Utilizing Prediction Analytics in the Optimal Design and Control of Healthcare Systems

Yue Hu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Utilizing Prediction Analytics in the Optimal Design and Control of Healthcare Systems

Yue Hu

In recent years, increasing availability of data and advances in predictive analytics present new opportunities and challenges to healthcare management. Predictive models are developed to evaluate various aspects of healthcare systems, such as patient demand, patient pathways, and patient outcomes. While these predictions potentially provide valuable information to improve healthcare delivery, there are still many open questions considering how to integrate these forecasts into operational decisions. In this context, this dissertation develops methodologies to combine predictive analytics with the design of healthcare delivery systems.

The first part of dissertation considers how to schedule proactive care in the presence of patient deterioration. Healthcare systems are typically limited resource environments where scarce capacity is reserved for the most urgent patients. However, there has been a growing interest in the use of proactive care when a less urgent patient is predicted to become urgent while waiting. On one hand, providing care for patients when they are less critical could mean that fewer resources are needed to fulfill their treatment requirement. On the other hand, due to prediction errors, the moderate patients who are predicted to deteriorate in the future may self cure on their own and never need the treatment. Hence, allocating limited resource for these patients takes the capacity away from other more urgent ones who need it now. To understand this tension, we propose a multi-server queueing model with two patient classes: moderate and urgent. We allow patients to transition classes while waiting. In this setting, we characterize how moderate and urgent patients should be prioritized for treatment when proactive care for moderate patients is an option.

The second part of the dissertation focuses on the nurse staffing decisions in the emergency departments (ED). Optimizing 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 the second part of the dissertation, 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. 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.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the help and support of many people.

First and foremost, I am deeply indebted to my advisors, Carri W. Chan, Jing Dong, and Ohad Perry. I cannot express enough of my gratitude, respect, and admiration for their amazing qualities and all-time support. Carri has never failed to impress me with her strong enthusiasm for research and wholehearted care for students. I am always amazed at her capability to strike a good balance among research work, availability to students, services as journal editor and conference coordinator, administrative duties as the director of the Healthcare and Pharmaceutical Management Program and the director of our PhD program during my course of stay, and family responsibilities. I look up to her with the hope that one day I would become as confident, charismatic, and efficient, and focused as she is. Jing is a perfect supervisor who is brilliant and sharp, yet approachable and patient. I cannot thank her enough for spending numerous hours mentoring me over the five years of PhD pursuit and the last two years of undergraduate study. At work, she shapes my research vision in working on fundamentally important research questions and thinking deeply into each problem. She also tirelessly encourages me to practice writing, presentation, and communication. In life, she truly cares about my well-being and is aware of my ups and downs. The many meals, conversations, and outings that we had together are most memorable and will always be deeply cherished in my heart. Ohad gave me the first exposure to research in my junior year of the undergraduate study. I still vividly remember our conversation on Sheridan Road seven years ago when he asked me if I would be interested

in starting a research project together. An important lesson I have learned from Ohad is to be ambitious, hardworking, and fearless. When I was concerned about registering for PhD courses as a junior undergraduate, Ohad encouraged me and said: “The PhDs were just like you two years ago." Throughout the years, whenever I was in difficulties and having self-doubts, I kept telling myself: “They (the established ones) were just like you, and you can do it." I am extremely proud to be advised by Carri, Jing, and Ohad. I hope to pass the positive influence and learning I gained from these great minds on to future generations.

I would like to thank David D. Yao and Assaf Zeevi for agreeing to serve on my committee. I also owe a debt of gratitude to other faculty and staff members at Columbia University, including Nick Arnosti, Santiago Balseiro, Omar Besbes, Mark Broadie, Elizabeth Elam, Adam Elmachtoub, Awi Federgruen, Paul Glasserman, Vineet Goyal, Yash Kanoria, Winnie Leung, Hongyao Ma, Will Ma, Costis Maglaras, Ciamac Moallemi, Hongseok Namkoong, Daniel Russo, Ke Zeng, and Fanyin Zheng, for their constructive feedback. My special thanks goes to my faculty and peer mentors in my undergraduate study at Northwestern University, including David Morton, Karen Smilowitz, Jill H. Wilson, and Haoxiang Yang.

Life in New York would not have been enjoyable without the support from many friends. I would like to express my profound gratitude to my best friend, Yifan (Emma) Wang, who I have known for fourteen years since middle school. Emma has all the traits that I aspire to possess: intelligent, energetic, enthusiastic, dedicated, and inspiring. The years after Emma obtained her JD from Stanford Law School and started working in New York are most memorable—I will always remember our joy and tears, struggles and achievements, peaks and valleys in New York. Regardless of where we are, we will always be the biggest supporter, listener, and cheerleader for each other. I am privileged to by accompanied by many sincere friends at Columbia University, including Paolo Baudissone, Yi Cao, Xin Chen, Yuri Fonseca, Anand Kalvit, Mengxuan Liu, Zhe Liu, and Qi (Jimmy) Qin, who unreservedly support me and invariably wish me the best. I also very much enjoyed many meaningful conversations with Francisco Castro, Yilun Chen, Boris Epstein, Yuanling (Judy) Gan, Oliver Giesecke, Jiaqi Lu, Omar Mouchtaki, Pengyu Qian, Yiwen Shen, Ran Snitkovsky, and Noa Zychlinski, who inspired me in both research and life.

Last but not the least, I would like to thank my family for their unconditional love and support. My mother, Xiaohong Sun, puts my happiness and well-being as her top priority. She is always willing to listen to the ebbs and flows of my life and never fails to enlighten me with new perspectives. She also exemplifies the virtue of resilience and optimism, with which she is able to overcome any obstacle, no matter how challenging it is. My father, Dingxiang Hu, has taught me since my early childhood the importance of diligence and persistence. He believes that getting the knack of any skill requires incessant learning and practice, day in and day out. I am glad that I can now resonate with his viewpoint. I conclude by thanking my beloved grandmother, Yanhua Shao, who is an incredibly positive and tough woman. She is extraordinarily open-minded and willing to embrace new concepts and skills. Getting to spend so little time with me after I moved to the US is hard for her, yet she loves me with all her heart and supports me every step of the way. I dedicate this dissertation to them.

New York, NY
April, 2022

To my family

Introduction

Healthcare delivery systems face convoluted operational challenges different from those in conventional service systems. On the micro level, patients can experience complex disease progression and interact simultaneously or sequentially with a myriad of heterogeneous resources. On the macro level, healthcare systems are typically networks of many interacting elements that exert mutual influence on each other and on the system as a whole. These complexities result in significant uncertainties in demand, supply, and patient pathways. In the era of burgeoning information and data science, predictive analytics have revealed inspiring opportunities to mitigate uncertainties in various aspects of healthcare delivery and revolutionize traditional operations. For example, to increase throughput in the emergency department (ED), tests can be ordered proactively for patients at triage based on their likelihood of needing them. In addition, patients can be advanced to the inpatient ward before having official admission orders if their disposition can be accurately predicted. These prediction-based practices have the potential to improve operational efficiency without comprising quality of care.

In this dissertation, we aim to develop methodologies to combine predictive analytics with the design and control of healthcare delivery systems. The dissertation has two parts. Part I (Chapter 1) focuses on designing the optimal scheduling policy of proactive care in the presence of predicted patient deterioration and improvement. Part II (Chapters 2 and 3) addresses the nurse staffing problem in the ED based on demand forecasts.

In Part I (Chapter 1), we begin with the observation that in healthcare systems, it is typical that scarce medical resources are reserved for the most severe patients. In recent years, there has been a growing interest in the use of proactive care when a less urgent patient is predicted to become urgent while waiting. On one hand, advancing care for patients when they are less critical could mean that fewer resources are needed to fulfill their treatment requirement. On the other hand, utilizing limited resources for patients who are less critical may take capacity away from the more critical ones. Moreover, due to prediction errors, some of these less critical patients may self cure on their own without ever needing critical care. Thus, providing proactive care to them may end up generating more workload for the system.

In Chapter 1, we aim to develop a better understanding of the key tradeoffs in providing preventative care. To this end, we propose a multi-class queueing system that explicitly models patients' deterioration and improvement behavior, and study the optimal scheduling policy for proactive care. We analyze both the long-run and transient performance, with the focus on developing structural insights on the optimal policy. The long-run average optimization problem provides guidance on scheduling proactive care when the system is in its “normal" state of operation. The transient analysis further sheds light on the most cost-effective way to bring the system back to normal after a surge in demand due to random shocks such as disease outbreaks and mass casualty events. Our analysis quantifies the merits of proactive care and the impact of prediction errors on the optimal scheduling policy. Since the tradeoffs for proactive care can be similarly observed in other service sectors, the exposition of Chapter 1 is applicable to service systems in general and not limited to healthcare settings. There, “customer" and “proactive service" can be understood as the counterpart for “patient" and “proactive care," respectively.

During the global pandemic caused by coronavirus disease (COVID-19) in 2020, critical care physicians from New York and Florida reached out to us with the question of when to apply different levels of respiratory support for patients with COVID-19-associated respiratory failure. We applied insights from Chapter 1 to help the physicians derive allocation policies of high-flow nasal cannula and mechanical ventilators based on projected patient deterioration. Results of this application are summarized in Gershengorn et al. (2021).

In Part II (Chapters 2 and 3), we focus on the ED nurse staffing problem. ED crowding is a significant problem across the world, leading to adverse effects on patient outcomes, patient satisfaction, and staff morale. Nurses provide a substantial portion of patient care and are often a bottleneck resource in the ED. Despite its central role in reducing patient waiting time, the nurse staffing problem has been a time-honored challenge for hospitals, especially because there is a high level of demand uncertainty and staffing decisions have to be made ahead of time. In recent years, rapid progress of machine learning provides an opportunity to mitigate demand uncertainty by building advanced prediction models for ED demand.

In Chapter 2, we evaluate the effectiveness of rich real-time information in predicting shift-level ED patient volume. We aim to understand which real-time information has predictive power, and what prediction techniques are appropriate for forecasting ED demand. To this end, we conduct a retrospective study in an ED site in a large academic hospital in New York City. We examine various prediction techniques including linear regression, regression tree, extreme gradient boosting, and time series models. By comparing models with and without real-time predictors, we assess the potential gain in prediction accuracy from real-time information. We find that real-time predictors improve prediction accuracy upon models without contemporary information. Among extensive real-time predictors examined, recent patient arrival counts, weather, Google trends, and concurrent patient comorbidity information have significant predictive power. Out of all the forecasting techniques explored, SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) achieves the smallest out-of-sample RMSE (Root-Mean-Square Error) of 13.803 and MAPE (Mean Absolute Percentage Error) of 8.482%. Linear regression is the second best with out-of-sample RMSE and MAPE equal to 14.089 and 8.633%, respectively.

In Chapter 3, we propose a two-stage staffing framework that integrates the prediction models developed in Chapter 2 into ED nurse staffing decisions. In particular, the base staffing decision is determined weeks in advance, when the demand information is relatively crude. The surge staffing decision is set hours before the beginning of the nursing shift, when the demand forecast is much more accurate. We find that when the ED faces significant demand uncertainty, reasonably accurate demand prediction leads to significant cost savings (11%–16% or $2 M–$3 M) while guaranteeing timely access to care. Our proposed prediction-driven staffing rule lends to itself an intuitive interpretation and achieves near-optimal performance. Preliminary results have been presented to ED management in New York for the possibility of running a pilot study of our proposed prediction model and two-stage staffing rule.

Chapter 1: Optimal Scheduling of Proactive Service with Customer Deterioration and Improvement

1.1 Introduction

With recent advancements of predictive analytics and data availability, considerable efforts have been made to develop predictive tools for service systems. For example, in healthcare settings, predictive models have been created to evaluate the risk of ICU admission (Churpek et al., 2014), hospital acquired infection (Chang et al., 2011), Cardiovascular events (Rumsfeld et al., 2016), and various other adversarial patient deterioration. In call centers, predictive models have been developed to identify customers who are likely to contact their insurance company based on past claims data (Jerath et al., 2015).

From the operations perspective, predictive information on customers' future service needs brings the opportunity of developing approaches to provide effective proactive service and, potentially, improve system performance. In healthcare, there is well-documented evidence that delayed treatment can lead to worse medical outcomes such as longer length of stay or higher mortality rate (Chan et al., 2008; Chalfin et al., 2007; Chan et al., 2016). Proactive care, with the help of the predictive models that forecast patient deterioration, can help reduce treatment delays and improve patient outcomes (Hu et al., 2018). In the insurance company call center example, Jerath et al. (2015) advocate reaching out proactively to customers who have a high probability of calling to increase customer satisfaction and reduce peak demand.

Isolating the potential impact of proactive service is not straightforward. On one hand, advancing service for customers when they are less urgent could mean that fewer resources are needed to fulfill their service requirement. This has the potential benefit of reducing the overall workload of the system. On the other hand, utilizing limited capacity for customers who are less critical may take capacity away from other more critical customers whose service needs are more urgent. Moreover, some of these less critical customers may be satisfied without ever needing the critical service. Thus, providing proactive service to them may end up generating more workload for the system. In this chapter, to develop a better understanding of the key tradeoffs in proactive service, we propose a multi-class queueing system that explicitly models customers' deterioration and improvement behavior, and study the optimal scheduling policy for proactive service based on the model.

While proactive service has long been considered in manufacturing settings where preventative maintenance effectively reduces the demand for future repair services (McCall, 1965; Pierskalla and Voelker, 1976), in service systems, there are very few works analyzing proactive service with predictive information about customers' future needs (see Section 1.1.1 for a detailed review of some related works). Our modeling approach aims to provide a systematic way to capture the key tradeoffs in the limited resource environment: the potential benefit of serving customers early on with fewer resources versus the potential cost of delaying service for the more urgent customers and generating more overall workload to the system. Moreover, our analysis provides insights on how the accuracy of the predictive information affects the prioritization of services.

We conduct analysis on both the long-run average performance and the transient performance, with the focus on developing structural insights into the optimal scheduling policy. The long-run average performance analysis provides guidance on scheduling proactive service when the system is in its “normal" state of operation. That said, service systems often operate in a highly non-stationary environment. A surge in demand due to random shocks, e.g., disease outbreaks or mass casualty events for hospitals and insurance companies, weather patterns resulting in mass flight cancellations for airline call centers, etc., can bring the system far from its normal state of operation. It is thus important to study the transient optimal control and to develop an understanding of the most cost-effective way to bring the system back to normal.

Our analysis quantifies the merits of proactive service. We are able to characterize settings where proactive service can be beneficial and others where it is better to focus all resources on the most urgent customers. Our main contributions can be summarized as follows.

Queueing model with dynamic class types. We propose a Markovian multi-server queue with two customer classes: urgent and moderate. The key feature we incorporate is that a moderate customer who does not receive timely service may resolve their problem and leave without requiring service, or may deteriorate and become an urgent customer. Similarly, an urgent customer who does not receive service may leave the system, e.g., through adversarial events such as abandonment, or may improve to the moderate class. If we assume there is a classifier (e.g. an early warning system) that classifies potentially risky customers into the moderate class, then the proportion of moderate customers who will actually deteriorate into the urgent class measures the true positive rate of the classifier. Our analysis, which builds on a deterministic fluid approximation of this queueing model, provides insights on how different model parameters affect the optimal scheduling policy for proactive service.

Equilibrium analysis. To minimize the long-run average cost for the fluid model, we show that the decision to prioritize the urgent class versus the moderate class is governed by what we refer to as the modified $c\mu/\theta$-rule. In particular, the corresponding modified $c\mu/\theta$-index accounts for the class-transition dynamics in addition to the holding costs, service rates, and abandonment rates. The exact expression of this index lends itself to a very intuitive interpretation of which parameters – pre or post transition of class types – impact the performance.

Transient optimal control. To minimize the cumulative transient cost (until reaching the equilibrium point with zero queue) for the fluid model, we show that the optimal policy may switch priority depending on the interplay between two indices: the $c\mu$-index and the modified $c\mu/\theta$-index. In particular, it is optimal to schedule according to the modified $c\mu/\theta$-rule when the system state is far away from the equilibrium, and follow the $c\mu$-rule when the state gets close to the equilibrium. Furthermore, if the same class is prioritized by both the $c\mu$-rule and the modified $c\mu/\theta$-rule, then it is optimal to assign strict priority to this class throughout the transient time horizon. On the other hand, if one class is prioritized near the equilibrium and the other is prioritized far away from the equilibrium, then the optimal scheduling policy switches priority at most once along the trajectory. After characterizing the structure of the optimal scheduling policy, calculating the optimal policy curve where priority switches can be done relatively easily. We conduct sensitivity analysis on the policy curve and quantify the effect of prediction accuracy on the optimal scheduling policy.

Our transient analysis also provides a paradigm for solving transient control problems in queues. In particular, the analysis can be summarized by three steps: (i) Approximate the transient dynamics using a proper fluid model; (ii) Derive the structure of the optimal scheduling policy for the fluid model. As the fluid model is a deterministic dynamical system, this step is done utilizing Pontryagin's Minimum Principle and special techniques to deal with state constraints; (iii) Based on the structure of the optimal policy, solve a simpler version of the optimal control problem, i.e. solve for the optimal policy curve.

The rest of the chapter is organized as follows. We conclude this section with a brief review of related literature (Section 1.1.1). The model and detailed problem formulation are introduced in Section 1.2. We derive the optimal scheduling policy to minimize the long-run average cost in Section 1.3, and the optimal scheduling policy to minimize the cumulative transient cost until reaching the equilibrium point in Section 1.4. Section 1.5 considers some model extensions. Lastly, we conclude in Section 1.6. All the proofs are provided in the appendix.

1.1.1 Related Literature

Our work is mainly related to three streams of literature. From the problem context, our problem is related to i) proactive service for managing service systems and ii) scheduling in multi-class queues, especially queues with dynamic class types. From the methodology perspective, our work is related to iii) transient queueing control. In what follows, we briefly review related works in these areas.

Proactive Service. There are a number of works on proactive service in service systems, most of which focus on optimal screening strategies in healthcare. For example, Özekici and Pliska (1991) study the optimal scheduling of inspection in the context of screening for cancerous tumors. They take false positives into account but not the limited resource environment, i.e. they do not consider the externality each patient places on other patients. Örmeci et al. (2015) study the optimal scheduling of screening where the screening service shares resources with the more urgent diagnostic service. They model the benefit of screening through its effect on improving the “environment". Sun et al. (2017) study whether to perform triage under austere conditions, where triage occupies scarce resources but can provide more information on how to prioritize patients. Hu et al. (2018) take an empirical approach to examine the cost and benefit of proactively transferring “risky" patients to the ICU. In various service settings, there are also works modeling proactive service when providers have advance information about customers’ future service needs, but they do not model the dynamic change of customer class types as we do. Examples include Xu and Chan (2016), Yom-Tov et al. (2018), Delana et al. (2019) and Cheng et al. (2019). Our work complements this literature by providing a general modeling framework that takes several key aspects of proactive service into account. These aspects include a limited resource environment, customer deterioration and amelioration, different service needs, and different waiting costs. We also derive structural insights on the optimal scheduling policy for proactive service.

Optimal scheduling of multi-class queues. Our modeling approach falls into the category of multi-class queues. There is a growing literature on optimal scheduling of multi-class queues; see, for example, Mandelbaum and Stolyar (2004); Harrison and Zeevi (2004); Stolyar et al. (2004), and Puha and Ward (2019) for a recent review of works on scheduling multi-class queues with impatient customers. Due to the linear structure in system dynamics, in a lot of cases, a simple index-based policy can be shown to be optimal. For example, the $c\mu$-rule is shown to be optimal for a single server queue without abandonment (Cox and Smith, 1991). The $c\mu/\theta$-rule is shown to be asymptotically optimal for multi-class queues with exponential patience time distribution in the many-server overloaded regime (Atar et al., 2011). We also note that due to the prohibitively large state-space and policy-space for these problems, approximation techniques are often employed to develop structural insights on the optimal policy, (e.g., Van Mieghem (1995); Tezcan and Dai (2010); Gurvich and Whitt (2010)).

The most relevant multi-class queueing models to ours are queues with dynamic class types. Sharing similar motivation to our work, Akan et al. (2012) model the wait list for donated organs as a multi-class overloaded queue. Disease evolution is captured by allowing customers to transition between different classes representing different health levels. Xie et al. (2017) conduct performance analysis for systems where delayed customers may renege the current queue and transfer to a higher-priority class. Cao and Xie (2016) derive the optimal scheduling policy for a single-server two-class model with holding and transferring costs. Down and Lewis (2010) study an N-model in which customers from the class with flexible servers (low-priority) can be upgraded to the one with dedicated servers (high-priority). Most of these works rely on exact or numerical analysis of the corresponding Markov decision process (MDP), where the analysis can become prohibitively challenging when the scale of the system becomes large or more features are added to the model. In this chapter, we adopt a fluid approximation approach, which borrows insights from the conventional heavy-traffic asymptotic analysis under the fluid scaling (Whitt, 2002).

Transient Queueing Control. Analyzing transient queueing dynamics is often very challenging, even without the added complexity of optimizing over different control policies. Only a limited set of numerical and approximation techniques have been developed for transient performance analysis. These include inverting Laplace transforms (Abate and Whitt, 1988, 2006), heavy-traffic asymptotics (Honnappa et al., 2015), etc. Our study uses a fluid approximation and employs tools from the optimal control theory for dynamical systems to derive the optimal transient scheduling policy; see (Sethi and Thompson, 2000; Grass et al., 2008) for an overview of continuous-time control theory and its wide applications. In particular, we utilize Pontryagin's Minimum Principle (Hartl et al., 1995), which is a common methodology used for both linear and nonlinear continuous control problems. The most relevant works to ours are Larrañaga et al. (2013) and Larrañaga (2015), where they consider a multi-class single-server queue with abandonment but static (fixed) class types. Aiming to minimize the cumulative transient holding cost for the fluid approximation, the authors show that the optimal policy may switch priority depending on the interplay between the $c\mu$-index and the $c\mu/\theta$-index. We note that adding the component of dynamic class types is a highly nontrivial extension due to the more complicated boundary behavior (when the state constraints are binding). Moreover, the optimal trajectories in our case cannot be characterized in closed form. We highlight that the analysis laid out in Section 1.4 substantially extends the framework for navigating optimal control problems with state constraints; this approach may shed insights for other queueing control problems.

1.2 The Model

To explore the potential benefits of proactive service, we propose a Markovian two-class multi-server queueing system as depicted in Figure 1.1. Customers (jobs) are defined by their need for service. Without loss of generality, we refer to Class 1 as the *urgent class*: those with immediate need for service. Focusing resource allocation to just these customers is a common approach in the service operations literature. In this work, we also consider a *moderate class* (Class 2): those who currently do not need as high level of service as Class 1, but are at risk of becoming urgent. The novel feature we incorporate is dynamic class types. We allow Class 2 customers to transition to Class 1 while waiting and refer to this as a degradation. *Proactive service* (preventive service), i.e. providing service to Class 2 customers, can prevent Class 2 customers from becoming Class 1 customers. We also allow Class 1 customers to transition to Class 2 while waiting, and refer to this as improvement. Note that, mathematically, our model is symmetric. We differentiate customers as urgent and moderate to better facilitate discussions of real-world applications and derive managerial insights.

These type of dynamics may arise in a lot of service operations applications. For example, in hospitals, Class 1 customers may correspond to patients who are physiologically unstable and in need of care in an Intensive Care Unit (ICU), while Class 2 customers may correspond to patients in the general medical ward who are *at risk* of deteriorating. Those who are in the general medical ward, but are known to have no risk of needing ICU care, would be outside of our modeling framework. Many patients in the general medical ward will never need ICU care, while others may decompensate and be transferred up to the ICU. With improving accuracy of early warning systems, proactive ICU admission *before* a patient is severely critical is becoming a reality (Hu et al., 2018). What remains is to understand when and how such care should be utilized.

Another example is airline call centers following massive flight cancellations, e.g. due to severe weather issues. In this case, urgent customers are those with complicated and urgent travel needs, and thus require immediate assistance from the agents. Moderate customers are those who can either rebook through an agent or rebook online themselves. However, some moderate customers may develop negative emotions while trying to find another flight themselves and may require more service time to satisfactorily address their needs once they have joined the urgent queue for agent assistance (Altman et al., 2019).

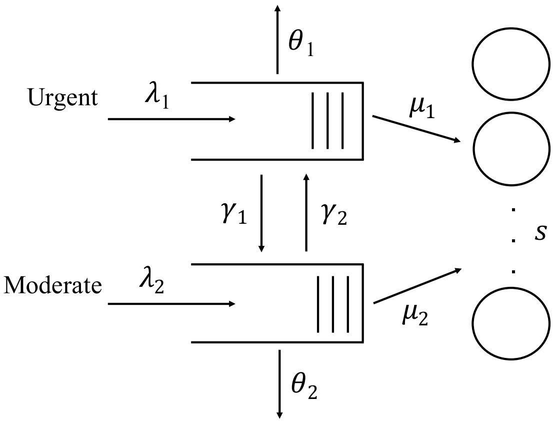
We consider a system with *s* identical servers, i.e. they offer the same quality of service.

Class i customers, $i = 1, 2$, arrive to the system according to a time-homogeneous Poisson process with rate λ_i. Class 1 customers have independent and identically distributed service requirements following an exponential distribution with mean $1/\mu_1$. While waiting to receive service, a Class 1 customer may improve and transition to the Class 2 queue according to an exponentially distributed clock with rate γ_1. A Class 1 customer can also abandon the queue if its waiting time exceeds its patience time. The patience time is exponentially distributed with mean $1/\theta_1$ and is independent of everything else. For Class 1 customers, one can interpret this abandonment as an undesirable event. For example, in the healthcare setting, urgent patients could be placed in an off-service unit, transferred to another hospital, or even die. In a call center setting, customers may abandon and their patronage may be lost.

Class 2 customers can either be proactively served (i.e. before transitioning to Class 1), abandon the system, or deteriorate into Class 1. Should the system administrator choose to provide proactive service to a Class 2 customer, its service time is exponentially distributed with mean $1/\mu_2$. Deterioration and abandonment happen according to two independent exponential clocks with rate γ_2 and θ_2, respectively. For Class 2 customers, one can interpret the abandonment as a desirable outcome. For example, in the healthcare example, the abandonment for moderate patients can be the event that the patient is no longer at risk for deterioration, i.e., the patient self-cures.

