying transient-shift adjustment to $u_{2,ERR}$. Given the high proximity of the tradeoff curves yielded by the heuristic and optimized adjustments, applying the simple heuristic is effective and circumvents additional computation need.

Figure 11 **Comparison of tradeoff curves under the proposed and optimized adjustment parameters**