Remark 1. *We make two remarks about our modeling assumptions. First, the Markovian assumption on system primitives, including exponential deterioration and upgrade times, is quite common in the literature; see, for example, Down and Lewis (2010); Cao and Xie (2016); Xie et al. (2017). This is in part because the assumption greatly facilitates the theoretical analysis of system dynamics. Second, in practice, it is natural to assume that the service times while in Class 1 and Class 2 for the same customer should be correlated. This can be achieved by assuming that the "base" service requirement for a customer is characterized by a rate* 1 *exponential random variable,* V_0*. When the customer is served* *as Class i, $i = 1,2$, its service time is V_0/μ_i. In this case, we keep the marginal service time distribution in Class i as exponential with rate μ_i while maintaining the order of the service times, e.g., if $\mu_1 < \mu_2$, then $V_0/\mu_1 > V_0/\mu_2$ with probability* 1. *Due to the memoryless property of exponential random variables, introducing such correlation will not affect the dynamics of the system. For simplicity in the subsequent development, we treat these service times as independent random variables.*

Figure 1.1: Two-class queue



We next provide a useful interpretation for the ratio $\phi := \gamma_2/(\theta_2+\gamma_2)$. Note that if no proactive service is provided to Class 2 customers, $\gamma_2/(\theta_2+\gamma_2)$ of them will deteriorate into the urgent class. Suppose Class 2 customers are identified via a classifier that determines customers who are “at risk" of deteriorating (e.g., Escobar et al. (2012)), then $\gamma_2/(\theta_2+\gamma_2)$ can be interpreted as the true positive rate of this classifier. That is, it measures the accuracy of the classifier. For example, if we know with certainty that Class 2 customers will eventually deteriorate into Class 1 customers, then $\theta_2 = 0$ and $\gamma_2/(\theta_2+\gamma_2) = 1$.

To understand the key tradeoffs we are trying to capture with this model, we start by discussing the extreme case where $\gamma_1 = \theta_1 = \theta_2 = 0$. In this case, if no service is provided to Class 2 customers, each Class 2 customer generates an average workload of $\gamma_2/(\mu_1(\theta_2+\gamma_2))$ to the system. This is because $\gamma_2/(\theta_2+\gamma_2)$ of the Class 2 customers will deteriorate into Class 1 and all Class 1 customers must be served. On the other hand, if we can provide proactive service to all Class 2 customers, then each Class 2 customer will generate an average workload of $1/\mu_2$. The magnitude of $\gamma_2/(\theta_2+\gamma_2)$ impacts whether it may be more or less beneficial, from a workload perspective, to provide proactive service to Class 2 customers. Of course, the actual problem we are facing is more complicated than minimizing the system workload. In particular, the different waiting, abandonment, and/or class-transition costs incurred by the two classes can also have a substantial impact on the optimal scheduling policy.

Let $X_i(t)$ denote the number of Class i customers in the system at time t, $t \geq 0$. We denote by $Z_i(t)$ the number of servers assigned to Class i customers, and by $Q_i(t)$ the queue length of Class i at time t. Clearly, $Z_1(t)+Z_2(t) \leq s$ and $X_i(t)-Z_i(t) = Q_i(t) \geq 0$ for $i = 1,2$. We also write $X(t) = (X_1(t),X_2(t))$, $Z(t) = (Z_1(t),Z_2(t))$, and $Q(t) = (Q_1(t),Q(t))$. Note that the state of the system at time t can be described by $(X(t),Q(t))$. A scheduling policy Π is defined as a rule for allocating servers to customers, i.e. Z_i's are the control variables. We consider Markovian policies under which the server allocations are made based on the current state (X,Q) only. In particular, the policy is non-anticipating. Under this class of scheduling policies, which we denote by set $\mathscr{S}$, $\{(X(t),Q(t)) : t \geq 0\}$ forms a Markov process.

As the process $\{(X(t),Q(t)) : t \geq 0\}$ actually depends on the scheduling policy Π, we can more explicitly mark the dependence by writing the stochastic process as $\{(X^{\Pi}(t),Q^{\Pi}(t)) : t \geq 0\}$. We also denote $R_i^{\Pi}(t)$ as the cumulative number of the customers that have abandoned the Class i queue by time t, and $\Gamma_i^{\Pi}(t)$ as the cumulative number of customers that have changed type from Class i to the other by time t. In what follows, we shall drop the superscript Π when it can be understood from the context.

We incur costs for all customers who wait, abandon, or transition classes. In particular, for each Class i customer, we denote h_i as the holding cost per unit time waiting in queue, α_i as the fixed cost of abandonment, and ν_i as the fixed cost of changing class types. Our goal is to minimize the aggregated cost incurred, namely,

$$\mathbb{E}\left[\int_0^T \sum_{i=1,2} h_i Q_i(t)dt + \sum_{i=1,2} (\alpha_i R_i(T) + \nu_i \Gamma_i(T))\right]. \tag{1.1}$$

Note that under the Markovian modeling assumption, we have

$$\mathbb{E}[R_i(T)] = \theta_i \mathbb{E}\left[\int_0^T Q_i(t)dt\right] \quad \text{and} \quad \mathbb{E}[\Gamma_i(T)] = \gamma_i \mathbb{E}\left[\int_0^T Q_i(t)dt\right], \quad i = 1,2.$$

Thus, (1.1) can be equivalently written as

$$\mathbb{E}\left[\int_0^T (c_1 Q_1(t) + c_2 Q_2(t))\, dt\right], \quad \text{where } c_i = h_i + \alpha_i \theta_i + \nu_i \gamma_i \text{ for } i = 1,2.$$

This implies that we can incorporate the abandonment costs and the class-transition costs into the holding costs. In what follows, we shall use c_1 and c_2 to denote the “generalized" holding costs. Note that we defined Class 1 as the *urgent* class in order to facilitate interpretation and draw managerial insights. For example, this can correspond to defining Class 1 customers as those having a higher generalized holding cost, i.e., $c_1 > c_2$.

Remark 2. *Due to our Markovian assumptions and our holding cost criteria, the system performance is agnostic to the order customers are served within a class. The policy development focuses on which class to prioritize; customers within the same class can be served in any order, e.g., first-come-first-served. That said, when looking at individual customers, depending on the transition dynamics and scheduling policies, it is possible that a customer's waiting time may* increase *or* decrease *after changing class. Indeed, our policy development leverages the fact that customers may be able to afford to wait longer after transition (due to the smaller holding cost) and so we can focus resources to the higher priority customers. If we wanted to take waiting-time related fairness into account, we would need to modify our objective function to add some cost of fairness or adapt the optimization problem to incorporate a fairness constraint. Quantifying the fairness of a scheduling policy is an interesting and challenging problem which is outside the scope of this work. We refer to Wierman (2011) for more discussions on the topic.*

In this chapter, we focus on two cost measures. One is the long-run average cost; the other is the cumulated transient cost. The two cost formulations have different focuses and are both relevant in practice. The long-run average cost formulation involves minimizing the cost when the system is in its "normal" state of operation. When shocks bring the system far from its normal state of operation, the transient cost formulation aims to minimize the cost incurred to bring the system back to normal. More precisely, the **long-run average cost minimization problem** is

$$\min_{\Pi \in \mathscr{S}} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\int_0^T \left(c_1 Q_1^\Pi(t) + c_2 Q_2^\Pi(t)\right) dt\right]. \tag{S1}$$

It is significant that the long-run average problem is not capacity specific, namely, the system can be staffed to operate in an underloaded or overloaded regime. For the **cumulated transient cost minimization problem**, we define

$$\mathscr{T} := \inf\{t \geq 0 : Q_1(t) + Q_2(t) = 0\}.$$

That is, $\mathscr{T}$ is the time until the total queue is emptied. We assume that for the transient problem, we have ample capacity such that $\mathbb{E}[\mathscr{T}] < \infty$ for any fixed initial state $(X(0), Q(0)) = (x_0, q_0)$. Then the transient optimization problem can be written as

$$\min_{\Pi \in \mathscr{S}} \mathbb{E}\left[\int_0^{\mathscr{T}} \left(c_1 Q_1^\Pi(t) + c_2 Q_2^\Pi(t)\right) dt\right]. \tag{S2}$$

These cost minimization problems are MDP's. Due to the large (infinite) state-space and policy-space, they are prohibitively hard to solve from a computational standpoint. Even if we solve it numerically, limited insights about the optimal policy can be gained. Various approximation techniques have been developed in the literature to solve large-scale MDPs. With the goal of gaining structural insights into the optimal scheduling policy, we employ a fluid approximation approach; a similar method has been used in, for example, Whitt (2006a); Perry and Whitt (2009); Atar et al. (2010).

1.2.1 The Fluid Model

To construct the fluid model, we replace the stochastic arrival, service, abandonment and class-transition processes by their corresponding deterministic flow rates. We use the lowercase q to denote the fluid queue length process, and a fluid scheduling policy π specifies the service capacity allocation process (z_1, z_2). Under π, the fluid dynamics take the form

$$\begin{aligned} \dot{q}_1(t) &= \lambda_1 - z_1(t)\mu_1 - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\ \dot{q}_2(t) &= \lambda_2 - z_2(t)\mu_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t). \end{aligned} \tag{1.2}$$

Let $\mathscr{F}$ denote the set of fluid admissible scheduling policies. We say that a policy belongs to $\mathscr{F}$ if the server allocation only depends on the current state of the system (Markovian), and satisfies the following constraints:

$$\begin{aligned} & z_i(t) \geq 0, \quad i = 1, 2, \ t \geq 0 \\ & z_1(t) + z_2(t) \leq s, \quad t \geq 0 \\ & \dot{q}_i(t) \geq 0 \text{ whenever } q_i(t) = 0, \quad i = 1, 2, \ t \geq 0. \end{aligned} \tag{1.3}$$

The first and second constraints in (1.3) require that a non-negative amount of service capacity is assigned to each class, and the total amount of allocated resource does not exceed service capacity. The third constraint guarantees that the resulting queue length process $q_i(t)$ is non-negative for all $t \geq 0$. Note that the queue length process $\{q(t) : t \geq 0\}$ actually depends on the scheduling policy π. We can more explicitly mark the dependence by writing it as $\{q^\pi(t) : t \geq 0\}$. To keep the notation concise, we shall drop the superscript when it can be understood from the context.

We comment that the fluid dynamics capture the mean dynamics of the stochastic system well, as we will demonstrate later with numerical experiments. In addition, this type of fluid model often arises in the literature as the functional law of large numbers limit for a sequence of properly scaled stochastic systems under the conventional heavy traffic scaling (Whitt, 2002; Reed and Ward, 2008). In this limiting regime, we scale up the arrival rates and the service rates while scale down the space (Alternatively, we can scale up time while scale down the abandonment rates, the class-transition rates, and the space). The number of servers is held fixed[1].

1.2.2 Problem Formulation

In this section, we introduce the fluid counterparts of the stochastic cost minimization problems. Note that for the long-run average optimization problem, we only require that the amount of service capacity is non-negative, $s \geq 0$.

Fluid long-run average cost optimization problem:

$$\min_{\pi \in \mathscr{F}} \limsup_{T \to \infty} \frac{1}{T} \int_0^T (c_1 q_1^\pi(t) + c_2 q_2^\pi(t))\, dt. \tag{F1}$$

For the transient optimization problem, let $\tau := \inf\{t \geq 0 : q_1(t) + q_2(t) = 0\}$, which is the first time when the total fluid queue reduces to 0. We assume that there is ample capacity s such that for any $q(0) = q_0$, $\tau < \infty$. As will be explained in Section 1.4, the precise condition is $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$.

Fluid transient optimization problem:

$$\min_{\pi \in \mathscr{F}} \int_0^\tau (c_1 q_1^\pi(t) + c_2 q_2^\pi(t))\, dt. \tag{F2}$$

Our analysis relies on understanding the long-run regularity of the fluid model. We thus provide the following definition.

Definition 1. *Consider the autonomous dynamical system $\dot{q}(t) = f(q(t))$ with $q(0) = q_0$. Suppose f has an equilibrium point q_e, i.e. $f(q_e) = 0$. Let $||\cdot||$ be the Euclidean norm in $\mathbb{R}^2$. Then*

(1) q_e is **locally asymptotically stable** *if there exists $\delta > 0$, such that if $||q_0 - q_e|| < \delta$, then $\lim_{t\to\infty} ||q(t) - q_e|| = 0$.*

(2) q_e is **globally asymptotically stable** *if for any initial condition q_0, $\lim_{t\to\infty} ||q(t) - q_e|| = 0$.*

[1] In particular, we do not scale up the number of servers as in the many-server heavy traffic regime.

We shall start by solving the long-run average cost minimization problem (F1) in Section 1.3. We then solve the transient cost minimization problem (F2) in Section 1.4.

1.3 Optimal Long-Run Scheduling Policy

In this section, we solve the fluid long-run average cost minimization problem. To ensure system stability for any arrival rates and service capacity, we impose the following assumption on the abandonment and class-transition rates.

Assumption 1. *(i)* $\theta_1 + \gamma_1 \theta_2 > 0$ *and* $\theta_2 + \gamma_2 \theta_1 > 0$. *(ii)* $\frac{1}{\mu_2} \neq \frac{1}{\mu_1} \frac{\gamma_2}{\theta_2 + \gamma_2}$ *and* $\frac{1}{\mu_1} \neq \frac{1}{\mu_2} \frac{\gamma_1}{\theta_1 + \gamma_1}$.

Part (i) of Assumption 1 requires that the system has the “necessary" abandonment for stability even with no service. For example, if the abandonment rate from Class 1 is zero ($\theta_1 = 0$), then a Class 1 customer can leave the system by converting to Class 2 and eventually abandoning the Class 2 queue ($\gamma_1 \theta_2 > 0$). Part (ii) of the assumption requires that there is a workload difference based on when (i.e. before versus after class-transitions occur) service is provided. This imposes a tradeoff when deciding which class to prioritize (see, Appendix A.1.1 for more details).

The long-run average cost minimization problem can be explicitly written as

$$
\begin{aligned}
\min_{z_1, z_2, q_1, q_2} \quad & \limsup_{T \to \infty} \frac{1}{T} \int_0^T (c_1 q_1(t) + c_2 q_2(t))\, dt \\
s.t. \quad & \dot{q}_1(t) = \lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\
& \dot{q}_2(t) = \lambda_2 - \mu_2 z_2(t) - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t) \\
& z_1(t) + z_2(t) \leq s, \quad t \geq 0 \\
& z_1(t), z_2(t), q_1(t), q_2(t) \geq 0, \quad t \geq 0.
\end{aligned}
$$

This is an infinite dimensional linear program (LP). We first make an important observation that allows us to reformulate the problem as a finite dimensional LP. This observation will be made rigorous in Theorem 1. If the fluid dynamical system converges to an equilibrium point as $t \to \infty$, then minimizing the long-run average cost can be reformulated as finding the optimal equilibrium point. In particular, we have the following alternative problem formulation.

$$\begin{aligned}
\min_{z_1^e, z_2^e, q_1^e, q_2^e} \quad & c_1 q_1^e + c_2 q_2^e \\
s.t. \quad & \lambda_1 - \mu_1 z_1^e - \theta_1 q_1^e - \gamma_1 q_1^e + \gamma_2 q_2^e = 0 \\
& \lambda_2 - \mu_2 z_2^e - \theta_2 q_2^e - \gamma_2 q_2^e + \gamma_1 q_1^e = 0 \\
& z_1^e + z_2^2 \leq s \\
& z_1^e, z_2^e, q_1^e, q_2^e \geq 0.
\end{aligned} \tag{1.4}$$

Note that the first two constraints in (1.4) characterize the equilibrium point: rate-in equals rate-out. By rearranging (1.4), we have an equivalent optimization problem:

$$\begin{aligned}
\max_{z_1^e, z_2^e} \quad & \left(\frac{c_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\gamma_2 + \theta_2}} + \frac{c_2 \frac{\gamma_1}{\gamma_1 + \theta_1}}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1 + \theta_1}} \right) \mu_1 z_1^e + \left(\frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1 + \theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2 + \gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2 + \gamma_2}} \right) \mu_2 z_2^e \\
s.t. \quad & z_1^e + z_2^e \leq s \\
& \frac{(\theta_2 + \gamma_2)\lambda_1 + \gamma_2 \lambda_2}{(\theta_2 + \gamma_2)\theta_1 + \gamma_1 \theta_2} - \frac{(\theta_2 + \gamma_2)\mu_1}{(\theta_2 + \gamma_2)\theta_1 + \gamma_1 \theta_2} z_1^e - \frac{\gamma_2 \mu_2}{(\theta_2 + \gamma_2)\theta_1 + \gamma_1 \theta_2} z_2^e \geq 0 \\
& \frac{(\theta_1 + \gamma_1)\lambda_2 + \gamma_1 \lambda_1}{(\theta_1 + \gamma_1)\theta_2 + \gamma_2 \theta_1} - \frac{(\theta_1 + \gamma_1)\mu_2}{(\theta_1 + \gamma_1)\theta_2 + \gamma_2 \theta_1} z_2^e - \frac{\gamma_1 \mu_1}{(\theta_1 + \gamma_1)\theta_2 + \gamma_2 \theta_1} z_1^e \geq 0 \\
& z_1^e, z_2^e \geq 0.
\end{aligned} \tag{1.5}$$

It is easy to see that the optimal solution to (1.5) tends to assign a larger value to the z_i^e with a larger coefficient in the objective function. Motivated by this observation, we define *the modified $c\mu/\theta$-index* as follows: **The modified $c\mu/\theta$ index for Class 1** is

$$r_1 := \left(\frac{c_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\gamma_2 + \theta_2}} + \frac{c_2 \frac{\gamma_1}{\gamma_1 + \theta_1}}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1 + \theta_1}} \right) \mu_1, \tag{1.6}$$

and **the modified $c\mu/\theta$ index for Class 2** is

$$r_2 := \left(\frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1 + \theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2 + \gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2 + \gamma_2}} \right) \mu_2. \tag{1.7}$$

From (1.6) and (1.7), we observe that when $\gamma_i = 0$ we recover the standard $c\mu/\theta$-index (Atar et al., 2010). When $\gamma_i \neq 0$, the extra terms are to account for the class-transition dynamics. To interpret the index r_1 in (1.6) (r_2 in (1.7) follows by symmetry), we note that the first term corresponds to the standard $c\mu/\theta$-structure for Class 1 customers. In particular, these customers incur a cost at rate c_1. The effective abandonment rate is $\theta_1 + \gamma_1\theta_2/(\theta_2+\gamma_2)$. This is because “abandonment" in this case consists of the nominal abandonment, which happens at rate θ_1, as well as the improvement. The improvement happens at rate γ_1, but we also have to adjust for the fact that $\gamma_2/(\gamma_2+\theta_2)$ of those customers may deteriorate and transition back to Class 1. Thus, the net improvement rate is $\gamma_1\theta_2/(\gamma_2+\theta_2)$. The second term takes into account the Class 1 customers who improve to Class 2. These customers will incur a cost at rate c_2 when in Class 2. Because the proportion of Class 1 customers who improve to Class 2 is $\gamma_1/(\theta_1+\gamma_1)$, the expected cost rate is $c_2\gamma_1/(\theta_1+\gamma_1)$. When in Class 2, these customers abandon at rate θ_2, and deteriorate at rate γ_2 with a feedback probability $\gamma_1/(\theta_1+\gamma_1)$.

Formally, we have the following theorem characterizing the optimal scheduling policy based on the modified $c\mu/\theta$-index.

Theorem 1. *Under Assumption 1, giving strict priority to the class with a higher modified $c\mu/\theta$-index minimizes the long-run average cost* (F1)*. That is, if $r_1 \geq r_2$, for r_1, r_2 defined in* (1.6) *and* (1.7)*, then it is optimal to give strict priority to Class 1. Otherwise, it is optimal to give strict priority to Class 2.*

To prove Theorem 1, we need to ensure that the fluid dynamical system converges to the desired equilibrium point under the strict priority rule implied by the modified $c\mu/\theta$-index. We provide detailed analysis on the long-run regularity of the fluid model under the strict priority rules in Appendix A.1. These convergence analyses are interesting in their own right, as they reveal important characteristics of the system dynamics. Moreover, we show that an interesting *bi-stability* phenomenon, i.e. the presence of two equilibria, can arise when the sub-optimal strict priority rule is employed. We provide more discussions about this phenomenon in Section 1.3.1.

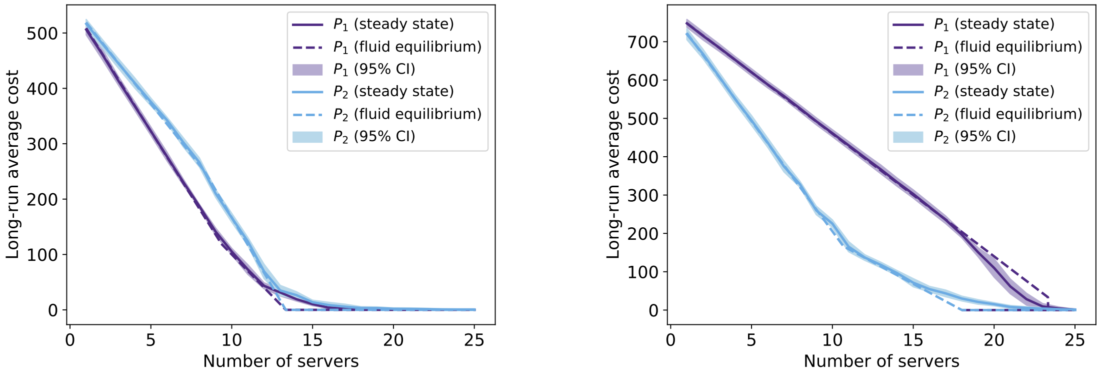
We next numerically compare the long-run average costs of the fluid models to those of the corresponding stochastic systems under different strict priority rules. We denote P_1 as strict priority to Class 1 and P_2 as strict priority to Class 2[2]. Figure 1.2 plots the long-run average costs for systems with different numbers of servers s. The fluid costs are plotted in dashed lines while the costs for the stochastic systems are plotted in solid lines. As the long-run average costs for the stochastic systems are estimated using simulation, we also provide the corresponding 95% confidence interval. Figure 1.2a illustrates the scenario where the modified $c\mu/\theta$-index suggests prioritizing Class 1, while Figure 1.2b has the modified $c\mu/\theta$-index suggesting prioritizing Class 2. We first note that the long-run average fluid cost approximates the long-run average cost of the stochastic system reasonably well, especially when s is small (the system is in the so-called overloaded regime) and when s is large (the system is in the so-called underloaded regime). Second, we observe that when comparing the strict priority rules, prioritizing the class with a larger modified $c\mu/\theta$-index always leads to a lower cost in the stochastic system. Thus, even when the cost of fluid system may deviate from that of the corresponding stochastic system, the resulting policy recommendations are consistent. Lastly, we note that in Figure 1.2b, when $18 \leq s \leq 22$, the fluid model under strict priority to Class 1 has two different equilibria (bi-stability). Which equilibrium the fluid system converges to depends on its initial condition. For the corresponding stochastic system, it will fluctuate around one equilibrium point for a while before transitioning to the region around the other equilibrium point. Thus, the corresponding long-run average cost is a weighted average of the costs around the two equilibria.

1.3.1 Bi-Stability

Due to the dynamic class types, applying the strict priority rule that does not agree with the modified $c\mu/\theta$-index can lead to a bi-stability phenomenon. Motivated by proactive

[2] Throughout this manuscript, all numerical experiments for the stochastic system are conducted with preemption, though we emphasize this has no impact on the fluid analysis.

Figure 1.2: Optimal long-run scheduling policy
((a): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.5, \mu_2 = 3, \gamma_1 = 0.1, \gamma_2 = 0.1, \theta_1 = 0.1, \theta_2 = 0.4, c_1 = 5, c_2 = 3$
(b): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$)



(a) The modified $c\mu/\theta$-rule: P_1

(b) The modified $c\mu/\theta$-rule: P_2

service applications, in this section, we study in more depth a special case where bi-stability arises. Specifically, we consider the system parameters for which Theorem 7 in Appendix A.1.1 suggests that if we prioritize the urgent class, the system exhibits *bi-stability*. While, in this parameter regime, following the modified $c\mu/\theta$-rule is the optimal policy, from a practical standpoint, the service provider may prefer to give priority to the urgent class, as long as it does not degrade system performance. We explore whether it is reasonable to (sometimes) give priority to the urgent class even though one of the optimal long-run average policies indicates priority should be given to the moderate class.

The parameter regime we are interested in is when the urgent class (Class 1) has a higher $c\mu$-index, i.e., $c_1\mu_1 > c_2\mu_2$, but $\mu_1 < \frac{\gamma_2}{\theta_2+\gamma_2}\mu_2$, which implies that the moderate class (Class 2) has a higher modified $c\mu/\theta$-index, i.e., $r_2 \geq r_1$. In this case, from the workload perspective, it is more efficient to serve moderate customers before they deteriorate, i.e.,

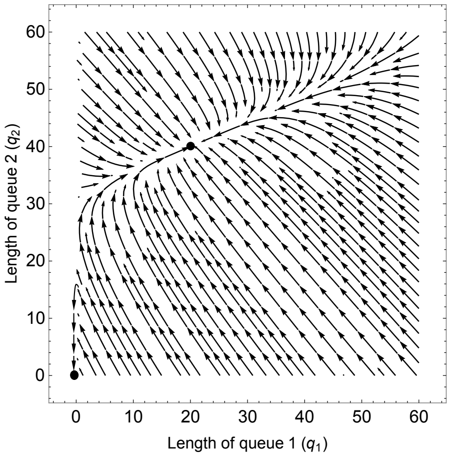
$$\frac{1}{\mu_2} < \frac{\gamma_2}{\gamma_2 + \theta_2}\frac{1}{\mu_1}.$$

Additionally, the capacity is in the critical region

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2 + \gamma_2}\frac{\lambda_2}{\mu_1}. \tag{1.8}$$

Figure 1.3 provides an illustration of the vector field under bi-stability. We note that there are two locally asymptotically stable equilibrium points. Which equilibrium point the queue process converges to depends on its initialization.

Figure 1.3: Vector field under bi-stability
($\lambda_1 = 10, \lambda_2 = 20, s = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2$)



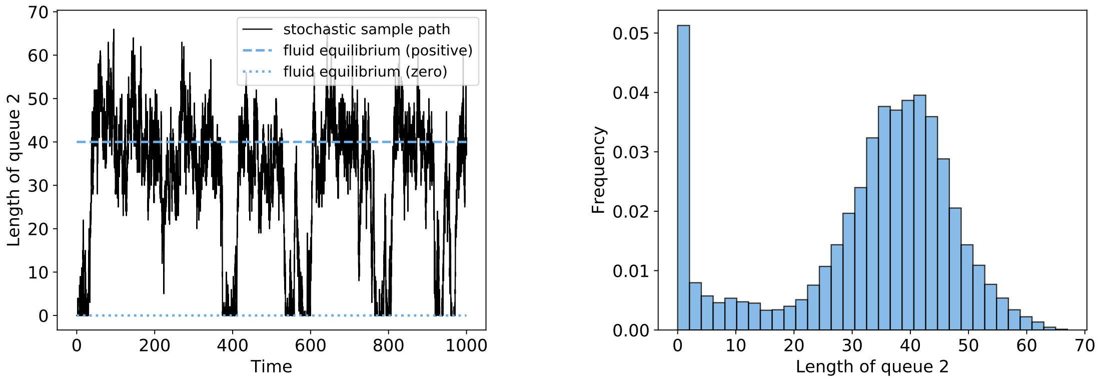
Intuitively, the bi-stability arises because if we delay service for moderate customers, they will end up generating more workload on average when they deteriorate into the urgent class. When the system is critically loaded as in (1.8), even though we have enough capacity to serve both classes when service is provided in a timely manner, i.e., $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, we do not have enough capacity to serve all the customers when service for Class 2 is delayed, i.e.,

$$s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2 + \gamma_2}\frac{\lambda_2}{\mu_1}.$$

Under bi-stability, we note that one of the equilibrium points leads to very good performance – zero holding cost, while the other equilibrium point has positive queues for at least one class (Figure 1.3 and Theorem 7). Ideally, we want to avoid the “bad" equilibrium regardless of where we start. One way to ensure global convergence to the “good" equilibrium is to switch priority to the moderate class as suggested by the modified $c\mu/\theta$-rule. However, there are many systems where it may be preferable to give priority to the urgent class for obvious administrative reasons. Thus, we propose an alternative intervention, which we refer to as the bi-stability control. For a fixed threshold $\alpha_0 > 0$, when $\bar{q}_1(t) + \bar{q}_2(t) \leq \alpha_0$, we prioritize the urgent class; otherwise, we prioritize the moderate class. Following a similar Lyapunov argument as in Appendix A.1.1, when α_0 is sufficiently small, $q(t)$ will converge to $(0,0)$ regardless of its initialization q_0, i.e., $(0,0)$ is a globally asymptotically stable equilibrium under this control. As such, both the modified $c\mu/\theta$ rule and the bi-stability control with properly chosen threshold lead to the same optimal long-run average cost in this case. However, when studying the transient cost, i.e., the cost incurred to restore system to zero when it is initialized far from zero, the bi-stability control can lead to a lower cost than the modified $c\mu/\theta$ rule as we will explain in Section 1.4.

We next elaborate on the implications of the fluid bi-stability phenomenon for the stochastic system. When bi-stability arises in the fluid system, the queue length process of the corresponding stochastic system will fluctuate around one equilibrium for a while before transitioning to the region around the other equilibrium. Figure 1.4a shows a typical sample path of the stochastic queue length process, i.e., we plot $Q_2(t)$ for $t \in [0,1000]$. Figure 1.4b provides the histogram of $Q_2(t)$. We observe that it follows a bi-modal distribution where the two peaks are around the two fluid equilibria.

Figure 1.4: Bi-stability in the stochastic system
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, s = 20$)

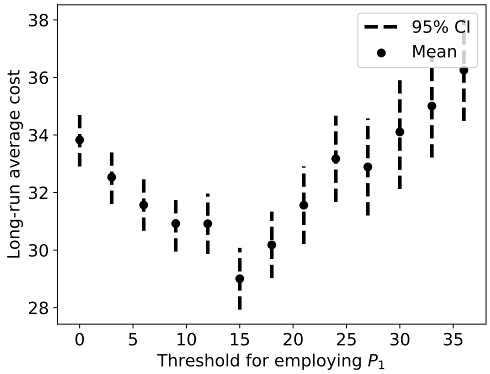


(a) Sample path of $Q_2(t)$

(b) Histogram of $Q_2(t)$

Figure 1.5 plots the long-run average cost for the stochastic system under the bi-stability control for different values of α_0 (point estimates together with the corresponding 95% confidence intervals). In the stochastic system, if $Q_1(t) + Q_2(t) \leq \alpha_0$, we prioritize the urgent class. Otherwise, we prioritize the moderate class. Note that $\alpha_0 = 0$ is equivalent to assigning strict priority to the moderate class, i.e., the modified $c\mu/\theta$-rule. Interestingly, we observe that for certain values of α_0, the bi-stability control achieves a smaller long-run average cost than the modified $c\mu/\theta$-rule. As surprising as the observation may seem at first glance, this phenomenon is due to stochastic fluctuations that bring the system away from the equilibrium, i.e., zero queue. To restore the system to zero in the most cost-effective way, the experiments suggest that we should prioritize the moderate class when the queues are large, and prioritize the urgent class when the queues are small. We explore this more formally in our transient analysis in Section 1.4.

Figure 1.5: Long-run average cost under the bi-stability control
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.2, \theta_1 = 0.1, \theta_2 = 0.2, s = 19, c_1 = 10, c_2 = 1$)



1.4 Optimal Transient Scheduling Policy

Service systems often operate in highly non-stationary environments. In healthcare settings, for example, random shocks like disease outbreaks or mass casualty events can push the system far from its normal state, i.e., equilibrium. When such a demand shock happens, the key question we wish to address is how to bring the system back to its normal state of operation in the most cost-effective way. In this section, we study the transient optimal control problem (F2) to find the optimal clearing of backlogs. In particular, we derive the optimal scheduling policy to help the system recover from demand shocks.

We start by focusing on the after-shock control. In particular, we assume $q(0) = q_0 > 0$

[3] but we now have abundant capacity to bring the fluid queues to zero in a finite amount of time under some admissible control. In particular, we make the following assumption on the capacity s.

Assumption 2. $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$.

Later in Section 1.5.1, we generalize the arrival-rate pattern to include the period of demand shock in our planning horizon. In particular, the shock raises the arrival rates for a fixed amount of time, during which the service capacity is insufficient and so the backlogs increase. Importantly, the arrival rates during the demand shock period can be time-varying and violate Assumption 2. After the initial shock, the arrival rates restore to normal and satisfy Assumption 2. It is significant that the structure of the optimal control does not change under this more general arrival-rate model (see Theorem 3). The after-shock control studied in this section builds the basis for the cases with more general arrival rates.

Recall that $\tau = \inf\{t \geq 0 : q_1(t) + q_2(t) = 0\}$ is the first time both of the fluid queues are emptied. Based on Theorem 7, Assumption 2 implies that there exists a scheduling policy π, under which, for any $q(0) = q_0 > 0$, $\tau < \infty$. We also note from our long-run regularity analysis in Appendix A.1.1 that under Assumption 2, both strict priority to Class 1 and strict priority to Class 2 lead to the same long-run average holding cost – zero. However, our following analysis will reveal important differences in their transient performance.

The optimal transient scheduling policy depends on the interplay between two index rules. We define the *$c\mu$-rule* as a policy that prioritizes the class with a higher $c_i\mu_i$ value, $i = 1, 2$, i.e., the $c\mu$-index. Similarly, the *modified $c\mu/\theta$-rule* is a policy that prioritizes the class with a higher r_i value, $i = 1, 2$, i.e., the modified $c\mu/\theta$-index as defined in (1.6) and (1.7). To capture the differential effect of each of these rules, we impose the following assumption on the indices.

[3] We define a vector $a > 0$ if all its components are nonnegative and there is at least one component that is strictly positive.

Assumption 3. *$c_1\mu_1 \neq c_2\mu_2$ and $r_1 \neq r_2$.*

We next introduce a few more notations to simplify the presentation of the problem. From the fluid dynamics (1.2), we define $f(q,z) = (f_1(q,z), f_2(q,z))$ where $f_1(q,z) = \lambda_1 - z_1\mu_1 - \theta_1 q_1 - \gamma_1 q_1 + \gamma_2 q_2$ and $f_2(q,z) = \lambda_2 - z_2\mu_2 - \theta_2 q_2 - \gamma_2 q_2 + \gamma_1 q_1$. From the constraints on the admissible controls (1.3), we define $g(q) = (g_1(q), g_2(q))$, where $g_i(q) = -q_i$, for $i = 1,2$, and $h(z) = (h_1(z), h_2(z), h_3(z))$, where $h_1(z) = z_1 + z_2 - s$, $h_2(z) = -z_1$, and $h_3(z) = -z_2$. We also define $F(q) = c_1 q_1 + c_2 q_2$. Then the transient optimal control problem can be explicitly written as:

$$\begin{aligned} \min_{z} \quad & \int_0^\tau F(q(t))\,dt \\ s.t. \quad & \dot{q}(t) = f(q(t), z(t)) \\ & g(q(t)) \le 0 \\ & h(z(t)) \le 0. \end{aligned} \tag{F2$'$}$$

In optimal control theory, optimization problems of the form (F2′) are referred to as *optimal control with state constraints*. Despite a rich body of literature in optimal control, problems with state constraints are, in general, very difficult to solve explicitly as they impose extra boundary conditions (Trélat, 2012). While some results can be derived in special cases, there is no systematic way to deal with these problems; we refer to the survey paper Hartl et al. (1995) for an overview.

We combine several techniques from optimal control theory to derive the optimal transient control. Our solution strategy is to first derive the structure of the optimal scheduling policy. In particular, as we shall explain in Theorem 2, the optimal scheduling policy switches priority at most once and priorities can be characterized by two simple index rules. Then solving for the optimal scheduling policy reduces to finding the policy curve that governs where in the state space the switch in priority happens. We provide a closed form characterization of the policy curve in Proposition 4 for a special case, and provide an efficient numerical scheme to construct the policy curve for the other cases.

The next theorem characterizes the structure of the transient optimal scheduling policy. Let τ^* denote the time to empty the queue under the optimal policy.

Theorem 2. *Under Assumptions 1, 2, and 3, for the transient optimal control problem* (F2$'$)*:*

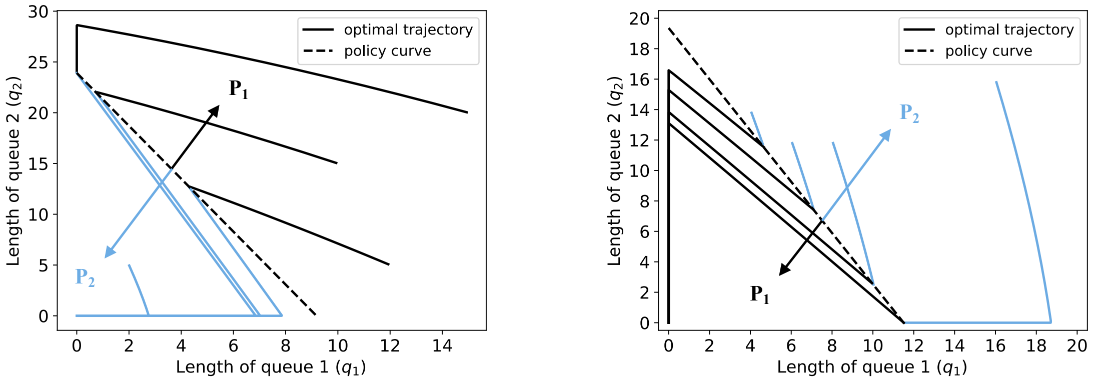
I. If the $c\mu$-rule and the modified $c\mu/\theta$-rule both prioritize Class i, $i = 1,2$, then the strict priority rule to Class i is optimal for any $t \in [0,\tau^*]$.

II. If the $c\mu$-rule prioritizes Class i but the modified $c\mu/\theta$-rule prioritizes Class j, for $i \neq j$, $i, j = 1,2$, then there exist positive real numbers ε and M with $0 < \varepsilon < M$, such that it is optimal to prioritize Class i when $q_1(t)+q_2(t) < \varepsilon$ and prioritize Class j when $q_1(t)+q_2(t) > M$. Furthermore, the optimal scheduling policy switches priority at most once over the transient time horizon $[0,\tau^*]$.

Based on Theorem 2, if the $c\mu$-rule and the modified $c\mu/\theta$-rule agree with each other, it is optimal to give strict priority to the class with a higher $c\mu$-index (and correspondingly a higher modified $c\mu/\theta$-index) for any $q \in \mathbb{R}^2_+$. If the two index rules do not agree, we will follow the $c\mu$-rule when we are close enough to the equilibrium point $(0,0)$; when we are far from the equilibrium point, we should follow the modified $c\mu/\theta$-rule. Moreover, in this case, we switch priority at most once, and the time at which the switch occurs depends on the value of q_0. This indicates that there exists a *policy curve* $\{q : u(q) = 0\}$, where we switch from the modified $c\mu/\theta$-rule to the $c\mu$-rule.

The remaining task is to characterize the policy curve. In Figure 1.6, we provide a numerical illustration of the optimal trajectory of the queue length process. Figure 1.6a shows the case where the modified $c\mu/\theta$-rule prioritizes Class 1 while the $c\mu$-rule prioritizes Class 2. We plot four optimal fluid trajectories starting from different initial values (derived by solving the a discretized version of (F2$'$)). We also plot the corresponding policy curve (dashed line). Figure 1.6b shows the case where the modified $c\mu/\theta$-rule prioritizes Class 2 while the $c\mu$-rule prioritizes Class 1. We will provide more discussions about the policy curve in Section 1.4.3.3.

Figure 1.6: Optimal transient queue length trajectory
((a): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.5, \mu_2 = 3, \gamma_1 = 0.1, \gamma_2 = 0.1, \theta_1 = 0.1, \theta_2 = 0.4, s = 17, c_1 = 5, c_2 = 3$
(b): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, s = 26, c_1 = 5, c_2 = 1$)



(a) The $c\mu$-rule: P_2, the modified $c\mu/\theta$-rule: P_1

(b) The $c\mu$-rule: P_1, the modified $c\mu/\theta$-rule: P_2

Remark 3. *Even though Theorem 2 is stated under Assumption 1, following the same lines of analysis, we can show that if $\theta_1 = \theta_2 = \gamma_1 = \gamma_2 = 0$, we can recover the well-known optimality of the $c\mu$-rule throughout the transient time horizon (see Corollary 3 in Appendix A.4). Furthermore, if $\gamma_1 = \gamma_2 = 0$ but $\theta_1, \theta_2 > 0$, we should follow the $c\mu$-rule when we are close to the origin and the ordinary $c\mu/\theta$-rule when we are far from the origin (This is a special case of Theorem 2). In this case, we recover the results in Larrañaga (2015). Nevertheless, the approaches utilized in literature to study the special cases are not directly generalizable to our setting with dynamic class types.*

We next provide the general strategy of proving Theorem 2. It includes three main parts. We first provide some formal definitions to describe the boundary behavior and rule out some “irregular" behaviors in Section 1.4.1. We then establish the optimal scheduling policy when $q_1 + q_2 < \varepsilon$ for ε sufficiently small in Section 1.4.2. This is done by solving the optimal control problem directly. Lastly, we establish the optimal scheduling policy for the rest of the state space in Section 1.4.3, utilizing Pontryagin's Minimum Principle. We believe this framework can be applied to derive the structure of the optimal policy for other transient control problems for queues.

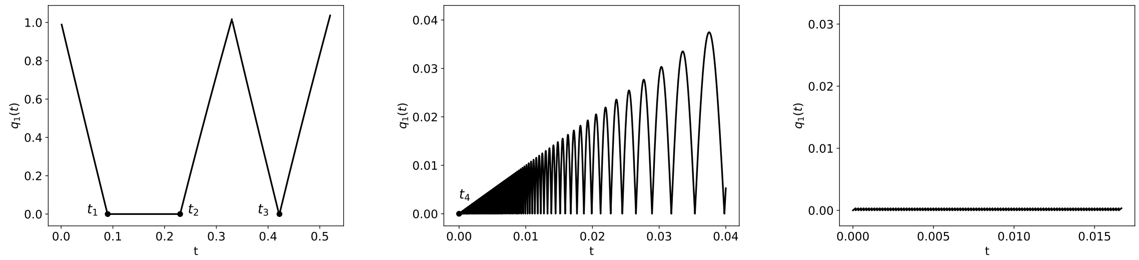
1.4.1 Boundary Behavior

The main challenge in dealing with an optimal control problem of the form (F2$'$) is to characterize the system behavior on the boundary where the state constraints hold tight. In our case, the state constraint $g(q(t)) = -q(t) \leq 0$ requires the queue length process $q(t)$ to stay non-negative for all $t \in [0, \tau]$.

To characterize the boundary behavior, we would ideally like to identify when the trajectory enters the boundary and when it exits the boundary. In particular, we would like to characterize the time points t_k's when $g_i(q(t_k)) = 0$ for some $i = 1,2$, but for any $\delta > 0$, there exists $t \in (t_k - \delta, t_k + \delta)$ such that $g_i(q(t)) > 0$. An important class of points of this type is known as the *junction time* (Hartl et al., 1995). We next provide some formal definitions to characterize the junction times. An interval $\mathscr{I} := [t_1, t_2] \subseteq [0, \tau]$ (or $[t_1, t_2)$, $(t_1, t_2]$, (t_1, t_2)) is called an *interior arc* if $g(q(t)) < 0$ holds for all $t \in \mathscr{I}$. Correspondingly, an interval $\mathscr{I} := [t_1, t_2] \subseteq [0, \tau]$ (or $[t_1, t_2)$, $(t_1, t_2]$, (t_1, t_2)) is called a *boundary arc* if $g_i(q(t)) = 0$, for some $i = 1,2$, holds for all $t \in \mathscr{I}$. A time instant t_1 is called an *entry time* if an interior interval ends at and a boundary interval starts at t_1. A time instant t_2 is called an *exit time* if a boundary interval ends and an interior interval starts at t_2. Furthermore, if the trajectory of $q_i(t)$, $i = 1,2$, only “touches" the boundary at time t_3, i.e., $q_i(t_3) = 0$, but there exists $\delta > 0$ such that $q_i(t) > 0$ for any $t \in (t_3 - \delta, t_3 + \delta)$ and $t \neq t_3$, then t_3 is called a *contact point*. Entry, exit, and contact times taken together are called *junction times*. Figure 1.7a provides a pictorial illustration of different types of junction times for $q_1(t)$. In particular, t_1, t_2, and t_3 in Figure 1.7a are an entry, exit, and contact point respectively. In addition, the interval $[t_1, t_2]$ is a boundary arc, and the interval $[0, t_1)$ is an interior arc.

Not all boundary trajectories can be characterized by the junction times. A class of boundary behaviors that is often hard to deal with is known as chattering, which happens when the trajectory $q_i(t)$, $i = 1,2$, oscillates between zero and positive values infinitely fast. Specifically, a time instant t_4 is said to be a *chattering point* of the state trajectory q_i, if $q_i(t_4) = 0$, and for any $\delta > 0$ there exists s' and $s'' \in (t_4 - \delta, t_4 + \delta)$ such that $q_i(s') > 0$ and $q_i(s'') = 0$. In addition, an interval is said to be a *chattering interval* if any sub-interval of it contains at least one chattering point. Figure 1.7b provides an example where the state trajectory has a chattering point t_4, and Figure 1.7c provides an example of a chattering interval.

Figure 1.7: Different types of junction times and chattering behavior



(a) Entry/exit/contact point

(b) Chattering point

(c) Chattering interval

Chattering behavior can arise in many different optimal control problems. One classical example is Fuller's problem (Fuller, 1963). Noticeably, for non-constrained linear control problems with compact polyhedral control space, it has been shown that there always exists an optimal solution that switches finitely many times among the vertices of the control polyhedron; see, for example, Chapter 2.8 in Schättler and Ledzewicz (2012). However, the pathological situation of chattering has not been ruled out for linear systems with state constraints, which is the case of our problem (F2$'$). We overcome the difficulty here by showing that for (F2$'$), it is without loss of optimality to consider trajectories without chattering points or chattering intervals.

Lemma 1. *For the transient optimal control problem* (F2$'$)*, it is without loss of optimality to consider state trajectories without chattering behavior.*

1.4.2 The $c\mu$-Rule Near the Origin

When the state is close enough to the origin $(0,0)$, which is also an equilibrium point for the fluid system under Assumption 2 and an appropriate control, we establish that the $c\mu$-rule is optimal.

Proposition 1. *Under Assumptions 1, 2, and 3, for the transient optimal control problem* (F2$'$)*, if $q_1(t), q_2(t) \in [0, \varepsilon)$, with $\varepsilon > 0$ sufficiently small, then the $c\mu$-rule is optimal on the transient time interval* $[t, \tau^*]$.

The result in Proposition 1 is derived based on the observation that when the queue length is sufficiently small, the dominant dynamic for the system comes from service completion, which has an order ε effect. The effect of abandonment and class-transition is only second-order, namely, order ε^2. Focusing on service completion only, $c_i\mu_i$ is the rate at which we can reduce the holding cost per unit time and per unit capacity allocated to serving Class i jobs, $i = 1, 2$. In order to reduce holding cost as fast as possible, the class with a larger $c\mu$-index should be prioritized.

1.4.3 The Optimal Policy for the Rest of the State Space

When the states are far away from the origin, we have to take abandonment and class-transition into account, and these substantially complicate the analysis. To develop structural insights in this region, we utilize a necessary characterization for the optimal solution to the control problem, which is known as Pontryagin's Minimum Principle (Hartl et al., 1995).

To understand the underlying mechanism, we first note that if we view the optimal control problem (F2$'$) as an infinite dimensional linear program, then we can write down its dual problem and study the optimal primal-dual structure. There are two classes of "dual variables". One is referred to as the *adjoint vectors* (also known as the *co-state vectors*), which are the dual variables for the fluid dynamics, i.e. $\dot{q}(t) = f(q(t), z(t))$. The other is called the *Lagrangian multipliers*, which are the dual variables for the state constraints, i.e. $g(q(t)) \leq 0$, and the pure-control constraints i.e. $h(z(t)) \leq 0$. More precisely, let $p \in \mathbb{R}^2$ denote the adjoint vector, and $\eta \in \mathbb{R}^2$ and $\xi \in \mathbb{R}^3$ denote the Lagrangian multipliers for the state and control constraints, respectively. The *Hamiltonian* $H : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ of the system is defined as:

$$\begin{aligned}
H(q(t),z(t),p(t)) &:= p(t)^T f(q(t),z(t)) + F(q(t)) \\
&= p_1(t)\dot{q}_1(t) + p_2(t)\dot{q}_2(t) + c_1 q_1(t) + c_2 q_2(t) \\
&= p_1(t)\left(\lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)\right) \\
&\quad + p_2(t)\left(\lambda_2 - \mu_2 z_2(t) - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)\right) + c_1 q_1(t) + c_2 q_2(t).
\end{aligned}$$

The *augmented Hamiltonian* $L : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}$ is defined as

$$\begin{aligned}
&L(q(t),z(t),p(t),\eta(t),\xi(t)) \\
&:= H(q(t),z(t),p(t)) + \eta(t)^T g(q(t)) + \xi(t)^T h(z(t)) \\
&= p_1(t)\left(\lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)\right) \\
&\quad + p_2(t)\left(\lambda_2 - \mu_2 z_2(t) - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)\right) + c_1 q_1(t) + c_2 q_2(t) \\
&\quad - \eta_1(t) q_1(t) - \eta_2(t) q_2(t) + \xi_1(t)(z_1(t) + z_2(t) - s) - \xi_2(t) z_1(t) - \xi_3(t) z_2(t).
\end{aligned}$$

Pontryagin's Minimum Principle states a number of necessary conditions which the optimal solution to the optimal control problem (F2′) satisfies. The actual theorem can be found in Appendix A.2.3. Here we provide a brief overview of the conditions.

1) *Ordinary Differential Equation condition* (ODE) specifies the dynamics of the "optimal primal trajectory" $q^*(t)$:

$$q^*(0) = q_0, \quad \dot{q}^*(t) = f(q^*(t), z^*(t)). \tag{ODE}$$

2) *Adjoint Vector condition* (ADJ) specifies the dynamics of the "optimal dual trajectory" $p^*(t)$:

$$\dot{p}_1^*(t) = (\theta_1+\gamma_1)p_1^*(t) - \gamma_1 p_2^*(t) - c_1 + \eta_1^*(t), \;\; \dot{p}_2^*(t) = (\theta_2+\gamma_2)p_2^*(t) - \gamma_2 p_1^*(t) - c_2 + \eta_2^*(t). \tag{ADJ}$$

In general, we cannot fully characterize $p^*(t)$ due to the fact that $p_i^*(0)$ and $\eta_i^*(t)$ are "unspecified", i.e., we cannot fully specify their values or dynamics based on the necessary conditions.

3) *Minimization condition* (M) characterizes the optimal control $z^*(t)$ as a minimizer of the Hamiltonian:

$$H(q^*(t), z^*(t), p^*(t)) = \min_z \{H(q^*(t), z(t), p^*(t))\}. \tag{M}$$

As $H(q^*(t), z(t), p^*(t))$ is linear in $z(t)$, it is easy to see from (M) that the optimal control strictly prioritizes one class at any given time. As $z_1^*(t) + z_2^*(t) = s$ for $t \in [0, \tau^*]$, we can write $z_1^*(t) = s - z_2^*(t)$. Then, we define

$$\psi(t) := \frac{\partial H(q^*(t), s - z_2(t), z_2(t), p^*(t))}{\partial z_2} = \mu_1 p_1^*(t) - \mu_2 p_2^*(t).$$

$\psi(t)$ is referred to as the *switching curve*, because the sign of $\psi(t)$ determines which class we should give priority to. In particular, to minimize H, when $\psi(t) > 0$, priority should be given to Class 1 at time t, i.e.,

$$z_1^*(t) = \begin{cases} s & \text{if } q_1^*(t) > 0 \\ \min\left\{s, \frac{\lambda_1 + \gamma_2 q_2^*(t)}{\mu_1}\right\} & \text{if } q_1^*(t) = 0, \end{cases} \quad \text{and} \quad z_2^*(t) = s - z_1^*(t). \tag{1.9}$$

When $\psi(t) < 0$, priority should be given to Class 2, i.e.,

$$z_1^*(t) = s - z_2^*(t), \quad \text{and} \quad z_2^*(t) = \begin{cases} s & \text{if } q_2^*(t) > 0 \\ \min\left\{s, \frac{\lambda_2 + \gamma_1 q_1^*(t)}{\mu_2}\right\} & \text{if } q_2^*(t) = 0. \end{cases} \tag{1.10}$$

However, when $\psi(t) = 0$, the optimal control is undetermined. We also note that $\psi(t)$ can be fully characterized by $p_i^*(t)$'s, $i = 1, 2$. Thus, analyzing the structure of the optimal dual trajectory $p^*(t)$ can reveal important information about the optimal scheduling policy $z^*(t)$.

4) For optimal control problems with state constraints, if F, f, g, h do not depend on t explicitly, *Hamiltonian condition* (H) requires that $H(q^*(t), z^*(t), p^*(t))$ is a constant for all $t \in [0, \tau^*]$. Further, if the problem has a fixed termination state but free termination time, as in our case, then the constant is equal to zero (Cristiani and Martinon, 2010). In particular, we have

$$H(q^*(t), z^*(t), p^*(t)) = 0. \tag{H}$$

5) *Transversality condition* (T) requires that

$$-\mu_1 p_1^*(t)+\xi_1^*(t)-\xi_2^*(t)=0, \quad -\mu_2 p_2^*(t)+\xi_1^*(t)-\xi_3^*(t)=0. \tag{T}$$

6) *Complementarity condition* (C) requires that

C1) $\eta_1^*(t)=0$ if $q_1^*(t)>0$; $\eta_1^*(t)\geq 0$ if $q_1^*(t)=0$.

C2) $\eta_2^*(t)=0$ if $q_2^*(t)>0$; $\eta_2^*(t)\geq 0$ if $q_2^*(t)=0$.

C3) $\xi_1^*(t)=0$ if $z_1^*(t)+z_2^*(t)<s$; $\xi_1^*(t)\geq 0$ if $z_1^*(t)+z_2^*(t)=s$.

C4) $\xi_2^*(t)=0$ if $z_1^*(t)>0$; $\xi_2^*(t)\geq 0$ if $z_1^*(t)=0$.

C5) $\xi_3^*(t)=0$ if $z_2^*(t)>0$; $\xi_3^*(t)\geq 0$ if $z_2^*(t)=0$.

7) *Jump condition* (J) characterizes the potential discontinuity of the adjoint vector $p^*(t)$ and the Hamiltonian $H(q^*(t),z^*(t),p^*(t))$ at junction times or in the boundary arcs. Specifically, For any time β in a boundary arc or a junction time, the adjoint vector $p^*(t)$ and the Hamiltonian $H(q^*(t),z^*(t),p^*(t))$ may have a discontinuity, but they must satisfy the following jump conditions: There exits a vector $\omega^*(\beta)=(\omega_1^*(\beta),\omega_2^*(\beta))\in\mathbb{R}^2$, such that

$$\begin{aligned}(J1): p^*(\beta-) &= p^*(\beta+)+\omega_1^*(\beta)\nabla_q g_1(q^*(\beta))+\omega_2^*(\beta)\nabla_q g_2(q^*(\beta))\\ (J2): H(q^*(\beta-),z(\beta-),p^*(\beta-)) &= H(q^*(\beta+),z(\beta+),p^*(\beta+))-\omega_1^*(\beta)\nabla_t g_1(q^*(\beta))\\ &\quad -\omega_2^*(\beta)\nabla_t g_2(q^*(\beta))\\ (J3): \omega^*(\beta)\geq 0, &\quad \omega^*(\beta)^T g(q^*(\beta))=0,\end{aligned} \tag{J}$$

where $\nabla_x g$ denote the derivative of g with respect to x.

From the discussion of the necessary conditions, we highlight that if we can characterize the switching curve $\psi(t)$, then we will be able to unfold the corresponding optimal policies. However, this is a highly nontrivial task, as we are not able to fully characterize $p^*(t)$.

1.4.3 The Modified $c\mu/\theta$-Rule Far from the Origin

We now derive several key properties of the switching curve $\psi(t)$ from Pontryagin's Minimum Principle. These properties together allow us to establish the optimal scheduling policy when the states are large (far from the origin).

The first property characterizes the switching curve on the boundary arc.

Lemma 2. *Let $[t_1, t_2]$ be a boundary arc along the optimal state trajectory with entry point t_1 and exit point t_2. For any $t \in (t_1, t_2)$, the switching curve $\psi(t) = 0$.*

The second property establishes the continuity of the switching curve.

Lemma 3. *The switching curve $\psi(t)$ is continuous over $[0, \tau^*]$.*

Assume there exists an optimal control to problem (F2′) under which the state trajectory only has a finite number of junction points. Let N denote the total number of entry and contact points in the optimal state trajectory $q_1^*(t)$ and $q_2^*(t)$ before τ^*. These N entry or contact points are ordered and denoted by τ_j, $j = 1, ..., N$. In particular, τ_1 is the first time when one of the queues gets emptied from the initial queue length q_0; τ_N is the last time before τ^* when one of the queues gets emptied. Naturally, the queue that gets emptied at time τ_N is maintained at zero until the other queue reaches zero at time τ^*. From Lemmas 2 and 3, we know that $\psi(\tau_j) = 0$ for entry/exit point τ_j. To this end, we examine the switching curve backward in time from each entry point and derive the following characterization of the switching curve.

Lemma 4. *For any entry and contact point τ_j, $j = 1, ..., N$, there exists an interval $(0, \alpha_j)$, $0 < \alpha_j < \tau_j$, such that for $t \in (0, \alpha_j)$, the backward switching curve $\psi(\tau_j - t)$ takes the form*

$$\psi(\tau_j - t) = r_1 - r_2 + \big(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\big)\, e^{\upsilon_1(\tau_j - t)} - \big(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\big)\, e^{\upsilon_2(\tau_j - t)},$$

where r_1, r_2 are defined in (1.6) *and* (1.7)*, υ_1, υ_2 are positive constants that depend on γ_i's and θ_i's, and $A_1(\tau_j)$,$A_2(\tau_j)$ are constants that depend on the values of τ_j and $p^*(\tau_j)$.*

Following Lemma 4, we define the *pseudo switching curve* backward from τ_j as

$$\begin{aligned} D^{\tau_j}(t) := r_1 - r_2 + \left(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\right) e^{\upsilon_1(\tau_j - t)} \\ - \left(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\right) e^{\upsilon_2(\tau_j - t)}, \quad \text{for } t \geq 0,\ j = 1, \ldots, N. \end{aligned}$$

In particular, the pseudo switching curve removes the constraint that $t \in (0, \alpha_j)$ from Lemma 4 and it agrees with the switching curve $\psi(\tau_j - t)$ as long as the multipliers $\eta_1^*(\tau_j - t)$ and $\eta_2^*(\tau_j - t)$ stay at zero. However, if one of the multipliers becomes strictly positive at some time β, i.e., $\eta_i^*(\tau_j - \beta) > 0$ for some $i = 1, 2$, the switching curve may deviate from the pseudo switching curve for $t \geq \beta$.

The significance of Lemma 4 is that even though the constants $A_1(\tau_j)$ and $A_2(\tau_j)$ are unspecified, there are only a very few possibilities for the shape of $D^{\tau_j}(t)$, and thus for the part of $\psi(\tau_j - t)$ that coincides with $D^{\tau_j}(t)$. Now, consider the first (forward in time) entry point τ_1. By the definition of τ_1, both classes have strictly positive queues before τ_1, so the multipliers $\eta_1^*(\tau_1 - t)$ and $\eta_2^*(\tau_1 - t)$ are zero for all $t \in (0, \tau_1]$. In this case, the backward switching curve $\psi(\tau_1 - t)$ and the pseudo switching curve $D^{\tau_1}(t)$ coincide over the interval $t \in (0, \tau_1]$. Note that for $t > \tau_1$, the queue length trajectory is beyond its initialization, and thus $\psi(\tau_1 - t)$ is not defined. On the other hand, the pseudo switching curve $D^{\tau_1}(t)$, as a function of t, is well-defined for all $t \geq 0$. Sending t to infinity in the pseudo switching curve $D^{\tau_1}(t)$, we get

$$\lim_{t \to \infty} D^{\tau_1}(t) = r_1 - r_2, \quad \text{for } r_1, r_2 \text{ in (1.6) and (1.7).} \tag{1.11}$$

The sign of the right-hand-side of (1.11) is governed by the modified $c\mu/\theta$-index, which is positive if the modified $c\mu/\theta$-index for Class 1 is larger. It is important to correctly interpret the limit in (1.11) for the backward switching curve $\psi(\tau_1 - t)$. Because $\psi(\tau_1 - t)$ only equals to $D^{\tau_1}(t)$ on $(0, \tau_1]$ and is not defined for $t > \tau_1$, one may hypothesize that if the initial queue lengths, q_0, are large enough, then τ_1, the amount of time needed to empty one of the queues, is also large, and we might be able to send t large enough such that the sign of $\psi(\tau_1 - t)$ will be governed by the modified $c\mu/\theta$-index. However, we need to note that the constants $A_1(\tau_1)$ and $A_2(\tau_1)$ also depend on q_0 through τ_1 and $p^*(\tau_1)$. We thus need to rigorously establish that $A_1(\tau_1)$ and $A_2(\tau_1)$ are properly bounded. Putting all these analyses together, we are able to establish the following result.

Proposition 2. *Under Assumptions 1, 2, and 3, for the transient optimal control problem (F2′), there exists a positive real number M such that when $q_1(t)+q_2(t) > M$, the modified $c\mu/\theta$-rule is optimal at time t, $t \geq 0$.*

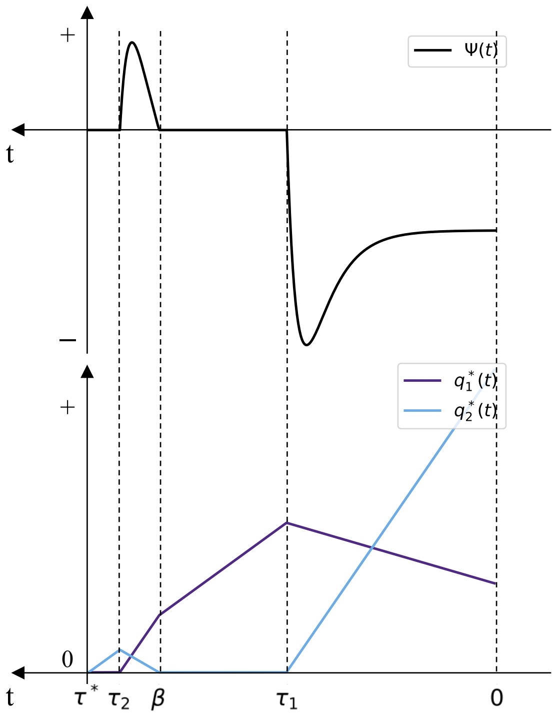
1.4.3 Number of Priority Switches

Propositions 1 and 2 imply that the $c\mu$-rule is optimal when the states are close enough to the origin, and the modified $c\mu/\theta$-rule is optimal when the states are far away from the origin. We now specify what happens in between these two extreme regions. By analyzing possible shapes of the switching curve characterized in Lemma 4, we are able to establish the following proposition.

Proposition 3. *Under Assumptions 1, 2, and 3, for the transient optimal control problem* (F2′)*, if the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, the optimal transient scheduling policy does not switch priority. If the two index rules prioritize different classes, the optimal transient scheduling policy switches priority at most once over the transient time horizon* $[0,\tau^*]$.

Figure 1.8 illustrates the interaction between the switching curve and the optimal transient system dynamics. In particular, we plot the switching curve $\psi(t)$ and the corresponding optimal state trajectory $q^*(t)$ for $t \in [0,\tau^*]$ backward in time. In this example, over the initial time interval $[0,\tau_1)$, $\psi(t)$ is negative, so strict priority is given to Class 2 (following the modified $c\mu/\theta$-rule). Class 2 queue empties at time τ_1 and is given priority to be maintained at zero over the interval $[\tau_1,\beta)$. Immediately after β, the switching curve becomes strictly positive and priority is switched to Class 1 (following the $c\mu$-rule). Note that the Class 1 queue decreases and the Class 2 queue increases over $[\beta,\tau_2)$. Lastly, priority is kept at Class 1 on $[\tau_2,\tau^*]$ to maintain its queue at zero.

Figure 1.8: Example backward switching curve and state trajectory



1.4.3 The Policy Curve

In this section, we first focus on the special case where there is only one-way transition from Class 2 to Class 1, namely, $\gamma_1 = 0$. From Theorem 2, the optimal scheduling rule switches priority at most once. This implies there exists a policy curve $\mathscr{P}$ that divides the state space and governs where the priority switches. Note that this curve is distinct from, but intimately related to, the switching curve $\psi(t)$. Suppose the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1. By utilizing the Hamiltonian condition (H), we are able to characterize (and approximate) the policy curve $\mathscr{P}$ for switching from P_1 to P_2 explicitly. Namely, if the states are initialized "above" $\mathscr{P}$, then the server prioritizes Class 1 until $q(t) \in \mathscr{P}$ at some t. From time t onwards, the server prioritizes Class 2 until the system is emptied at τ^*.

Proposition 4. *Under Assumptions 1, 2, and 3, for the transient optimal control problem* (F2′) *with* $\gamma_1 = 0$, *if* $c_1\mu_1 < c_2\mu_2$ *and* $r_1 > r_2$, *the policy curve* $\mathscr{P}$ *for switching from* P_1 *to*

P_2 is given by

$$\mathscr{P} := \left\{ (\mathbf{a_1}, \mathbf{a_2}) \in \mathbb{R}^2_+ : \frac{1}{\mu_2} \left(\frac{c_1(\lambda_1\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\theta_1} + \frac{B_1(\mathbf{a_2})B_2(\mathbf{a_1},\mathbf{a_2})}{B_3(\mathbf{a_1},\mathbf{a_2})} \right) = 0 \right\},$$

where

$$B_1(\mathbf{a_2}) := (c_1(-\mathbf{a_2}(\theta_2+\gamma_2)+\lambda_2)\mu_1 + c_2\mathbf{a_2}\theta_1\mu_2 + c_1(\mathbf{a_2}\gamma_2+\lambda_1-s\mu_1)\mu_2)$$

$$B_2(\mathbf{a_1},\mathbf{a_2}) := \Bigg(-\mu_2(\mathbf{a_2}\gamma_2\theta_1 + \mathbf{a_1}\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2)$$
$$+(\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2)\left(1+\frac{\mathbf{a_2}(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg)$$

$$B_3(\mathbf{a_1},\mathbf{a_2}) := (\gamma_2-\theta_1+\theta_2)\left(\theta_1(\mathbf{a_2}(\theta_2+\gamma_2)-\lambda_2)\mu_1 + \theta_1(-\mathbf{a_2}\gamma_2 + \mathbf{a_1}\theta_1 - \lambda_1 + s\mu_1)\mu_2\right).$$

If $\gamma_2 = 0$, $c_1\mu_1 > c_2\mu_2$ and $r_1 < r_2$, we can derive the policy curve $\mathscr{P}$ for switching from P_2 to P_1 by symmetry from Proposition 4.

If $c_1\mu_1 > c_2\mu_2$ and $r_1 < r_2$ (still with $\gamma_1 = 0$), the policy curve for switching from P_2 to P_1 cannot be characterized in closed form. This is due to the class-transition dynamics. In particular, we lack information of the Lagrange multiplier $\eta_1^*(t)$ on the boundary arc when $q_1^*(t) = 0$. Due to the deterioration, $\eta_1^*(t)$ not only affects $p_1^*(t)$ but also $p_2^*(t)$ through $p_1^*(t)$, see (ADJ). As such, the condition that $H(q^*(t), z^*(t), p^*(t)) = 0$ is not enough to pin down the value of policy curve. Note that this is not the case in Proposition 4, because on the boundary arc when $q_2^*(t) = 0$, $\eta_2^*(t)$ affects $p_2^*(t)$ only. See Appendix A.2.10 for a more detailed discussion.

We note that the policy curve characterized in Proposition 4 is close to being, but not exactly, linear. More generally, to characterize the policy curve for switching from the modified $c\mu/\theta$-rule to the $c\mu$-rule in the presence of transitions from both class types, i.e., $\gamma_1, \gamma_2 > 0$, we propose the following numerical scheme:

Step 1. Construct n optimal trajectories $q^*(t)$ starting from different initial conditions that are far from the origin. This can be done by solving a discretized version of (F2$'$). Record the n corresponding switching points.

Step 2. Fit the best curve that goes through the n switching points.

We conduct extensive numerical experiments on $\mathscr{P}$ for different system parameters. In all cases, the curve appears to be close to linear. Thus, we suggest setting n to be around 5, setting the discretization step size to be around $0.1\mu_1$, and fitting the best line to the n switching points.

We next provide some sensitivity analysis on the policy curve through numerical experiments. In particular, we use the numerical scheme outlined above to construct the policy curve. For simplicity of illustration, we focus on the case where the $c\mu$-rule prioritizes the urgent class, Class 1, and the modified $c\mu/\theta$-rule prioritizes the moderate class, Class 2, i.e., $c_1\mu_2 > c_2\mu_2$ and $r_1 < r_2$.

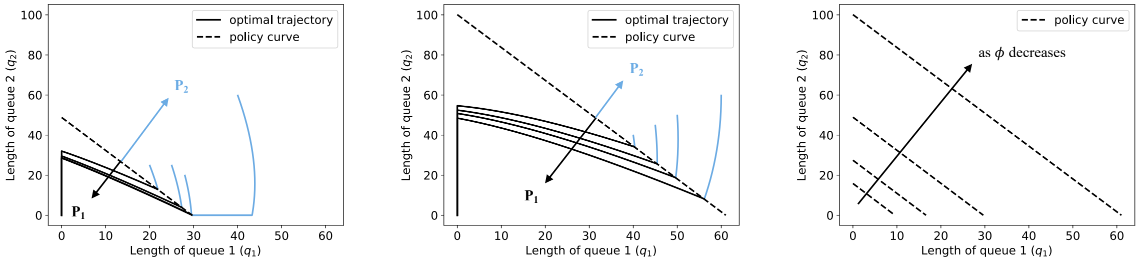
Our first group of numerical experiments is on the value of $\phi := \gamma_2/(\theta_2+\gamma_2)$. As mentioned earlier, if Class 2 patients are identified by a classifier, e.g., an early warning system, ϕ measures the true positive rate of the classifier. Figures 1.9 and 1.10 illustrate how the policy curve changes as ϕ decreases. Since γ_2 and θ_2 both affect the value of ϕ, we first keep θ_2 fixed and vary the value of γ_2 (Figure 1.9). Then, we keep γ_2 fixed and vary the values of θ_2 (Figure 1.10). In both figures, we vary the values of ϕ from 0.7 to 0.4 in increments of size -0.1. We first observe that the policy curve contracts inwards as ϕ increases. In the example of a classifier, this observation suggests that as the quality of the classifier improves, the region in which the optimal scheduling policy prioritizes Class 2 increases. When $\phi = 1$, the size of the region where it is optimal to prioritize Class 1 is minimized, but the region is still non-trivial. On the other hand, as ϕ decreases, a phase transition in the prioritization rule occurs because r_1 will become larger than r_2. In particular, there exists a threshold ϕ_0 such that once $\phi < \phi_0$, $r_1 > r_2$ and the policy curve “vanishes", namely, we should give strict priority to Class 1 for all states. We also note that given the complex nature of system dynamics, the effect of increasing θ_2 and decreasing γ_2 would be different. In particular, when comparing Figure 1.9 to Figure 1.10, we observe that even for the same value of ϕ, the policy curves in the two figures are different. To look further into this, in Figure 1.11, we fix $\phi = 0.6$ and vary θ_2 and γ_2 simultaneously. We observe that as θ_2 and γ_2 decrease, the policy curve contracts outwards. This is because as abandonment and class-transition occur at slower rates, the effect of service completion is more dominant. Thus, the region in which we adopt the $c\mu$-rule increases.

Similar to the above sensitivity analysis on the policy curve with respect to ϕ via θ_2 or γ_2, we also conduct sensitivity analysis for different values of capacity s; see Figure 1.12. We observe that the policy curve contracts inwards as s decreases. Define the nominal traffic intensity as

$$\rho := (\lambda_1/\mu_1 + \lambda_2/\mu_2)/s.$$

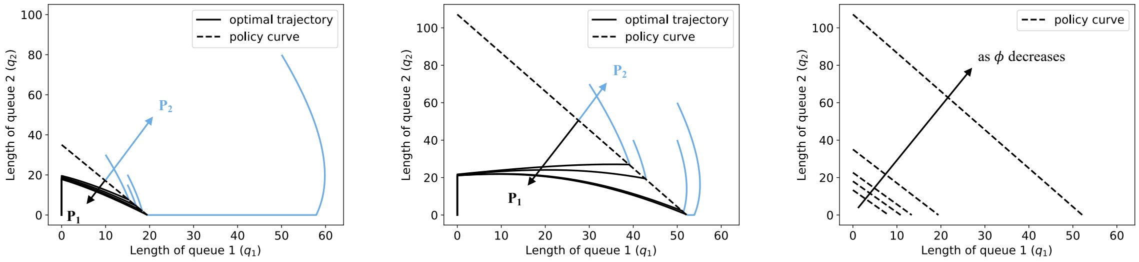
Figure 1.12 indicates that as the system becomes more heavily loaded (i.e., as ρ increases), the region where we prioritize according to the $c\mu$-rule shrinks.

Figure 1.9: Sensitivity analysis of the policy curve with respect to $\gamma_2/(\theta_2+\gamma_2)$ by varying γ_2 $(\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \theta_1 = 0.1, \theta_2 = 0.2, s = 26, c_1 = 5, c_2 = 1)$



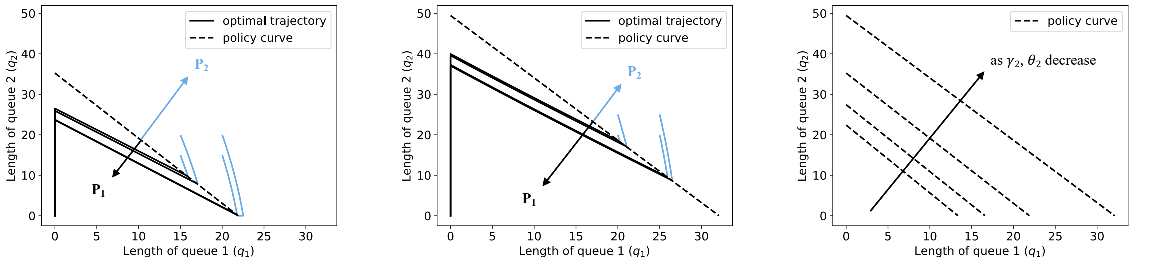
(a) $\phi = 0.5, \gamma_2 = 0.2$ **(b)** $\phi = 0.4, \gamma_2 = 0.13$ **(c)** Decreasing $\phi = 0.7, 0.6, 0.5, 0.4$

Figure 1.10: Sensitivity analysis of the policy curve with respect to $\gamma_2/(\theta_2+\gamma_2)$ by varying θ_2 $(\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, s = 26, c_1 = 5, c_2 = 1)$



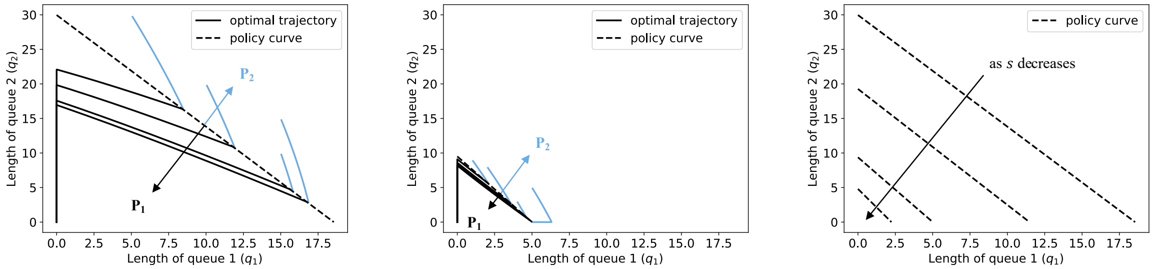
(a) $\phi = 0.5, \theta_2 = 0.4$ **(b)** $\phi = 0.4, \theta_2 = 0.6$ **(c)** Decreasing $\phi = 1, 0.7, 0.6, 0.5, 0.4$

Figure 1.11: Sensitivity analysis of the policy curve with respect to γ_2 and θ_2 for fixed $\gamma_2/(\theta_2+\gamma_2)$ $(\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \theta_1 = 0.1, \phi = 0.6, s = 26, c_1 = 5, c_2 = 1)$



(a) $\gamma_2 = 0.2, \theta_2 = 0.13$ **(b)** $\gamma_2 = 0.1, \theta_2 = 0.07$ **(c)** Decreasing $\gamma_2 = 0.4, 0.3, 0.2, 0.1$

Figure 1.12: Sensitivity analysis of the policy curve with respect to s
$(\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1)$



(a) $s = 30$, $\rho = 0.6$ **(b)** $s = 22$, $\rho = 0.8$ **(c)** Decreasing $s = 30$, 26, 22, 20; increasing $\rho = 0.6$, 0.7, 0.8, 0.9

1.4.4 Numerical Experiments for the Stochastic System

We conclude this section by generalizing the insights from the fluid model analysis to the original stochastic system. The quality of the generalization depends how closely the fluid model approximates the corresponding stochastic system.

As mentioned in Section 1.2.1, the fluid model can arise as a functional law of large numbers limit for a sequence of properly scaled stochastic systems in the conventional heavy traffic regime. In what follows, we first elaborate on the scaling under heavy traffic and then conduct numerical comparisons between the fluid trajectory and scaled stochastic sample paths.

Consider a sequence of stochastic systems indexed by n, $n \in \mathbb{N}$. For Class i in the nth system, the arrival and service rates satisfy $\lambda_i^n := \lambda_i n$, $\mu_i^n := \mu_i n$, $i = 1, 2$. Moreover, we scale down space by considering the fluid-scaled queue length process $\bar{Q}_i^n(\cdot) := Q_i^n(\cdot)/n$ for the nth stochastic system. Given the initial fluid queue length q_0, the nth system has initial queue length $Q^n(0) := \lceil q_0 n \rceil$. Given the fluid policy curve $\mathscr{P}$, for the nth stochastic system, a switch in priority will happen at time t if $\bar{Q}^n(t) \in \mathscr{P}^n$, where

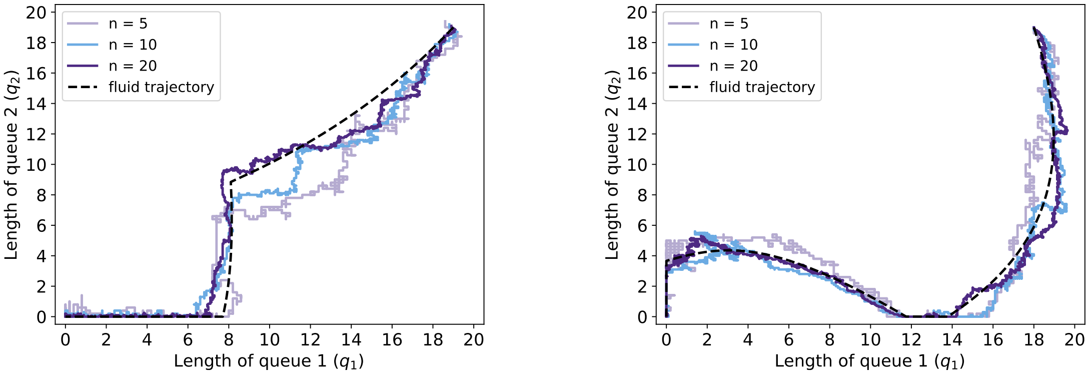
$$\mathscr{P}^n := \left\{ (\bar{Q}_1^n, \bar{Q}_2^n) : \bar{Q}_1^n \in [q_1 - 1/n, q_1 + 1/n], \bar{Q}_2^n \in [q_2 - 1/n, q_2 + 1/n], (q_1, q_2) \in \mathscr{P} \right\}.$$

Figure 1.13 compares the fluid trajectory with a simulated sample path for the corresponding stochastic system for different values of n. We observe from the plots that for a relatively small scaling parameter, e.g., $n = 10$, the stochastic sample path is already well approximated by the fluid model. Furthermore, if we plot the trajectory of the average queue length over multiple sample paths of the stochastic system, the behavior of the "average trajectory" mimics the fluid model even more closely.

For systems with a very small number of servers, we can solve the MDP (S2) numerically; see Appendix A.5 for details about our solution method. In Figure 1.14, we plot the MDP solutions together with the fluid policy curves for four 3-server systems with the nominal traffic intensity ρ varying from 0.6 to 0.9. We observe that for lightly and moderately loaded systems (with $\rho = 0.6, 0.7, 0.8$), the optimal scheduling policy for the stochastic system shares the same structure as the optimal fluid control, i.e., it switches priority once from the modified $c\mu/\theta$-rule to the $c\mu$-rule. We also plot the corresponding fluid policy curve (solid line in Figure 1.14). We observe that when the system is lightly loaded, i.e., $\rho = 0.6, 0.7$, the policy curve of the MDP solution matches the fluid policy curve well. For critically loaded system (with $\rho = 0.9$), the optimal policy follows the modified $c\mu/\theta$-rule throughout; namely, the neighborhood near the origin where the $c\mu$-rule is optimal does not exist. In addition, note that the region where the fluid policy employs the $c\mu$-rule is also very small in this case. Despite some slight deviations between the MDP solution and the fluid-translated policy, in all cases, the optimality gap of the fluid-translated policy is very small, as shall be seen next.

For each of the four stochastic systems in Figure 1.14, we randomly select a set $\mathscr{I}$ of initial points; see Appendix A.5 for details on this initialization. For each initialization $q_0 \in \mathscr{I}$, we compare the average transient cost under (i) the MDP policy (the optimal policy), (ii) the fluid-translated policy with an approximating linear policy curve $\mathscr{P}$, (iii) strict priority to Class 1, P_1, and (iv) strict priority to Class 2, P_2. Each cost is estimated based on 1000 independent sample paths. In Table 1.1, we present the average, minimum and maximum optimality gap for polices (ii), (iii), and (iv). We observe that the fluid-translated policy has a very small optimality gap in all cases. In particular, the maximum optimality gaps are less than 3.8% and the mean optimality gaps are less than 1.6%.

Figure 1.13: Comparison of the transient fluid trajectory and the stochastic sample path
((a) 2 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.5, \mu_2 = 3, \gamma_1 = 0.1, \gamma_2 = 0.1, \theta_1 = 0.1, \theta_2 = 0.4$
(b) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.4, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2$)



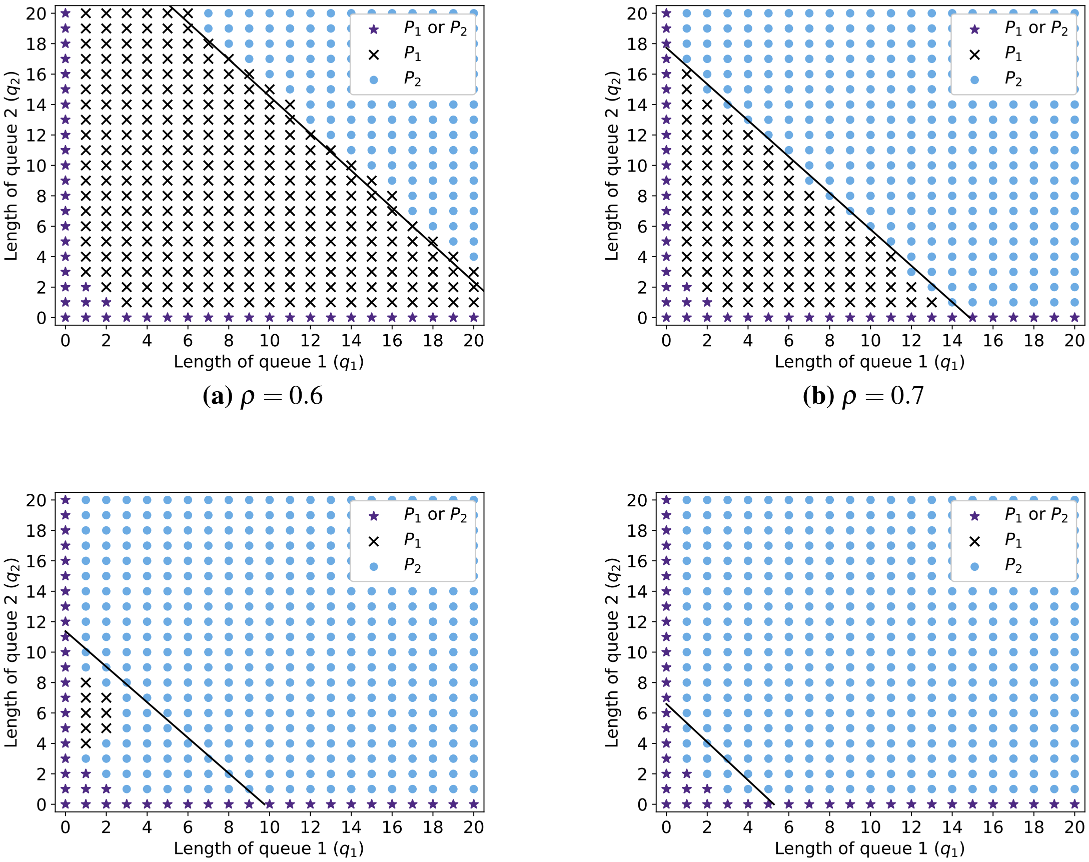
(a) The $c\mu$-rule: P_2, the modified $c\mu/\theta$-rule: P_1

(b) The $c\mu$-rule: P_1, the modified $c\mu/\theta$-rule: P_2

Table 1.1: Stochastic optimality gap of different policies (percentage gap to the MDP)
((a) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.28, \mu_2 = 2.0, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(b) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.09, \mu_2 = 1.7, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(c) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.95, \mu_2 = 1.48, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(d) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.84, \mu_2 = 1.32, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$)

Case (a) $\rho = 0.6$				Case (b) $\rho = 0.7$			
	Fluid policy curve	P$_1$	P$_2$		Fluid policy curve	P$_1$	P$_2$
Mean gap	**0.22%**	0.69%	8.34%	Mean gap	**1.11%**	2.43%	4.27%
Min gap	**0.01%**	0.17%	3.92%	Min gap	**0.01%**	0.88%	0.33%
Max gap	**0.41%**	1.74%	14.74%	Max gap	**2.85%**	4.60%	16.51%
Case (c) $\rho = 0.8$				Case (d) $\rho = 0.9$			
	Fluid policy curve	P$_1$	P$_2$		Fluid policy curve	P$_1$	P$_2$
Mean gap	**1.53%**	6.92%	0.77%	Mean gap	**0.95%**	17.31%	0.00%
Min gap	**0.25%**	4.19%	0.10%	Min gap	**0.09%**	12.41%	0.00%
Max gap	**3.74%**	10.28%	1.70%	Max gap	**1.76%**	24.71%	0.00%

Figure 1.14: Exact MDP solutions (Solid line is the corresponding fluid policy curve)
((a) 3 servers:
$\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.28, \mu_2 = 2.0, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(b) 3 servers:
$\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.09, \mu_2 = 1.7, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(c) 3 servers:
$\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.95, \mu_2 = 1.48, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(d) 3 servers:
$\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.84, \mu_2 = 1.32, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1)$



(a) $\rho = 0.6$

(b) $\rho = 0.7$

(c) $\rho = 0.8$

(d) $\rho = 0.9$

1.5 Model Generalizations

In this section, we consider two generalizations of the model. In the first one, we consider the case with time-varying arrival rates to capture the full demand shock period. In particular, we assume the arrival rates are high for a certain period of time before returning to the “normal" level, and study the transient optimal control problem in this setting. In the second one, we study a system with more than two classes where transition can happen between adjacent classes. We generalize the modified $c\mu/\theta$-index to this setting. As we will demonstrate subsequently, many of the insights we derived in the previous sections still hold in these generalizations.

1.5.1 Time-Varying Transient Arrival Rate

The transient optimal control problem is motivated by demand shock scenarios. The analysis in Section 1.4 focuses on an after-shock optimal control problem where the arrival rates satisfy $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, but the system can have an arbitrarily large initial backlog. In this section, we consider a generalization where we include in our analysis the period of time during the shock. The shock raises the arrival rates for a fixed amount of time, during which the service capacity falls short and the queue increases. After the initial shock, the arrival rates restore to normal, i.e., the service capacity is able to meet demand and eventually empty the system. Formally, we impose the following assumption on the arrival rates.

Assumption 4. *The arrival rates to the system, $\{\lambda_1(t) : t \geq 0\}$ and $\{\lambda_2(t) : t \geq 0\}$, are non-negative, and there exists some $T \geq 0$ such that*

(1) $\lambda_1(t)$ and $\lambda_2(t)$ are continuously differentiable with respect to t over the time interval $[0,T)$;

(2) $q_1(t) > 0$ and $q_2(t) > 0$ for all $t \in [0,T]$ under any admissible scheduling policy $\pi \in \mathscr{F}$;

(3) $\lambda_1(t) = \lambda_1$ and $\lambda_2(t) = \lambda_2$ for some λ_1, λ_2 that satisfy $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, for all $t \geq T$.

Under Assumption 4, $\lambda_1(t)$ and $\lambda_2(t)$ can be any continuously differential functions with argument t over the initial interval $[0,T)$ (condition (1)). The demand shock needs to be high enough such that neither queues empties during the shock (condition (2)). Lastly, after the shock, we have enough resources to bring the queue all the way back to zero (condition (3)). With a slight abuse of notation, we define

$$\tau := \inf\{t \geq T : q_1(t) + q_2(t) = 0\} - T$$

Under Assumption 4 condition (3), $\tau < \infty$.

Recall that the fluid dynamics are defined via $f(q,z) = (f_1(q,z), f_2(q,z))$, where $f_1(q,z) = \lambda_1(t) - z_1\mu_1 - \theta_1 q_1 - \gamma_1 q_1 + \gamma_2 q_2$ and $f_2(q,z) = \lambda_2(t) - z_2\mu_2 - \theta_2 q_2 - \gamma_2 q_2 + \gamma_1 q_1$. For the initial period $[0,T)$ with potentially time-varying arrival rate, we will add a time component to f, i.e., $f(q,z,t)$, to reflect the time dependence. The transient fluid optimization problem can be formulated as a two-stage optimal control problem. In particular, the first-stage problem (over the initial time period $[0,T)$) is expressed as

$$\begin{aligned} \min_{\{z(t):0\leq t<T\}} \quad & \int_0^T F\left(q(t)\right) dt + \Xi(q(T)) \\ s.t. \quad & \dot{q}(t) = f\left(q(t), z(t), t\right) \\ & h(z(t)) \leq 0, \end{aligned} \tag{1.12}$$

where $\Xi(q(T))$ is the terminal cost and is the optimal objective value for the second-stage problem

$$\begin{aligned} \min_{\{z(t):T\leq t\leq T+\tau\}} \quad & \int_T^{T+\tau} F\left(q(t)\right) dt \\ s.t. \quad & \dot{q}(t) = f\left(q(t), z(t)\right) \\ & g(q(t)) \leq 0 \\ & h(z(t)) \leq 0. \end{aligned} \tag{1.13}$$

Note that the first-stage problem (1.12) is “explicit" without the state constraint $g(q(t)) \leq 0$, because under Assumption 4, there does not exist an admissible control under which either of the queues gets emptied during $[0,T]$. Let $q^*(T)$ denote the optimal queue length at the end of the initial time horizon in problem (1.12). The second-stage problem (1.13) is the same as (F2$'$) if we shift the time from $[T, T+\tau]$ to $[0,\tau]$ and set the initial condition $q(0) := q(T)$. Due to this connection, the structural insights from the case of constant arrival rates in Section 1.4 is maintained in this time-varying case.

Theorem 3. *Under Assumptions 1, 3, and 4, for the transient optimal control problem* (1.12)–(1.13)*:*

I. If the $c\mu$-rule and the modified $c\mu/\theta$-rule both prioritize Class i, $i = 1,2$, the strict priority rule to Class i is optimal for any $t \in [0, T+\tau^*]$.

II. If the $c\mu$-rule prioritizes Class i but the modified $c\mu/\theta$-rule prioritizes Class j, for $i \neq j$, $i, j = 1,2$, there exist positive real numbers ε and M with $0 < \varepsilon < M$, such that for $t \in [T, T+\tau^*]$, it is optimal to prioritize Class i when $q_1(t) + q_2(t) < \varepsilon$ and prioritize Class j when $q_1(t)+q_2(t) > M$. Furthermore, the optimal scheduling policy switches priority at most once over the entire transient time horizon $[0, T+\tau^*]$.

Theorem 3 indicates that for time-varying arrival rates satisfying Assumption 4, the optimal control switches priority at most once from the modified $c\mu/\theta$-rule to the $c\mu$-rule. However, different from the case of fixed arrival rates, the modified $c\mu/\theta$-rule can be optimal during the demand shock (i.e. $[0,T)$), even for very small queues if the demand rate and/or the duration of the shock are sufficiently large. This indicates that when the priority switches is not only state-dependent but also time-dependent. As a simple consequence of Theorem 3, the following corollary characterizes the optimal transient control for sufficiently large demand shocks.

Corollary 1. *For the two-stage transient control problem, let $\mathscr{P}$ be the policy curve from the second-stage problem, and $M \in \mathbb{R}_+$ be defined in Theorem 3. If the arrival rates $\{\lambda_1(t) : t \geq 0\}$ and $\{\lambda_2(t) : t \geq 0\}$ are such that $q_1(T) + q_2(T) > M$ under any admissible control, the optimal control employs the modified $c\mu/\theta$-rule over the interval $[0,T]$, and switches to the $c\mu$-rule when $t > T$ and the state crosses the policy curve $\mathscr{P}$, namely, when $q_1(t) + q_2(t) \in \mathscr{P}$.*

With general time-varying arrival rates, characterizing when and where the priority switches can be very complicated and highly case-dependent. For example, the switching point depends on where the system is initialized, how long the demand shock lasts, etc. As such, we leverage the insights from Theorem 3 and Corollary 1, and propose two heuristic policies. In both heuristics, we first derive the policy curve based on the optimal control problem (F2$'$) with the normal arrival rates, i.e., λ_1 and λ_2 (ignoring the demand shock). In Heuristic 1, we apply a time-homogeneous policy where we follow the $c\mu$-rule when the queues are “below" the policy curve, and follow the modified $c\mu/\theta$-rule when the queues are “above" the policy curve. In Heuristic 2, we modify the policy to be time-dependent. In particular, we employ the modified $c\mu/\theta$-rule for the initial demand-shock period $[0,T)$. Then, for $t \geq T$, we follow the $c\mu/\theta$-rule when the queue is “above" the policy curve, and follow the $c\mu$-rule when the queues are “below" the policy curve. In Table 1.2, we compare the performance of (i) Heuristic 1, (ii) Heuristic 2, (iii) the $c\mu$-rule, and (iv) the modified $c\mu/\theta$-rule. The problem instances we consider have piecewise constant arrival rates where the arrival rates switch from a fixed high level to a fixed low level, and we vary the duration of the high demand period, T. We observe that when the demand shock lasts for a sufficiently long time, i.e., $T \geq 0.4$, Heuristic 2 performs near optimal. However, when the demand shock lasts for only a short period of time, i.e., $T = 0.1, 0.2$, Heuristic 1 performs very well. This is because, in the later case, the queues barely build up during the demand shock, and it is optimal to apply the $c\mu$-rule throughout.

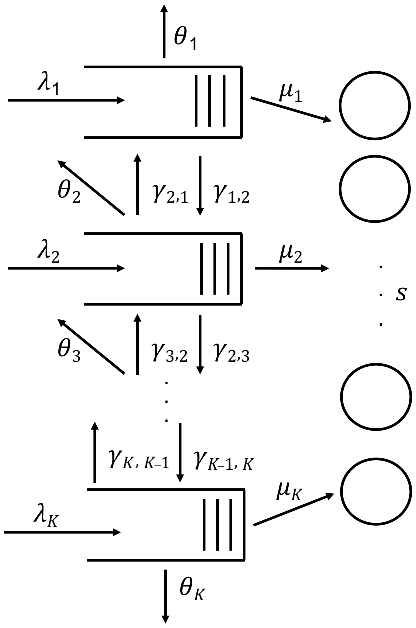
Table 1.2: Fluid optimality gap of different policies (percentage gap to (1.12)) ($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, s = 26, c_1 = 5, c_2 = 1, q(0) = (1,1), \lambda_i(t) = s\mu_i \text{ for } t \in [0,T), i = 1,2$)

Demand shock duration T	Heuristic 1	Heuristic 2	$c\mu$	Modified $c\mu/\theta$
0.1	**0.00%**	18.90%	**0.00%**	46.15%
0.2	**0.00%**	10.15%	**0.00%**	27.41%
0.4	6.14%	**0.00%**	18.48%	6.84%
0.6	9.29%	**0.00%**	72.19%	3.43%
0.8	10.20%	**0.00%**	127.39%	2.05%
1.0	10.33%	**0.00%**	160.89%	1.41%
1.2	10.27%	**0.00%**	179.99%	1.22%
1.4	10.08%	**0.00%**	190.41%	1.03%
1.6	9.54%	**0.00%**	195.29%	0.77%

1.5.2 Multi-Class System

Thus far, our analysis has focused on a two-class system. We next discuss an extension to a multi-class system with K customer classes as depicted in Figure 1.15. The customer classes can be interpreted as having different urgency levels, with Class 1 being the most urgent and Class K being the least. Class i is associated with its arrival rate λ_i, service rate μ_i, abandonment rate θ_i, and cost rate c_i, $i = 1,...,K$. To capture class-transitions, delayed Class i customers degrade into Class $i-1$ at rate $\gamma_{i,i-1}$, and improve to Class $i+1$ at rate $\gamma_{i,i+1}$, where $\gamma_{1,0}, \gamma_{K,K+1} := 0$. Note that this multi-class model can capture the case where some customers never transition type by setting the corresponding transition rates to 0.

Figure 1.15: Multi-class queueing network



1.5.2 Long-Run Average Analysis

Following similar lines of analysis as in Section 1.3, we first optimize over the set of equilibrium points. In particular, we have the following linear program:

$$\begin{aligned}
\min_{\{z_i^e,\, i=1,...,K\}} \quad & \sum_{i=1}^{K} c_i q_i^e \\
s.t. \quad & \lambda_i - \mu_i z_i^e - (\gamma_{i,i-1} + \gamma_{i,i+1} + \theta_i) q_i^e + \gamma_{i+1,i} q_{i+1}^e + \gamma_{i-1,i} q_{i-1}^e = 0, \quad i = 1,...,K \\
& \sum_{i=1}^{K} z_i^e \le s \\
& z_i^e, q_i^e \ge 0, \quad i = 1,...,K.
\end{aligned} \tag{1.14}$$

Set $\tilde{\gamma}_{K,K+1} := 0$, and for i decreasing from $K-1$ to 1, we define the modified class improvement rates sequentially as

$$\tilde{\gamma}_{i,i+1} := \gamma_{i,i+1} \frac{\theta_{i+1} + \tilde{\gamma}_{i+1,i+2}}{\gamma_{i+1,i} + \theta_{i+1} + \tilde{\gamma}_{i+1,i+2}}. \tag{1.15}$$

Similarly, set $\tilde{\gamma}_{1,0} := 0$, and for i increasing from 2 to K, the modified class deterioration rates are sequentially defined by

$$\tilde{\gamma}_{i,i-1} := \gamma_{i,i-1} \frac{\theta_{i-1} + \tilde{\gamma}_{i-1,i-2}}{\gamma_{i-1,i} + \theta_{i-1} + \tilde{\gamma}_{i-1,i-2}}, \tag{1.16}$$

The modified improvement and deterioration rates in (1.15) and (1.16) can be understood as the effective class-transition rates adjusted for potential feedback. For example, the nominal deterioration rate from Class 2 to Class 1 is adjusted from $\gamma_{2,1}$ to $\tilde{\gamma}_{2,1} = \gamma_{2,1}\theta_1/(\theta_1 + \gamma_{1,2})$. Intuitively, if no service is provided, out of the customers that have degraded to Class 1, a proportion $\theta_1/(\theta_1 + \gamma_{1,2})$ will be fed back to Class 2. Thus, the effective degradation rate is $\tilde{\gamma}_{2,1}$.

Rearranging the terms in (1.14), we can derive that the optimal solution to (1.14) assigns the maximum value to the z_i^e with a larger modified $c\mu/\theta$-index, r_i, where

$$\begin{aligned} r_i := \mu_i \Bigg(& \frac{c_i}{\theta_i + \tilde{\gamma}_{i,i-1} + \tilde{\gamma}_{i,i+1}} + \sum_{j=1}^{i-1} \frac{c_j}{\theta_j + \tilde{\gamma}_{j,j-1} + \tilde{\gamma}_{j,j+1}} \prod_{k=j+1}^{i} \frac{\gamma_{k,k-1}}{\gamma_{k,k-1} + \tilde{\gamma}_{k,k+1} + \theta_k} \\ & + \sum_{j=i+1}^{K} \frac{c_j}{\theta_j + \tilde{\gamma}_{j,j-1} + \tilde{\gamma}_{j,j+1}} \prod_{k=i}^{j-1} \frac{\gamma_{k,k+1}}{\gamma_{k,k+1} + \tilde{\gamma}_{k,k-1} + \theta_k} \Bigg), \quad \text{for } i = 1, ..., K. \end{aligned}$$

Note that while the expression is more complex, this index has similar interpretation to that when there are only two classes.

To establish the optimality of the modified $c\mu/\theta$-rule for the long-run average cost, we also need to verify that the optimal equilibrium point in (1.14) is an asymptotically stable equilibrium under the modified $c\mu/\theta$-rule. This requires extending the Lyapunov argument in Appendix A.1 to the multi-class setting. We note that this task will become prohibitively tedious, especially for a large number of classes, K.

1.5.2 Transient Analysis

The transient analysis for the two-class system can also be partially generalized to multi-class systems.

First, based on the insights from the two-class case, when the states are arbitrarily close to the origin, the effect of class-transition and abandonment on the system dynamics is only second-order. Focusing on the service completions, we can show that such that the $c\mu$-rule is optimal in the ε-neighborhood around the origin, i.e., when $q_i(t) \in [0, \varepsilon)$ for $i = 1, ..., K$, for ε sufficiently small.

Second, applying Pontryagin's Minimum Principle for the multi-class case, we see that at any time t, the optimal policy prioritizes the class with a larger $p_i^*(t)\mu_i$ value, where p_i^* is the optimal adjoint vector associated with Class i. Let τ_1 be the first time after initialization when one of the queues gets emptied. Using a similar backward construction as in Lemma 4, we can characterize $p_i^*(\tau_1 - t)$ and show that $\lim_{t\to\infty} p_i^*(\tau_1 - t)\mu_i = r_i$ (assuming we can extend the function to $t > \tau_1$), where r_i is the modified $c\mu/\theta$-index for Class i. This suggests that the modified $c\mu/\theta$-rule is likely to be optimal when the queues are far enough from the origin. However, we emphasize that this is only a heuristic argument. Rigorously establishing such a result requires highly non-trivial derivations.

Lastly, the optimal scheduling policy for areas between the ε-neighborhood of the origin and the far from the origin region remains unclear. Noticably, it is not necessarily true that the optimal policy switches priority at most once along the trajectory, as in the case of a two-class system. We perform extensive numerical experiments for a three-class model. The solutions to (F2$'$) confirm that the optimal solution follows the modified $c\mu/\theta$-rule when the state q is sufficiently far from the origin, and the $c\mu$-rule near the origin. In many problem instances, the optimal scheduling policy switches priority rule at most once. However, there are also instances where the optimal scheduling policy switches priority more than once, and it follows neither the modified $c\mu/\theta$-rule nor the $c\mu$-rule during part of the transient horizon.

To facilitate implementations, we propose a *one-switch policy*, where we switch priority at most once, and follow the the modified $c\mu/\theta$-rule when the system state is far from the equilibrium and the $c\mu$-rule when the state is close to the equilibrium. Table 1.3 compares the performance of the *one-switch policy*, the modified $c\mu/\theta$-rule, and the $c\mu$-rule. For the *one-switch policy*, we find the optimal policy curve when imposing that at most one switch is allowed. According to the system parameters, the modified $c\mu/\theta$-rule and the $c\mu$-rule prioritize in the order of Classes $3,2,1$ and Classes $1,2,3$ respectively. In these systems, the optimal LP solution may, under certain initial conditions, prioritize Class 2 over part of the transient horizon. Nevertheless, the sub-optimality gap of the one-switch policy is fairly small, i.e., less than 2.6%, while applying the modified $c\mu/\theta$-rule or the $c\mu$-rule throughout can sometimes lead to very large sub-optimality gaps. In general, we expect the one-switch policy to be a reasonable heuristic policy when the modified $c\mu/\theta$-index and the $c\mu$-index are relatively aligned.

Table 1.3: Fluid optimality gap of different policies (percentage gap to (F2′))
($\lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 30, \mu_1 = 4, \mu_2 = 5, \mu_3 = 6, \theta_1 = 0.2, \theta_2 = 0.1, \theta_3 = 0.1, \gamma_{2,1} = 0.1, \gamma_{1,2} = 0.2, \gamma_{3,2} = 0.1, \gamma_{2,3} = 0.3, s = 30, c_1 = 20, c_2 = 15, c_3 = 10, c\mu$-index $= \{80, 75, 60\}$, modified $c\mu/\theta$-index $= \{433, 583, 650\}$)

Initialization	One-switch	Modified $c\mu/\theta$	$c\mu$
(5, 5, 5)	**0.00%**	15.83%	**0.00%**
(10, 10, 10)	**0.00%**	14.80%	**0.00%**
(50, 50, 50)	**0.60%**	9.39%	**0.60%**
(100, 100, 100)	**2.21%**	6.56%	**2.21%**
(250, 250, 250)	**2.57%**	2.84%	5.09%
(500, 500, 500)	**0.95%**	1.04%	7.13%
(1000, 1000, 1000)	**0.36%**	0.42%	8.11%

1.6 Conclusion

In this work, we propose a novel multi-class queueing model to capture the class-transition behavior (e.g., degradation or improvement) in service systems. Our analysis provides insights into how proactive service should be utilized. We identify an important metric, the modified $c\mu/\theta$-index, which plays a critical role in specifying the optimal scheduling policy and lends itself to a very intuitive interpretation. In particular, as in the case of the conventional $c\mu/\theta$-index, the modified $c\mu/\theta$-index balances the relative importance of holding costs, service times, and abandonment rates. Moreover, it augments the standard $c\mu/\theta$-index by several important additional terms that account for class-transitions.

We study both the long-run average cost and the transient cost minimization problem. When planning the system in the long run, we show that following the modified $c\mu/\theta$-rule is optimal. When considering the most cost-effective way to clear backlogs created by demand shocks, one should employ the modified $c\mu/\theta$-rule when the system has a very large backlog (i.e., when it is essential to account for the abandonment and class-transition dynamics), and follow the $c\mu$-rule when the system has a sufficiently small backlog (i.e., when cost minimization is driven by service completions).

We assume, throughout the chapter, that class-transitions and abandonment happen according to independent exponential clocks, i.e., class-transition and abandonment have constant failure rates. It has been shown in a number of service settings, the patience times may have increasing or decreasing failure rates (Puha and Ward, 2019). One can potentially extend the long-run average cost minimization problem to incorporate non-exponential class-transitions and abandonments. In particular, characterizing the fluid equilibrium points in these cases follows similar lines of analysis as in Whitt (2006a). However, since the optimality of the conventional $c\mu/\theta$-rule may no longer hold for non-exponential patience-time distributions (Puha and Ward, 2019), we expect the optimality of the modified $c\mu/\theta$-rule may also not hold with general class-transitions. We leave this as an interesting direction for future research.

Chapter 2: Use of Real-Time Information to Predict Future Arrivals in the Emergency Department

2.1 Introduction

2.1.1 Background and Importance

Among the numerous operational and logistical challenges facing Emergency Departments (ED), ED overcrowding has been an endemic and unfortunately growing challenge across many acute care centers across the United States and globally. A large body of work has established that ED overcrowding is associated with adverse patient outcomes (Johnson and Winkelman, 2011) including reduced quality of care (Ball et al., 2017), reduced hospital revenue (Pines et al., 2011), increased mortality (Jo et al., 2012; McCusker et al., 2014) and even clinician burnout (Lall et al., 2021). As patient volumes continue to increase both in the acute care and inpatient setting (Lin et al., 2018), limited ability to scale or increase inpatient bed capacity dynamically in most hospital settings makes patient utilization forecasting critical. Past research indicates that ED crowding can be reduced by appropriate re-allocation of physician and nursing resources Joseph and White (2020). However, this approach relies on adequate, short-term, patient demand forecasting.

Forecasting Emergency Department (ED) arrivals and volumes offers the opportunity to improve efficient matching of clinical and operational resources with actual patient volume. Past work has used a variety of quantitative and statistical modeling to forecast ED arrivals. Several studies have utilized time-series models to forecast future arrivals based on recent arrival count information (Tandberg and Qualls, 1994; Morzuch and Allen, 2006; Schweigler et al., 2009; Boyle et al., 2012; Asheim et al., 2019; Choudhury, 2019). Additional work has tried other prediction models and found that other features such as day of the week, time of the year, holidays, and weather are effective in predicting ED demand Holleman et al. (1996); Batal et al. (2001); Zibners et al. (2006); Jones et al. (2008); Marcilio et al. (2013). There are also recent efforts that explore techniques to combine these aforementioned features with time-series models (Calegari et al., 2016; Whitt and Zhang, 2019).

Most of the existing literature utilizes classic predictors such as lagged arrival counts, temporal and seasonal variations, holidays, and weather. A few other studies have examined limited real-time information beyond weather and lagged arrival counts such as ambulance diversion status, chief complaints, and physician capacity (Brillman et al., 2005; McCarthy et al., 2008; Chase et al., 2012). However, to the best of our knowledge, little research has explored the vast amount of patient-level and regional data that are made available in near-real time by sources such as electronic health records of recent ED arrivals and Google trends (i.e., relative search volumes on Google for certain keywords).

2.1.2 Goals of This Investigation

The goal of this study was to explore and evaluate rich real-time information (including lagged arrival counts, temporal and seasonal variations, holidays, weather, electronic health records, and Google trends) to predict shift-level ED patient volumes. We sought to explore whether real-time information had predictive power, and what forecasting methods would be most appropriate for predicting ED demand. We also aimed to compare real-time information models to models that did not utilize real-time information, allowing an assessment of any potential gain in prediction accuracy from real-time information.

2.2 Methods

2.2.1 Study Setting and Objective

We conducted a retrospective study using data obtained from the electronic health records for an adult ED in a large academic hospital in New York City. A total of 164,858 adult patients who arrived at the ED from 12:00 AM January 1, 2018, through 11:59 PM August

26, 2019, are included in the analysis.

At the hospital, each day was divided into two main 12-hour nursing shifts that start at 7:00 AM and 7:00 PM, respectively. To facilitate relevant operational decision making (e.g., nurse staffing decisions), the subject of prediction was the shift-level arrival count defined as the total number of patients who arrived at the ED during each shift.

Model fitting and selection was performed using one year of data from January 1, 2018 to January 31, 2019, which we hereafter refer to as the training set. The performance of the prediction model was tested using the remaining data from February 1, 2019 to August 26, 2019, which we hereafter refer to as the test set. This study was approved by Columbia University Institutional Review Board: Protocol IRB-AAAT6452.

2.2.2 Data Source

We utilized three sources of data: patient electronic health records, weather data published by National Centers for Environmental Information (2020), and Google Trends (2020). These data sources were selected based on past work, extant models, and our own novel hypotheses. While the importance of weather information is well established in the literature as discussed in the Introduction section, the prediction power of real-time patient electronic health records and Google trends is relatively underexplored. The data extracted from the electronic patient tracking system specify for each patient: (i) the patientâĂŹs clinical time stamps in the ED, including arrival time, first evaluation time, admission decision time, and departure time; (ii) the arrival source of the patient, e.g., walking in or by ambulance; (iii) the patientâĂŹs chief complaint(s) (i.e., reason of visit); (iv) the patientâĂŹs Emergency Severity Index (ESI); (v) lab and imaging ordered: indicators for whether or not lab, CT, MRI, US, and XR were ordered; (vi) indicator for whether the patient was admitted into the hospital; (vii) the Charlson comorbidity index (CCI) of the patient based on a list of 17 comorbidities; (viii) age; and (ix) indicator for whether the patient left without being seen. In addition to the patient-level data, we obtain retrospective weather information for each day between January 1, 2018 and August 26, 2019, which includes the minimum temperature, precipitation, snow, wind, and a hot-weather indicator for whether the maximum temperature exceeds 86ÂřF (30ÂřC). The last source of data comes from the Google trends, which specifies, for each week between January 1, 2018 and August 26, 2019, the relative Google search volume for the words “flu", “weather", “depression", “heart attack", “hospital", “emergency room", “abuse", and “disorder" in New York State. We hypothesized that (i) the search volume for “flu", “weather", “depression", and “heart attack" signaled certain weather-triggered illnesses in the neighborhood; (ii) the search record for “hospital" and “emergency room" was directly correlated with ED demand; and (iii) the more search volume was for “abuse" and “disorder", the more demand the ED would see in the next few days, because studies had demonstrated that patients with a history of alcohol and substance abuse were more likely to return to the ED within 72 hours.19-20 When selecting the data sources and compiling them into shift-level predictors (see the Data Processing section below), we tried to be comprehensive by including as much potentially relevant information as possible. Later in the Model Training and Feature Selection section, we discuss procedures to train different prediction models and identify relevant predictors.

2.2.3 Data Processing

We processed the data into shift-level features to be used for prediction. The basic (not real-time) predictors included day vs. night, day of week, month, season, and near-holiday (within 3 days before and after a national holiday) indicators. With regards to real-time information, weather, Google trends data, and patient-level data were collected. For analytic purposes, we processed and classified the patient records into three categories to be used for prediction.

The first category was the previous-shift counts, which specified for each shift, the arrival count 1 day ago and 7 days ago, as well as the moving average of the shift-level arrival count over the last 30 days. More precisely, the arrival count on the previous day was the total number of patients who arrived during the previous 24 hours. The arrival count on the previous nth day was the two shifts between the previous $24 \times (n-1)$th and $24 \times n$th hour. For example, if the goal was to predict the arrival count for a Tuesday night shift, then the arrival count on the previous day was the sum of the arrival counts during the Tuesday day shift and Monday night shift.

The second category of predictors was the patient comorbidity information, which we processed into the following three sets. The first set specified for each comorbidity, the total number of patients with that comorbidity on the previous day, i.e., during the previous two shifts. The first set also included the sum and weighted sum of CCIs for all patients on the previous day. The second set contained similar information as the first set, but instead of considering the previous-day count, calculated the average daily number of patients with each comorbidity over the last 3 days, as well as the average daily sum and weighted sum of CCIs for all patients over the last 3 days. The third set calculated for each comorbidity, the percentage of patients with that comorbidity over the last 3 days, as well as the average sum and weighted sum of CCIs per patient over the last 3 days. Note that the difference between the second and third sets was that the third set considered average comorbidity measures on the individual level, and was not influenced by how many patients arrived over the last 3 days. The motivation to consider comorbidity information over the last 3 days was due to the existing findings that patients with certain comorbidities are more likely to be readmitted to the ED within 72 hours (Wang et al., 2007; Hong et al., 2019); see the Discussion section for more details. With the comorbidity information specified in multiple forms above, these three sets of information were likely to be correlated. Since it was a priori unclear which specification had the most predictive power, we left it to the model training and feature selection procedures to sift out redundant information and identify important features.

Lastly, the third category of predictors was the recent ED volume and patient severity information. This included the total number of patients who arrived by ambulance on the previous day, the total number of patients with ESI from 1 to 5 on the previous day, the total number of labs, CT, MRI, US, and XR ordered on the previous day, the total number of patients admitted to the hospital on the previous day, the total number of patients whose age exceeds 65 years old on the previous day, the total number of patients whose age exceeds 80 years old on the previous day, the total number of patients who left without being seen on the previous day, the average waiting time (from arrival time to first evaluation time) on the previous day, the average treatment time (from first evaluation time to discharge decision time) on the previous day, and the average boarding time (from discharge decision time to departure time) on the previous day. Intuitively, the waiting and boarding times captured how busy the ED was on the previous day. Formally, we defined the dataset with features including day vs. night, day of week, month, season, and holidays as the base dataset, i.e., dataset without real-time information. We refer to the dataset with real-time information as the full dataset, which contains all the predictors described above.

2.2.4 Model Evaluation

We focused on two measures of forecast accuracy for shift-level arrival counts—the root mean square error (RMSE) and the mean absolute prediction error (MAPE). Let $(y_1, y_2, , y_n)$ be the vector of observed arrival counts for a total of n shifts, and let $(\hat{y}_1, \hat{y}_2, , \hat{y}_n)$ be the corresponding vector of predicted arrival counts given by the prediction model. The RMSE was calculated as the normalized distance between the predicted and observed values:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}.$$

The MAPE was defined as the average percentage error of the prediction:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{y_i}.$$

Hereafter, we refer to the RMSE (MAPE) calculated on the training set as the training RMSE (MAPE), and on the testing set as the testing RMSE (MAPE).

2.2.5 Model Training and Feature Selection

Using the predictors developed in the Data Processing section, we employed state-of-the-art prediction models and feature selection tools. For the benchmark model without real-time information, as we had relatively few features in the base dataset, we trained simple linear regression models and regression tree models, only. When we incorporated real-time information in the full dataset, in addition to linear regression and regression tree, we also trained more sophisticated models including extreme gradient boosting (XGBoost), seasonal autoregressive integrated moving average (SARIMA), and SARIMA embedded with linear regression (SARIMAX).

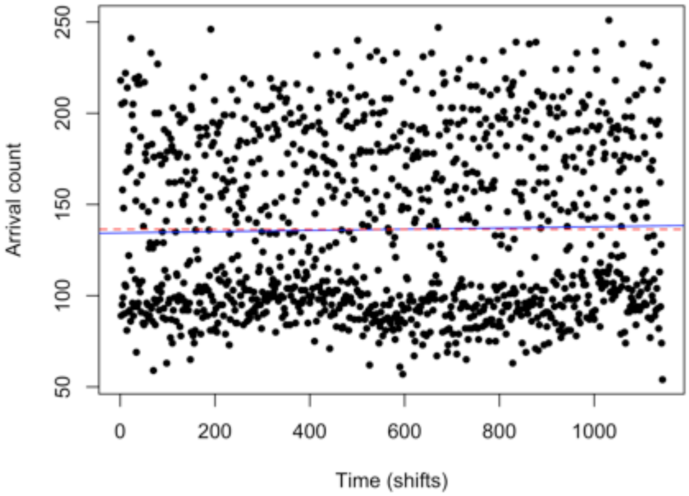
Linear Regression To train the linear regression model, we used a two-way stepwise model selection method based on the Akaike's information criterion (AIC). In particular, we started by including all the predictors in consideration, and in each step, we excluded or included one predictor that gave the largest reduction of the AIC, until the AIC could not be further reduced (Neter et al., 1996). We refer to the model that did not utilize real-time information as LR1, which included general shift-level information such as day vs. night, day of the week, month, and holidays. In comparison, we refer to the model that utilized real-time information as LR2, which could potentially include all the covariates described in the Data Processing section. Because the stepwise heuristic could be impeded by the extremely large number of predictors at initialization, in order to identify the covariates with the highest predictive power, we repeated the above procedure for different categories of covariates. In particular, we used the day vs. night, day of the week, season, and holidays as the base predictors, and respectively added the predictors in category of (i) weather, (ii) google trends, (iii) patient comorbidity information, (iv) previous-shift counts, and (vi) recent ED volume and patient severity information. For each of the five small-scale regression models, we performed the stepwise selection procedure as above and identified the significant covariates in each category. Covariates that were highly correlated were sifted out by comparing the correlation matrix. We refer to the model that combined all the remaining covariates from the small-scale regressions as LR2.

Regression Tree Regression tree model was implemented via the rpart package in R (Therneau et al., 2015). The following hyperparameters were tuned: (i) complexity parameter (cp) ranging from 0 to 0.08 in increment of 0.01, and (ii) maximum depth of any node of the final tree (maxdepth) ranging from 1 to 10 in increment of 1. The other hyperparameters are set to their default values (`https://stat.ethz.ch/R-manual/R-devel/library/rpart/html/rpart.control.html`). For each specification of hyperparameters, we evaluated the modelâĂŹs performance using 10-fold cross-validation on the training set and referred to the resulting average RMSE (MAPE) as the validation RMSE (MAPE). The hyperparameters that gave the smallest validation RMSE was selected. The final model was then trained with these hyperparameters on the training set and evaluated on the test set. We refer to the final model without real-time information as TR1, and to that with real-time information as TR2.

XGBoost XGBoost model was implemented via the xgboost package in python (`https://xgboost.readthedocs.io/en/latest/python/index.html`). The following hyperparameters were tuned: (i) number of boosting rounds (num_round) ranging from 10 to 200 in increment of 10, (ii) maximum tree depth for base learners (max_depth) ranging from 1 to 9 in increment of 1, (iii) boosting learning rate (eta) ranging from 0.1 to 0.5 in increment of 0.1, (iv) L1 regularization term on weights (alpha) ranging from 0.2 to 1 in increment of 0.2, and (v) L2 regularization term on weights (lamba) ranging from 0.2 to 1 in increment of 0.2. The other hyperparameters were set to their default values (`https://xgboost.readthedocs.io/en/latest/parameter.html`). For each specification of hyperparameters, we evaluated the model's performance using 10-fold cross-validation on the training set. The hyperparameters that gave the smallest validation RMSE was selected, and the final model was then trained with these hyperparameters on the training set and evaluated on the test set. Different from the other prediction models considered, XGBoost is a “black-box" model that does not specify explicitly how each covariate drives the prediction. We used relative importance, a measure that quantifies the improvement in prediction accuracy of a tree-based algorithm (including XGBoost) from a split based on a given covariate, to identify relevant predictors (Hastie et al., 2009). Note that relative importance does not specify directionality, but instead only indicates the predictive power of a covariate.

SARIMA and SARIMAX To train the SARIMA model, we set the seasonal term to 14 (i.e., $s = 14$), in order to distinguish the day vs. night and day of the week effects. In addition, since the time series had a stationary increasing trend (Figure 2.1), it was reasonable to conduct a difference for the original series. However, whether to conduct the difference directly (i.e., setting $d = 1, D = 0$) or seasonally (i.e., setting $d = 0, D = 1$) needed to be determined. For each of these two options, we conducted the Dickey-Fuller test to check whether the differenced time series was stationary. The resulting p-values were both 0.01, which suggested at 99% confidence level that the differenced time series under each option did not have a unit root and was therefore stationary. We then used a variation of the Hyndman-Khandakar algorithm (Hyndman and Athanasopoulos, 2018) to determine the hyperparameters. In particular, for each differencing method, we varied the AR term (p), the seasonal AR term (P), the MA term (q), and the seasonal MA term (Q) from 1 to 7 in increment of 1. We considered models where the highest-order AR and MA terms were statistically significant. The final model was then selected based on AIC on the training set and evaluated on the test set. As for the SARIMAX model, we used the same covariates in the selected linear regression model LR2 as external regressors, except that we excluded the previous-day arrival count in the covariates to avoid double counting the lag information. Since the embedded linear regression model already took into account the day vs. night and day of the week variations, as well as the lag information over the last 7 and 30 days, we set the seasonal hyperparameters (P, Q, s) in the SARIMAX model to 0 to avoid overfitting, which leads to an ARIMAX model. The explicit expression of the SARIMA and ARIMAX models is provided in Appendix A, which later facilitates interpretation of the estimated hyperparameters and covariates.

Figure 2.1: Shift-level arrival count from January 1, 2018 to August 26, 2019
The solid blue line is the best regression line where $y_t = 134.4 + 0.00339t$, and the dashed red line is the average arrival count.



2.3 Results

2.3.1 Models without Real-Time Information

After the stepwise selection procedure, the remaining covariates in LR1 were day vs. night, day of the week, month, and holidays. Without real-time information, LR1 achieved an R^2 value of 0.8998 and an adjusted R^2 value of 0.8959. On the test set, LR1 achieved an RMSE of 14.8840 and an MAPE of 9.3226%. Table 2.1 lists the estimated coefficients for the covariates in LR1. The final tree model TR1 had hyperparameters cp = 0.01 and maxdepth = 7. Figure 2.2 illustrates the structure of TR1 estimated on the training set, where the final tree was split by day vs. night indicator, and further by day of the week among all day shifts. Overall, TR1 performed similarly as LR1 on the test set and achieved test RMSE of 14.7560 and test MAPE of 9.5967%.

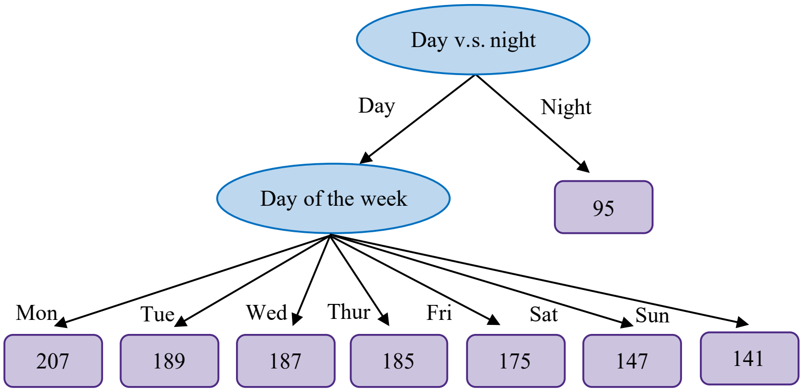
2.3.2 Models with Real-Time Information

Linear Regression The final linear regression model LR2 contained real-time predictors identified by the small-scale regressions, including day vs. night, day of the week, season,

Table 2.1: Estimated coefficients for LR1 without real-time information

	Estimate (standard error)
Monday day	119.674*** (2.970)
Tuesday day	97.095*** (2.981)
Wednesday day	96.109*** (2.955)
Thursday day	94.747*** (2.948)
Friday day	84.401*** (2.961)
Saturday day	56.539*** (2.960)
Sunday day	51.558*** (2.931)
Monday night	9.347*** (2.970)
Tuesday night	6.095** (2.981)
Wednesday night	2.878 (2.955)
Thursday night	4.728 (2.948)
Friday night	5.920** (2.961)
Saturday night	3.597 (2.960)
January	5.853** (2.692)
February	9.174*** (2.757)
March	−2.939 (2.699)
April	−2.449 (2.723)
May	2.457 (2.688)
June	−0.030 (2.721)
July	6.665** (2.687)
August	3.531 (2.701)
September	2.604 (2.709)
October	5.513** (2.688)
November	−3.527 (2.714)
Holiday	−22.565*** (3.530)
Holiday −1 day	−10.360*** (3.527)
Holiday +1 day	15.788*** (3.526)
Intercept	88.460*** (2.793)
Observations	730
R^2	0.900
Adjusted R^2	0.896
Residual Std. Error	14.944 (df = 702)
F Statistic	233.428*** (df = 27; 702)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 2.2: Structure of TR1 and TR2



holidays, weather, the total number of arrivals 1 and 7 days ago, the moving average of daily arrival count over the last 30 days, google trends for the words “depression" and “flu", and the average weighted sum of comorbidity indices per patient over the last 3 days. Overall, LR2 achieved an R^2 value of 0.9084 and an adjusted R^2 value of 0.9045. On the test set, LR2 achieved an RMSE of 14.0893 and an MAPE of 8.6335%. Table 2.2 lists the estimated coefficients for the covariates in LR2.

Regression Tree Among models that utilized real-time information, the hyperparameters cp = 0.01 and maxdepth = 7 again led to the smallest validation RMSE. In this case, TR1 and TR2 are identical, though TR1 was trained on the base dataset without real-time information, and TR2 was trained on the full dataset with real-time predictors. Figure 2.2 illustrates the structure of the final model estimated on the training set. Overall, TR2 achieved test RMSE of 14.7560 and test MAPE of 9.5967%.

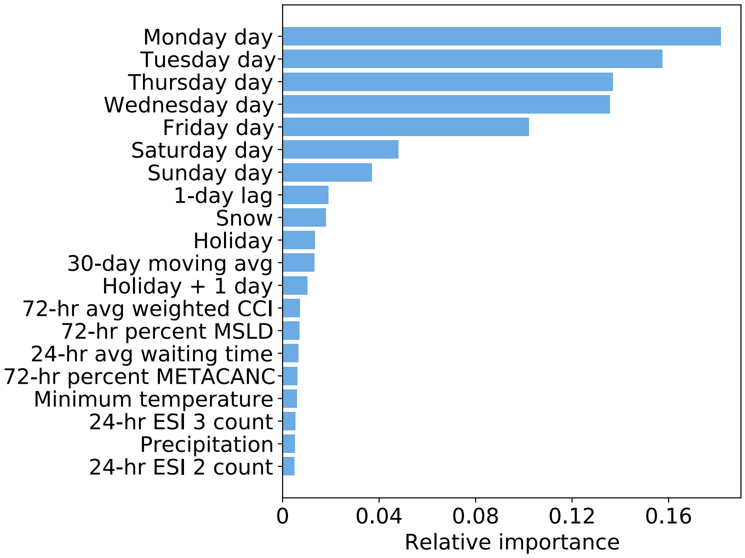
XGBoost The final model with the smallest validation RMSE had the following hyperparameters: number of boosting rounds (num_round) equal to 90, (ii) maximum tree depth for base learners (max_depth) equal to 1, (iii) boosting learning rate (eta) equal to 0.5, (iv) L1 regularization term on weights (alpha) equal to 0.2, and (v) L2 regularization term on weights (lamba) equal to 0.6. Figure 2.3 illustrates the top 20 most informative predictors identified by the selected model (estimated on the training set), including day vs. night, day of the week, the arrival count on the previous day, the moving average of daily arrival

Table 2.2: Estimated coefficients for LR2 with real-time information

	Estimate (standard error)
Monday day	119.972^{***} (2.855)
Tuesday day	97.307^{***} (3.421)
Wednesday day	96.277^{***} (3.174)
Thursday day	93.560^{***} (3.125)
Friday day	83.007^{***} (3.170)
Saturday day	57.421^{***} (3.327)
Sunday day	53.682^{***} (3.242)
Monday night	9.599^{***} (3.333)
Tuesday night	6.170^{*} (3.195)
Wednesday night	2.755 (3.183)
Thursday night	3.963 (3.160)
Friday night	5.650^{*} (3.306)
Saturday night	5.496^{*} (3.244)
Winter	3.021 (2.262)
Summer	-1.574 (2.034)
Fall	-2.355 (1.923)
Holiday	-22.392^{***} (3.512)
Holiday -1 day	-10.137^{***} (3.419)
Holiday $+1$ day	16.840^{***} (3.445)
Min temperature	0.532^{***} (0.114)
Precipitation	-0.160^{***} (0.054)
Snow	-0.169^{***} (0.033)
Wind	0.078^{**} (0.037)
Max temperature $\geq 86°$F	-5.761^{***} (2.146)
1-day lag	0.013 (0.026)
7-day lag	0.038 (0.024)
30-day moving average	0.012 (0.032)
Google trend “depression"	-0.098 (0.081)
Google trend “flu"	0.270^{**} (0.111)
Average weighted comorbidity score per patient over the last 3 days	14.848^{*} (8.817)
Intercept	57.365^{***} (20.894)
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:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

count over the last 30 days, holidays, weather, the average waiting time on the previous day, the average weighted sum of CCIs per patient over the last 3 days, and the percentages of patients with comorbidity “MSLD" (moderate or severe liver disease) and “METACANC" (metastatic solid tumor) over the last 3 days, respectively. The final model achieved test RMSE of 15.9276 and test MAPE of 9.5842%.

Figure 2.3: Top 20 informative predictors in the final XGBoost model



SARIMA and SARIMAX Among all SARIMA models, SARIMA(6,0,5)(7,1,7)14 was selected as the final model, achieving test RMSE of 14.980 and test MAPE of 9.080%. After incorporating the external regressors and setting the seasonal term to 0, the final ARIMAX(1,1,4) model achieved test RMSE of 13.803 and test MAPE of 8.482%. Table 2.3 lists the estimated coefficients for the variables in the ARIMAX(1,1,4) model. It was well expected that the coefficients for the covariates in the embedded linear regression model had the same signs (i.e., directional trends) as those for the final linear regression model (LR2). Moreover, as explicitly derived in Appendix A, the model suggests a positive correlation between the arrival count during the current shift and the arrival counts during the previous two shifts.

Table 2.3: Estimated coefficients for the selected ARIMAX model

	Estimate (standard error)
Monday day	121.172 *** (4.138)
Tuesday day	99.495*** (4.137)
Wednesday day	98.342*** (4.147)
Thursday day	95.473*** (4.152)
Friday day	84.982*** (4.190)
Saturday day	59.392*** (4.352)
Sunday day	55.298*** (4.388)
Monday night	10.417*** (2.606)
Tuesday night	6.889** (2.775)
Wednesday night	3.481 (2.814)
Thursday night	4.751* (2.843)
Friday night	6.331** (3.055)
Saturday night	5.808* (2.976)
Winter	3.064 (2.298)
Summer	-1.635 (2.077)
Fall	-2.152 (1.964)
Holiday	-22.732^{***} (3.280)
Holiday -1 day	-10.604^{***} (3.237)
Holiday $+1$ day	16.338*** (3.287)
Min temperature	0.525*** (0.112)
Precipitation	-0.157^{***} (0.050)
Snow	-0.170^{***} (0.031)
Wind	0.074** (0.035)
Max temperature $\geq 86°$F	-5.482^{***} (2.089)
7-day lag	0.042* (0.023)
30-day moving average	0.009 (0.033)
Google trend “depression"	-0.101 (0.082)
Google trend “flu"	0.261** (0.112)
Average weighted comorbidity score per patient over the last 3 days	13.128 (8.852)
AR1 (ϕ_1)	-0.987^{***} (0.014)
MA1 (θ_1)	-0.054 (0.041)
MA2 (θ_2)	-0.862^{***} (0.041)
MA3 (θ_3)	0.013 (0.039)
MA4 (θ_4)	-0.098^{**} (0.039)
σ^2	198.9
Note:	*p<0.1; **p<0.05; ***p<0.01

2.4 Discussion

Our novel model included a rich collection of real-time and operational-level factors to predict ED arrivals. There are many studies that apply different prediction techniques to forecast ED arrivals (Gul and Celik, 2020), but the majority of these studies only make use of classic predictors such as day vs. night, day of the week, month, holidays, weather, and previous-shift arrival counts (Jing et al., 2020). The predictive power of other rich real-time information has been relatively unexplored. Our work builds on this previous work by exploring a novel large set of real-time predictors from the concurrent patient electronic health records and Google trends. We demonstrated how real-time patient-level information could be processed into shift-level features and identify the relevant ones that were predictive of future patient volume. The estimation results of the selected models also lent themselves to intuitive interpretation about how various trends in the calling population affected future arrivals.

2.4.1 Implications for Emergency Department: Operations Level and Pairing of Resources

Real-time information was effective in improving prediction accuracy of ED arrivals. The linear regression model (LR2) identified previous-shift arrival counts, weather, Google trends, and patient comorbidity as informative covariates. According to the estimated coefficients in Table 2.2, ED arrivals were positively correlated with the patient volume 1 day and 7 days prior, as well as with the moving average of daily arrival count over the last 30 days. Severe weather such as snow, precipitation, and extremely cold or hot temperature could reduce ED arrivals. Nevertheless, the ED tended to see more patients on days with strong wind. In addition, ED arrivals increased during the weeks when there were more Google search records for the word “flu". Intuitively, the search volume for “flu" can be seen as the concurrent flu trend information in New York State. In contrast, ED arrivals lessened during the weeks when the Google trend for the word “depression" was higher. Thus, we infer that fewer patients arrived to the ED during the time periods when mental health issues (e.g., seasonal affective disorder) were salient. Moreover, the average weighted sum of CCIs per patient over the last 72 hours was positively correlated with the incoming patient volume. This trend could be corroborated by the findings that patients with higher weighted sum of CCIs, a history of chronic heart failure (CHF) or chronic obstructive pulmonary disease (COPD), and a history of alcohol and substance abuse were more likely to return to the ED within 72 hours (Wang et al., 2007; Hong et al., 2019). The estimated coefficients for the ARIMAX model in Table 2.3 further confirmed the aforementioned trends. The selected XGBoost model identified similar significant predictors as LR2 (Figure 2.3), with several new features such as the percentages of patients with comorbidities of moderate or severe liver disease (MSLD) and metastatic solid tumor (METACANC) over the last 3 days, as well as the numbers of patients with ESI 3 and 4 on the previous day. We remark that the percentages of patients with each type of comorbidity was correlated with (and thus could be partially captured by) the weighted sum of the CCIs. Similarly, the numbers of patients with ESI 3 and 4 on the previous day were highly correlated with the previous-day arrival count. Hence, the recommendations of informative features by LR2 and the XGBoost model were coherent.

2.4.2 Considerations for Staffing Other Dynamic Resources

Overcrowding due to lack of staffing capacity and other factors has been identified as a significant issue in EDs throughout the United States for over a decade and continues to deteriorate (Institute of Medicine Committee on the Future of Emergency Care in the US Health System, 2006). Appropriate staffing matched to demand, driven by both arrivals and ED census, is associated with improved patient safety and quality of care (Ball et al., 2017). Clinician wellbeing and risk for burnout is directly correlated with levels of overcrowding among other modifiable factors (Curtis and Puntillo, 2007; Anderson et al., 2021; Recio-Saucedo et al., 2015; Chang et al., 2018b). The ability to predict with some degree of certainty variations in ED volume would enable thoughtful and innovative staffing paradigms that would meet anticipated demand when it surges past capacity. As seen in this manuscript, real-time information can certainly be utilized to predict ED volumes, however, this must be matched by real-time resources that are available to be deployed on an ad hoc basis. Similarly, staffing paradigms should account for unanticipated low volume days.

2.4.3 Comparison of Different Prediction Models

Based on the performance measures reported in the Results section, Table 2.4 summarizes the RMSE and MAPE on the training and test sets for all the selected models. Among models that did not utilize real-time information, the final tree model (TR1) performed the best on the test set, achieving test RMSE of 14.756 and test MAPE of 8.917%. After incorporating real-time information and using more sophisticated model, the prediction accuracy on the test set could be further improved. Among models that were trained with real-time information, the linear regression model (LR2) and ARIMAX model performed the best. In particular, the ARIMAX model achieved the smallest test RMSE of 13.803 and the smallest test MAPE of 8.482%. The worse performance of the regression tree and SARIMA models was well expected due to their relatively simple structure, e.g., the SARIMA models only took lags information into account. On the other hand, the more advanced XGBoost model could be impeded by the limited number of observations available for training, e.g., the XGBoost model was trained with 134 features on 730 observations (shifts) only. Hence, when selecting effective forecast methods, we showed that the linear regression and ARIMAX models were not only effective in applying real-time information but also robust to avoid overfitting when the amount of data is limited. In addition, in practice, the linear regression model and ARIMAX model are more interpretable than XGBoost.

Table 2.4: Comparison of the selected models

Model	Utilize real-time information	Training RMSE	Training MAPE	Test RMSE	Test MAPE
LR1	No	14.654	9.275%	14.884	9.322%
TR1/TR2	No	15.967	9.597%	14.756	8.917%
LR2	Yes	14.009	8.960%	14.089	8.633%
XGBoost	Yes	13.280	8.554%	15.928	9.584%
SARIMA	Yes	13.745	7.709%	14.980	9.080%
ARIMAX	Yes	13.761	8.738%	13.803	8.482%

2.4.4 Limitations

Several limitations of the study include the limited amount of data. The training set only contained one year of data that provides 730 observations, which limited the performance of more sophisticated models that requires substantial hyper-parameter tuning such as the XGBoost. In addition, our study was performed for one ED in New York City at a quaternary care facility. An interesting future direction is to apply our analysis to multiple ED sites and compare the prediction accuracy and trends. Lastly, our analyses primarily focused on predicting ED demand in terms of patient arrival counts. An interesting extension is to apply similar approaches to explore the predictive power of real-time information on patient severity and patient length of stay.

2.5 Conclusion

We constructed and evaluated predictions models with rich real-time information to forecast ED patient volume. In alignment with the nursing shift structure in an ED site at a quaternary care facility in New York City, we aimed to predict the shift-level patient arrival count. Various prediction techniques were examined, including linear regression, regression tree, XGBoost, SARIMA, and (S)ARIMAX. Based on the data from our partner ED site, linear regression and ARIMAX when combined with real-time information achieved the highest prediction accuracy measured by RMSE and MAPE. Comparing to prediction models without real-time predictors, we found that contemporary information was able to improve prediction accuracy in near-real time. Among the extensive list of real-time predictors tested, recent patient arrival counts, weather, Google trends, and concurrent patient comorbidity information had the highest predictive power. The effectiveness of real-time information in improving demand forecast has policy implications for staffing. In particular, ED management can utilize real-time demand updates provided by the prediction model to make timely adjustments to staffing levels, which in turn can effectively mitigate ED overcrowding.

Chapter 3: Prediction-Driven Surge Planning with Application in the Emergency Department

3.1 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 being 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. (2018a); 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 chapter, 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

$$\Lambda = \lambda + \lambda^{\alpha} X, \tag{3.1}$$

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 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 Chapter 2, 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 (3.1). 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 savings 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 prediction model developed in Chapter 2, which utilizes concurrent information such as weather, patient comorbidity profile, recent arrival counts, 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.

3.1.1 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 (2006b) 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.

3.1.2 Organization

The rest of the chapter is organized as follows. In Section 3.2 we introduce the model and formulate the two-stage staffing problem. In Section 3.3 we quantify the cost saving from surge staffing. In Section 3.4 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 3.5, where we compare the performance of our proposed staffing rule to several benchmark policies. In Section 3.6, we extend the two-stage staffing rule to accommodate more general prediction errors at the surge stage. Lastly, in Section 3.7, 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 3.8. All the proofs appear in the appendix.

3.1.3 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 $n > 0$.

3.2 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 3.6 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.

$$\min_{\pi\in\Pi} \mathscr{C}_\pi = \min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,\Lambda)} \{ c_2 N_2(N_1,\Lambda) + (h + a\gamma)\,\mathbb{E}\left[Q(N_1 + N_2(N_1,\Lambda),\Lambda)|\Lambda\right]\} \right] \right\}. \tag{3.2}$$

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 (3.2). 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.

3.2.1 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 5.

Proposition 5. *For the optimal solution* $(N_1^*, N_2^*(N_1^*,\Lambda))$ *to problem* (3.2)*:*

(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 5, the cost parameters can be divided into four regimes as summarized in Table 3.1.

Table 3.1: 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 chapter, we are interested in the non-trivial regime that provides motivation to staff both base and surge servers.

Assumption 5. *The cost rates satisfy* $h\mu/\gamma + a\mu > c_2 > c_1$.

3.2.2 Arrival-Rate Uncertainty

Solving (3.2) 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

$$\Lambda = \lambda + X\lambda^{\alpha}\mu^{1-\alpha}, \tag{3.3}$$

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 (3.3) is equivalent to the form of arrival-rate uncertainty introduced in (3.1) 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 (3.3) 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 (3.3) 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.

3.3 When is Surge Staffing Beneficial?

As mentioned in Section 3.1, 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 (3.2), we define the single-stage optimal staffing problem as

$$\min_{\pi \in \Pi} \mathscr{C}_{\pi} = \min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[(h + a\gamma)\, Q(N_1, \Lambda) \right] \right\}. \tag{3.4}$$

Note that the single-stage problem is equivalent to the two-stage staffing problem (3.2) 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 $\mathscr{C}^{\lambda}_{1,*}$ to denote the optimal total cost for the single-stage optimization problem (3.4). Correspondingly, we use $\mathscr{C}^{\lambda}_{2,*}$ to denote the optimal total cost for the two-stage optimization problem (3.2).

Theorem 4 (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 $\mathscr{C}^{\lambda}_{1,*} - \mathscr{C}^{\lambda}_{2,*} = o(\sqrt{\lambda})$.

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

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

We next provide some intuition behind Theorem 4. 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 (3.3) 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 4 suggests the ED manager should only consider surge staffing when the arrival-rate uncertainty is high.

3.4 Near-Optimal Surge Staffing Policy

As derived in Section 3.3, 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 (3.2) 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 (3.2). These approximations can be viewed as asymptotic limits of (3.2) 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 3.4.3.

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 (3.2) 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 (3.2), 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 (3.4) can be approximated by

$$\min_{N_1} \left\{ c_1 N_1 + (h\mu/\gamma + a\mu)\mathbb{E}\left[(\Lambda/\mu - N_1)^+\right] \right\}. \tag{3.5}$$

Note that (3.5) 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

$$N_1 = \frac{\lambda}{\mu} + \bar{F}_X^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\alpha}, \tag{3.6}$$

where $\bar{F}_X$ is the ccdf of X. We remark that for all staffing rules discussed in the chapter, 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 $\mathscr{C}^{\lambda}_{1,NV}$ denote the expected total cost defined in (3.2) under the one-stage newsvendor solution. Recall that $\mathscr{C}^{\lambda}_{1,*}$ is the optimal total cost for the single-stage optimization problem (3.4). Theorem 1 in Bassamboo et al. (2010) establishes that

$$\mathscr{C}^{\lambda}_{1,NV} - \mathscr{C}^{\lambda}_{1,*} = O(\lambda^{1-\alpha}). \tag{3.7}$$

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 (3.2) takes the form

$$\min_{N_1}\left\{c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,\Lambda)}\left\{c_2 N_2(N_1,\Lambda) + (h/\gamma + a)\left(\Lambda - \mu(N_1 + N_2(N_1,\Lambda))\right)^+\right\}\right]\right\}. \tag{3.8}$$

Given N_1, Assumption 5 implies that the optimal surge-stage staffing level in (3.8) is given by

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

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

$$\min_{N_1}\left\{c_1 N_1 + c_2 \mathbb{E}\left[(\Lambda/\mu - N_1)^+\right]\right\}. \tag{3.9}$$

Similar to (3.5), (3.9) 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

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

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

Definition 2 (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 5 that $c_1/(h\mu/\gamma + a\mu) < c_1/c_2$. Thus, in comparison to the single-stage newsvendor solution described in (3.6), 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 $\mathscr{C}_{2,NV}^{\lambda}$ denote the expected total cost defined in (3.2) under the two-stage newsvendor solution. Recall that $\mathscr{C}_{2,*}^{\lambda}$ is the optimal total cost for the two-stage optimization problem (3.2).

Theorem 5 (optimality gap of $u_{2,NV}$)**.** *For $\alpha \in (1/2, 1)$, the two-stage newsvendor solution in Definition 2 has $\mathscr{C}^{\lambda}_{2,NV} - \mathscr{C}^{\lambda}_{2,*} = o(\lambda^{\alpha})$.*

Since $\alpha > 1/2$, Theorem 4 implies that $\mathscr{C}^{\lambda}_{1,NV} - \mathscr{C}^{\lambda}_{2,*} = \Theta(\lambda^{\alpha})$. This, together with Theorem 5 and the gap in (3.7), suggests that $\mathscr{C}^{\lambda}_{1,NV} - \mathscr{C}^{\lambda}_{2,NV} = \Theta(\lambda^{\alpha})$.

3.4.2 Refinement for The Two-Stage Newsvendor Solution

We have established in Theorem 5 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

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

In this case, the stochastic fluctuation of the queue process becomes a secondary effect. More specifically, as we will make clear in Appendix C.3.1 (see (C.11) in the proof of Lemma 8), 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

$$N_1 + N_2(N_1, \ell) = \ell/\mu + \eta\sqrt{\ell/\mu} + o(\sqrt{\ell/\mu}), \quad \text{for some } \eta \in \mathbb{R}. \tag{3.10}$$

Under the capacity prescription in (3.10), 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

6 in Appendix C.5. 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

$$\eta^* := \operatorname*{arg\,min}_{\eta \in \mathbb{R}} \; c_2\eta + \left(\frac{h\mu}{\gamma} + a\mu\right) \underbrace{\frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right) - \eta\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta)}}}_{(a)}. \tag{3.11}$$

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 (3.11) 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 (3.10) (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 3 (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 $\mathscr{C}^{\lambda}_{2,QED}$ denote the expected total cost in (3.2) 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 6 (optimality gap of $u_{2,QED}$)**.** *For $\alpha \in (1/2,1)$, the two-stage QED staffing rule in Definition 3 has $\mathscr{C}^{\lambda}_{2,QED} - \mathscr{C}^{\lambda}_{2,*} = o(\sqrt{\lambda})$.*

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

Theorem 6 establishes that any policy that belongs to the family of the two-stage QED staffing rules in Definition 3 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 6, 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

$$N_1 = \lambda/\mu + \beta^*(\lambda/\mu)^{\alpha} + k\sqrt{\lambda/\mu} \quad \text{and} \quad N_2(N_1,\Lambda) = (\Lambda/\mu + \eta^*\sqrt{\Lambda/\mu} - N_1)^+, \quad \text{for } k \in \mathbb{R}. \tag{3.12}$$

We consider systems with small arrival rates, namely, setting $\lambda = 25, 50, 75, 100$, and vary the value of k in (3.12) 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 3.2 and 3.3. For example, in Table 3.2, 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 3.2: System performance under different specifications of the two-stage QED staffing rule with $\beta^* = 0, \eta^* = 0.610$
$(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 2)$

k λ	-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 3.3: System performance under different specifications of the two-stage QED staffing rule with $\beta^* = 1.282, \eta^* = -0.140$
$(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 10)$

k λ	-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 3.2, $k = 1$ leads to the best performance across all system scales tested. In Table 3.3, $k = 0$ 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 3.3, 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 3.2, 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 3.2 and 3.3, we also summarize a few more sets of simulation results with different surge staffing cost in Appendix C.8.1. 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:

$$N_1 = \lambda/\mu + \beta^*(\lambda/\mu)^\alpha + \eta^*\sqrt{\lambda/\mu}, \quad \text{and} \quad N_2(N_1,\Lambda) = (\Lambda/\mu + \eta^*\sqrt{\Lambda/\mu} - N_1)^+. \tag{3.13}$$

The capacity prescription in (3.13) 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 4. *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 ($N_2 \leq N_2^0$) 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 5 and Theorems 4–6 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.*

3.5 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 (3.11).

(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

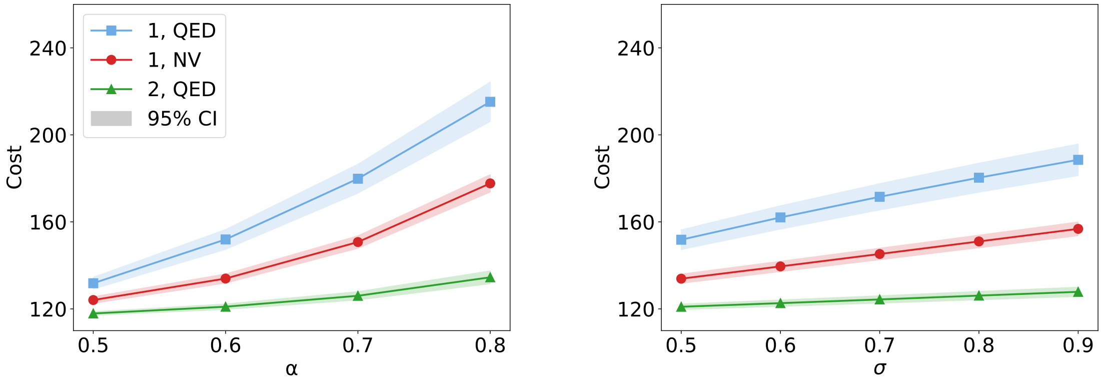
$$\eta^*_{1,QED} := \arg\min_{\eta\in\mathbb{R}} \; c_1\eta + \left(\frac{h\mu}{\gamma} + a\mu\right)\frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right) - \eta\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta)}}. \tag{3.14}$$

This policy ignores arrival-rate uncertainty. It is important to distinguish $\eta^*_{1,QED}$ in (3.14) (used in the single-stage square-root staffing rule) from η^* in (3.11) (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).

3.5.1 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 3.1 illustrates the expected total costs under the three policies, with α increasing from 0.5 to 0.8 in Figure 3.1a and σ increasing from 0.5 to 0.9 in Figure 3.1b. 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 3.1: 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) : $\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \alpha = 0.75$)



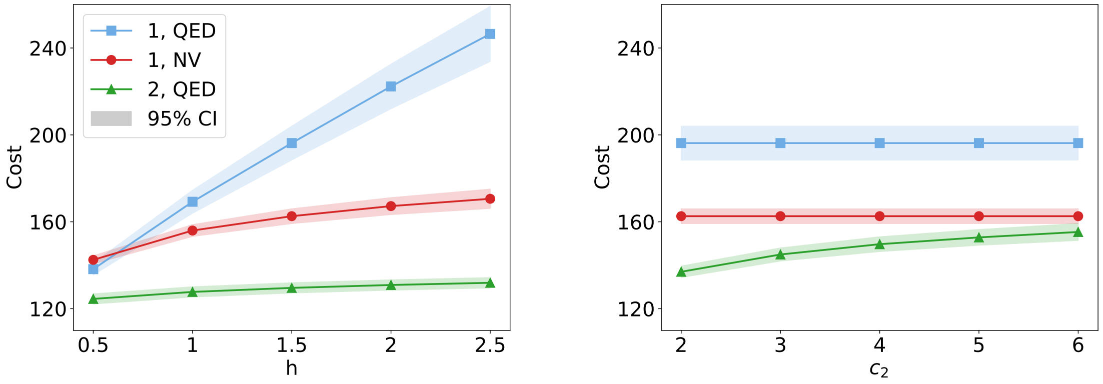
(a) Sensitivity with respect to α

(b) Sensitivity with respect to σ

3.5.2 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 3.2a. 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 larger queue when the realized arrival rate is large. Figure 3.3 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 5 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 3.2b, 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.

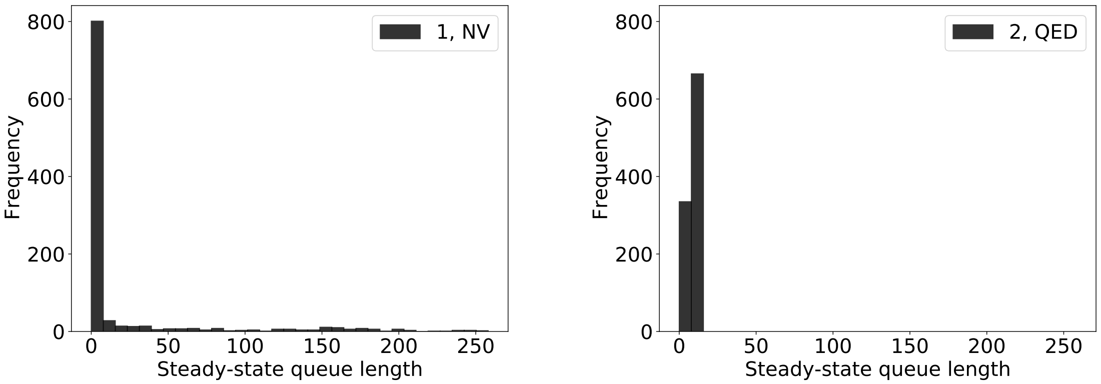
Figure 3.2: Sensitivity analysis with respect to the cost rates
((a): $\lambda = 100, \mu = 1, \gamma = 0.1, a = 2h, c_1 = 1, c_2 = 1.5, \alpha = 0.75, \sigma = 1$
(b): $\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, \alpha = 0.75, \sigma = 1$)



(a) Sensitivity with respect to h

(b) Sensitivity with respect to c_2

Figure 3.3: Distribution of the average steady-state queue length
($\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \alpha = 0.75, \sigma = 1$)



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

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

3.6 Model Extension: Incorporation of Surge-Stage Prediction Error

In the two-stage optimization problem (3.2), 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

$$\Lambda = \lambda + Y\lambda^{\alpha}\mu^{1-\alpha} + Z\lambda^{\nu}\mu^{1-\nu}, \tag{3.15}$$

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 (3.15), 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

$$\min_{N_1}\left\{c_1N_1+\mathbb{E}\left[\min_{N_2(N_1,Y)}\{c_2N_2(N_1,Y)+(h+a\gamma)\mathbb{E}\left[Q(N_1+N_2(N_1,Y),\Lambda)|Y\right]\}\right]\right\}. \quad (3.16)$$

To differentiate notation from that of problem (3.2), we denote the optimal objective value of (3.16) as $\mathscr{C}_{2,*}^{e,\lambda}$ when there is prediction error at the surge stage.

Similar to problem (3.2), we again compare to the single-stage optimization problem (3.4) for Λ in form of (3.15), and use $\mathscr{C}_{1,*}^{e,\lambda}$ to denote its optimal objective value. To draw connection between the arrival rates in (3.3) and (3.15), we can let X be such that

$$X\lambda^{\alpha}\mu^{1-\alpha}=Y\lambda^{\alpha}\mu^{1-\alpha}+Z\lambda^{\nu}\mu^{1-\nu}. \quad (3.17)$$

In this context, problem (3.2) can be seen as an “oracle" problem that knows the exact realized arrival rate at the surge stage. We use $\mathscr{C}_{2,*}^{o,\lambda}$ to denote the optimal objective value of the oracle problem (3.2) for Λ in form of (3.15). 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.

3.6.1 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 (3.16), we propose the *two-stage error policy* and denote it by $u_{2,ERR}$.

Definition 4 (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* (3.11).

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 $\mathscr{C}^{e,\lambda}_{2,ERR}$ denote the expected total cost under $u_{2,ERR}$ when the mean arrival rate is λ. Analogous results to Theorems 4 and 6 hold for $u_{2,ERR}$ when prediction error is small, namely, $\nu \in (0,1/2)$.

Proposition 6. *For* $\alpha \in (1/2,1)$ *and* $\nu \in (0,1/2)$*, we have*

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

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

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

Item (III) in Proposition 6 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.

3.6.2 Moderate to Large Prediction Error: $1/2 \leq \nu \leq \alpha$

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 (3.16), we consider the following stochastic-fluid optimization problem

$$\min_{N_1} \left\{ c_1 N_1 + \mathbb{E}\left[\min_{N_2(N_1,Y)} \left\{ c_2 N_2(N_1,Y) + (h\mu/\gamma + a\mu)\mathbb{E}\left[(\Lambda/\mu - N_1 - N_2(N_1,Y))^+ |Y \right] \right\} \right] \right\}. \tag{3.18}$$

Let $(\bar{N}_1, \bar{N}_2(\bar{N}_1, Y))$ denote an optimal solution to (3.18), whose existence is rigorously established in the proof of Proposition 7. 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 6. *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 6 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 6 may not hold if c_2 is large or Z has a large standard deviation. For a concrete example that violates Assumption 6, consider $Y \sim$ Uniform$[-1, 1]$, $Z \sim$ 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 6 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 7. *For $\alpha \in (1/2, 1)$ and $\nu \in [1/2, \alpha]$, we have*

(I) Cost saving: *If $\nu < \alpha$, then $\mathscr{C}^{e,\lambda}_{1,*} - \mathscr{C}^{e,\lambda}_{2,*} = \Theta(\lambda^\alpha)$. If $\nu = \alpha$ and Assumption 6 holds, then $\mathscr{C}^{e,\lambda}_{1,*} - \mathscr{C}^{e,\lambda}_{2,*} = \Theta(\lambda^\alpha)$. If $\nu = \alpha$ and Assumption 6 does not hold, then $\mathscr{C}^{e,\lambda}_{1,*} - \mathscr{C}^{e,\lambda}_{2,*} = o(\lambda^\alpha)$.*

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

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

Comparing item (III) in Proposition 7 to item (III) in Proposition 6, 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 $\Theta(\lambda^{\nu})$.

3.6.3 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 3.6.2. It has near-optimal performance as established in Proposition 7.

(II) The two-stage QED rule $u_{2,QED}$, which is a straightfoward extension of the two-stage QED rule defined in Definition 3 by ignoring the prediction error: For X defined in (3.17) (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 3.5, 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 3.5.

(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 (3.11). 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.

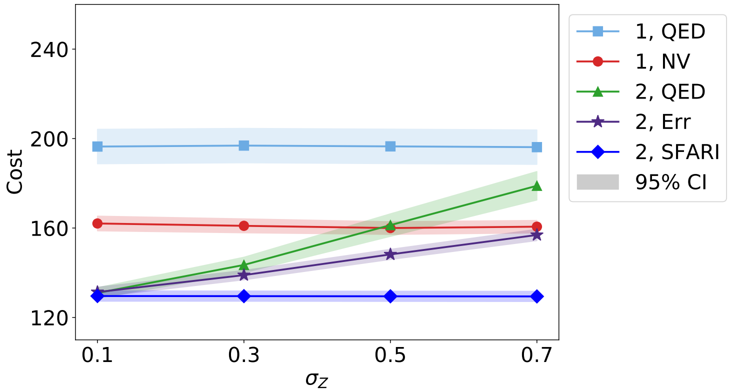
We assume that Y and Z are normally distributed with standard deviation σ_Y and σ_Z, respectively. We then fix the standard deviation of X to be equal to 1, i.e., $\sigma_Y^2 + \sigma_Z^2 = 1$, and vary the σ_Z from 0.1 to 0.7 in increment of 0.2. For each policy and each value of σ_Z, we simulate 1000 independent and identically distributed realizations of the random arrival rate, and use the average to approximate the expected total cost. Figure 3.4 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.

3.7 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

Figure 3.4: Sensitivity analysis with respect to prediction error
($\lambda = 100, \mu = 1, \gamma = 0.1, h = 1.5, a = 3, c_1 = 1, c_2 = 1.5, \alpha = \nu = 0.75$)



could be beneficial. In our partner hospital, α is estimated to be 0.769. According to Theorem 4, 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.

3.7.1 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.

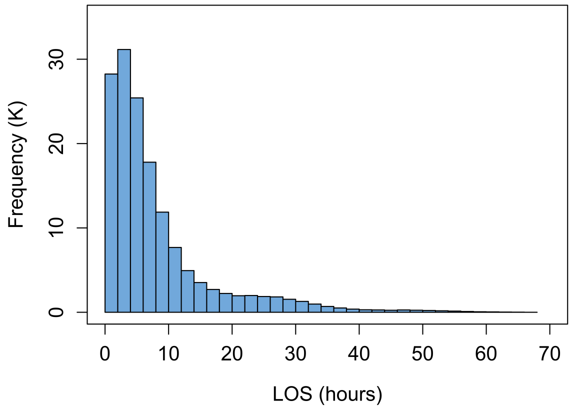
We first group the shifts into 14 different types based on day of the week and day versus night shift. Table 3.4 provides some summary statistics for different shifts. We observe that the day shifts see more arrivals than the night shifts, and weekday day shifts see more arrivals than the weekend day 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 3.4: 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 3.5, 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.

Figure 3.5: Patient LOS distribution



3.7.2 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 3.4. Let λ_i denote the mean arrival rate for type $i \in \mathscr{I} := \{1, ..., 14\}$ 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

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

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 $\Lambda_i \sim N(\lambda_i, \lambda_i^{2\alpha} \mu^{2(1-\alpha)} \sigma^2), i \in \mathscr{I}$.

Let $L_i^{(k)}$ denote the observed arrival count for the kth shift of type i, $1 \leq k \leq n_i$. 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

$$\bar{L}_i = \hat{\lambda}_i, \quad \Sigma_i^2 = \hat{\lambda}_i^{2\hat{\alpha}} \mu^{2(1-\hat{\alpha})} \hat{\sigma}^2, \quad i \in \mathscr{I}. \tag{3.19}$$

It follows from (3.19) that

$$\log \Sigma_i = \hat{\alpha} \log \bar{L}_i + \log(\mu^{1-\hat{\alpha}} \hat{\sigma}), \quad i \in \mathscr{I}. \tag{3.20}$$

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

$$\min_{\alpha \in (0,1),\, \gamma \in \mathbb{R}} \sum_{i=1}^{14} \left(\log \Sigma_i - \gamma - \alpha \log \bar{L}_i\right)^2. \tag{3.21}$$

In particular, let γ^* and α^* denote the optimal solution to the least squares problem (3.21). Then, $\hat{\alpha} = \alpha^*$ and $\mu^{1-\hat{\alpha}} \hat{\sigma} = \exp(\gamma^*)$.

In Table 3.5, the first row below header (with $|\mathscr{I}| = 14$) 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 $|\mathscr{I}| = 56$) 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 C.7). 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.

3.7.3 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.

Table 3.5: 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: $\lvert\mathscr{I}\rvert = 14$	0.769	(0.543, 0.994)	0.344	(0.114, 1.034)
Day-of-week, day/night and seasons: $\lvert\mathscr{I}\rvert = 56$	0.746	(0.558, 0.933)	0.362	(0.135, 0.902)

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 linear regression model developed in Chapter 2 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 2.2 in Chapter 2. This linear regression model is able to capture 90.8% 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., (3.15), 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 $\lambda_i^{\alpha}\mu^{1-\alpha}Z$, $i \in \mathscr{I}$. Using the point estimates for λ_i and α in Section 3.7.2, 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 3.7.2, the standard deviation of Y is 0.111.

3.7.4 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 simulate patient flow process in Milstein ED over 52 weeks from February 1, 2018 to January 31, 2019.

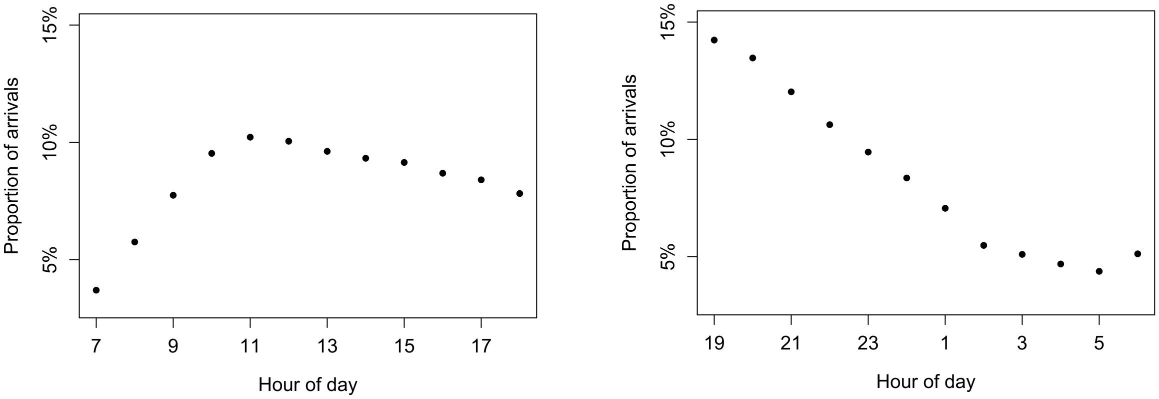
3.7.4 Model Calibration

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 3.6. 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 being 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 is 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 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 3.7 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 × 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 3.6: Proportion of patient arrivals by hour within day/night shift



(a) Day shift

(b) Night shift

3.7.4 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 3.2. In particular, based on our model calibration in Section 3.7.4.1, 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 C.8.2). 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 $1 \le i \le 14$, 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 $1 \le i \le 14$, $1 \le k \le 52$, 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 C.8.3, we show through numerical experiments that setting $\xi_1 = 5$ and $\xi_2 = 1$ achieves near-optimal and robust performance.

3.7.4 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 being 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 3.7. 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 3.6 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 being 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 3.7.3 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.

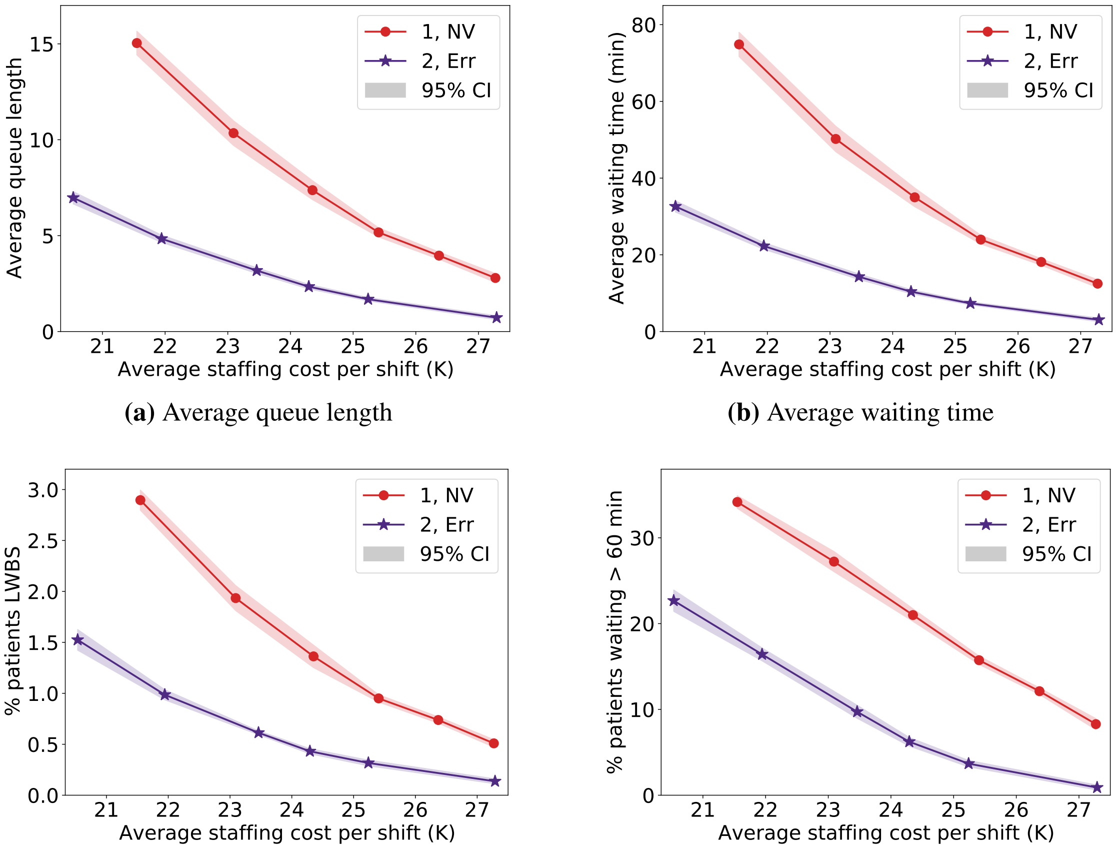
Table 3.6: Annual saving in staffing cost to achieve target performance

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%)

3.8 Conclusion

In this chapter, 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

Figure 3.7: Tradeoff between staffing cost and quality of service



(a) Average queue length

(b) Average waiting time

(c) % patients LWBS

(d) % paitents waiting $>$ 60 min

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 trade-offs 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

J. Abate and W. Whitt. A unified framework for numerically inverting laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.

Joseph Abate and Ward Whitt. Transient behavior of the m/m/1 queue via laplace transforms. *Advances in Applied Probability*, 20(1):145–178, 1988.

Mustafa Akan, Oguzhan Alagoz, Baris Ata, Fatih Safa Erenay, and Adnan Said. A broader view of designing the liver allocation system. *Operations research*, 60(4):757–770, 2012.

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

D Altman, S Ashtar, M Olivares, and GB Yom-Tov. Do customer emotions affect worker speed? an empirical study of emotional load in online customer contact centers. *Working Paper*, 2019.

Natalie Anderson, Fofoa Pio, Peter Jones, Vanessa Selak, Eunicia Tan, Sierra Beck, Suzanne Hamilton, Alice Rogan, Kim Yates, Mark Sagarin, et al. Facilitators, barriers and opportunities in workplace wellbeing: A national survey of emergency department staff. *International Emergency Nursing*, 57:101046, 2021.

Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, and Galit B Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems*, 5(1):146–194, 2015.

Andreas Asheim, Lars P Bache-Wiig Bjørnsen, Lars E Næss-Pleym, Oddvar Uleberg, Jostein Dale, and Sara M Nilsen. Real-time forecasting of emergency department arrivals using prehospital data. *BMC emergency medicine*, 19(1):1–6, 2019.

Rami Atar, Chanit Giat, and Nahum Shimkin. The cμ/θ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

Rami Atar, Chanit Giat, and Nahum Shimkin. On the asymptotic optimality of the cμ/θ rule under ergodic cost. *Queueing Systems*, 67(2):127–144, 2011.

Athanassios N Avramidis, Alexandre Deslauriers, and Pierre L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

J Ball, P Griffiths, and J Hope. Evidence on the effect of nurse staffing levels on patient outcomes. *Nursing Times*, 113(1):48–49, 2017.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

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

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

Holly Batal, Jeff Tench, Sean McMillan, Jill Adams, and Phillip S Mehler. Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine*, 8(1):48–53, 2001.

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

Steven L Bernstein, Dominik Aronsky, Reena Duseja, Stephen Epstein, Dan Handel, Ula Hwang, Melissa McCarthy, K John McConnell, Jesse M Pines, Niels Rathlev, et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10, 2009.

Dimitris Bertsimas, Jean Pauphilet, Jennifer Stevens, and Manu Tandon. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 2021.

Pol Boada-Collado, Sunil Chopra, and Karen Smilowitz. The value of information and flexibility with temporal commitments. *Available at SSRN 3452915*, 2020.

Sem Borst, Avi Mandelbaum, and Martin I Reiman. Dimensioning large call centers. *Operations research*, 52(1):17–34, 2004.

Justin Boyle, Melanie Jessup, Julia Crilly, David Green, James Lind, Marianne Wallis, Peter Miller, and Gerard Fitzgerald. Predicting emergency department admissions. *Emergency Medicine Journal*, 29(5):358–365, 2012.

Judith C Brillman, Tom Burr, David Forslund, Edward Joyce, Rick Picard, and Edith Umland. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. *BMC medical informatics and decision making*, 5(1):1–14, 2005.

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

Rafael Calegari, Flavio S Fogliatto, Filipe R Lucini, Jeruza Neyeloff, Ricardo S Kuchenbecker, and Beatriz D Schaan. Forecasting daily volume and acuity of patients in the emergency department. *Computational and mathematical methods in medicine*, 2016: 1–8, 2016.

Ping Cao and Jingui Xie. Optimal control of a multiclass queueing system when customers can change types. *Queueing Systems*, 82(3-4):285–313, 2016.

Donald B Chalfin, Stephen Trzeciak, Antonios Likourezos, Brigitte M Baumann, and R Phillip Dellinger. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical care medicine*, 35(6):1477–1483, 2007.

Carri W Chan, Vivek F Farias, and Gabriel J Escobar. The impact of delays on service times in the intensive care unit. *Management Science*, 63(7):2049–2072, 2016.

Carri W Chan, Michael Huang, and Vahid Sarhangian. Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. *Operations Research*, 2021.

Paul S Chan, Harlan M Krumholz, Graham Nichol, Brahmajee K Nallamothu, and American Heart Association National Registry of Cardiopulmonary Resuscitation Investigators. Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine*, 358(1):9–17, 2008.

Anna Marie Chang, Deborah J Cohen, Amber Lin, James Augustine, Daniel A Handel, Eric Howell, Hyunjee Kim, Jesse M Pines, Jeremiah D Schuur, K John McConnell, et al. Hospital strategies for reducing emergency department crowding: a mixed-methods study. *Annals of emergency medicine*, 71(4):497–505, 2018a.

Bernard P Chang, George Gallos, Lauren Wasson, and Donald Edmondson. The unique environmental influences of acute care settings on patient and physician well-being: a call to action. *Journal of Emergency Medicine*, 54(1):e19–e21, 2018b.

Y. Chang, M. Yeh, Y. Li, C. Hsu, M. Hsu, and W. Chiu. Predicting hospital-acquired infections by scoring system with simple parameters. *PLoS One*, 6(8), 2011.

Valerie J Chase, Amy EM Cohn, Timothy A Peterson, and Mariel S Lavieri. Predicting emergency department volume using forecasting methods to create a âĂIJsurge responseâĂİ for noncrisis events. *Academic Emergency Medicine*, 19(5):569–576, 2012.

Bert PK Chen and Shane G Henderson. Two issues in setting call centre staffing levels. *Annals of operations research*, 108(1):175–192, 2001.

Xin Chen, Melvyn Sim, and Peng Sun. A robust optimization perspective on stochastic programming. *Operations research*, 55(6):1058–1071, 2007.

Guang Cheng, Jingui Xie, and Zhichao Zheng. Optimal stopping for medical treatment with predictive information. *Available at SSRN 3397530*, 2019.

Avishek Choudhury. Hourly forecasting of emergency department arrivals: time series analysis. *arXiv preprint arXiv:1901.02714*, 2019.

M. Churpek, T. Yuen, S. Y. Park, and D. Edelson. Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards. *Critical care medicine*, 42(4):841, 2014.

David Roxbee Cox and Walter Smith. *Queues*, volume 2. CRC Press, 1991.

Emiliano Cristiani and Pierre Martinon. Initialization of the shooting method via the hamilton-jacobi-bellman approach. *Journal of Optimization Theory and Applications*, 146(2):321–346, 2010.

J Randall Curtis and Kathleen Puntillo. Is there an epidemic of burnout and post-traumatic stress in critical care clinicians?, 2007.

Andrew Daw and Jamol Pender. Queues driven by hawkes processes. *Stochastic Systems*, 8(3):192–229, 2018.

K. Delana, N. Savva, and T. Tezcan. Proactive customer service: operational benefits and economic frictions. forthcoming in Manufacturing & Service Operations Management, 2019.

Douglas G Down and Mark E Lewis. The n-network model with upgrades. *Probability in the Engineering and Informational Sciences*, 24(2):171–200, 2010.

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

Gabriel J. Escobar, Juan Carlos LaGuardia, Benjamin J. Turk, Arona Ragins, Patricia Kipnis, and David Draper. Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine*, 7(5):388–395, 2012.

Hélène Frankowska. Optimal control under state constraints. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 2915–2942. World Scientific, 2010.

AT Fuller. Study of an optimum non-linear control system. *International Journal of Electronics*, 15(1):63–71, 1963.

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

Hayley B Gershengorn, Yue Hu, Jen-Ting Chen, S Jean Hsieh, Jing Dong, Michelle Ng Gong, and Carri W Chan. The impact of high-flow nasal cannula use on patient mortality and the availability of mechanical ventilators in covid-19. *Annals of the American Thoracic Society*, 18(4):623–631, 2021.

Google Trends, 2020. Available Online: `https://www.ncdc.noaa.gov/cdo-web/cart` (accessed on 26 April 2020).

Dieter Grass, Jonathan P Caulkins, Gustav Feichtinger, Gernot Tragler, and Doris A Behrens. Optimal control of nonlinear processes. *Berlino: Springer*, 2008.

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

Linda V Green, Joao Soares, James F Giglio, and Robert A Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

Muhammet Gul and Erkan Celik. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, 9(4):263–284, 2020.

I. Gurvich and W. Whitt. Service-level differentiation in many-server service systems via queue-ratio routing. *Manufacturing and Service Operations Management*, 58(2):237–253, 2010.

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

J Michael Harrison and Assaf Zeevi. Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research*, 52(2):243–257, 2004.

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

Richard F Hartl, Suresh P Sethi, and Raymond G Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM review*, 37(2):181–218, 1995.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Donald R Holleman, Renee L Bowling, and Charlane Gathy. Predicting daily visits to a waik-in clinic and emergency department using calendar and weather data. *Journal of General Internal Medicine*, 11(4):237–239, 1996.

Woo Suk Hong, Adrian Daniel Haimovich, and Richard Andrew Taylor. Predicting 72-hour and 9-day return to the emergency department using machine learning. *JAMIA open*, 2(3):346–352, 2019.

H. Honnappa, J. Jain, and A. Ward. A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems*, 80(1-2):71–103, 2015.

Wenqi Hu, Carri W Chan, José R Zubizarreta, and Gabriel J Escobar. An examination of early transfers to the icu based on a physiologic risk score. *Manufacturing & Service Operations Management*, 20(3):531–549, 2018.

Yue Hu, Kenrick D Cato, Carri W Chan, Jing Dong, Nicholas Gavin, Sarah C Rossetti, and Bernard P Chang. Use of real-time information to predict future arrivals in the emergency department. 2021. Working Paper, Columbia Business School.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

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

Institute of Medicine Committee on the Future of Emergency Care in the US Health System. Hospital-based emergency care: at the breaking point, 2006.

Ganesh Janakiraman, Sridhar Seshadri, and Anshul Sheopuri. Analysis of tailored base-surge policies in dual sourcing inventory systems. *Management Science*, 61(7):1547–1561, 2015.

Otis B Jennings, Avishai Mandelbaum, William A Massey, and Ward Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

Kinshuk Jerath, Anuj Kumar, and Serguei Netessine. An information stock model of customer behavior in multichannel customer support services. *Manufacturing & Service Operations Management*, 17(3):368–383, 2015.

LI Jing, LI Bao Yu, WEI Zi Jian, ZHAO Yu Zhu, and LI Tan Shi. Application research on gated recurrent unit deep learning prediction and graded early warning of emergency department visits based on meteorological environmental data. *Biomedical and Environmental Sciences*, 33(10):817–820, 2020.

Sion Jo, Kyuseok Kim, Jae Hyuk Lee, Joong Eui Rhee, Yu Jin Kim, Gil Joon Suh, and Young Ho Jin. Emergency department crowding is associated with 28-day mortality in community-acquired pneumonia patients. *Journal of Infection*, 64(3):268–275, 2012.

Kimberly D Johnson and Chris Winkelman. The effect of emergency department crowding on patient outcomes: a literature review. *Advanced emergency nursing journal*, 33(1): 39–54, 2011.

Spencer S Jones, Alun Thomas, R Scott Evans, Shari J Welch, Peter J Haug, and Gregory L Snow. Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine*, 15(2):159–170, 2008.

Joshua W Joseph and Benjamin A White. Emergency department operations: An overview. *Emergency Medicine Clinics*, 38(3):549–562, 2020.

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

Yaşar Levent Koçağa, Mor Armony, and Amy R Ward. Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management*, 24(7):1101–1117, 2015.

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

Michelle D Lall, Bernard P Chang, Joel Park, Ramin R Tabatabai, Rita A Manfredi, Jill M Baren, and Jenny Castillo. Are emergency physicians satisfied? an analysis of operational/organization factors. *Journal of the American College of Emergency Physicians Open*, 2(6):e12546, 2021.

Maialen Larrañaga. *Dynamic control of stochastic and fluid resource-sharing systems*. PhD thesis, 2015.

Maialen Larrañaga, Urtzi Ayesta, and Ina Maria Verloop. Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation*, 70(10):841–858, 2013.

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

Michelle P Lin, Olesya Baker, Lynne D Richardson, and Jeremiah D Schuur. Trends in emergency department visits and admission rates among us acute care hospitals. *JAMA Internal Medicine*, 178(12):1708–1710, 2018.

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

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

A. Mandelbaum and A. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52(6):836–855, 2004.

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

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

John J McCall. Maintenance policies for stochastically failing equipment: a survey. *Management science*, 11(5):493–524, 1965.

Melissa L McCarthy, Scott L Zeger, Ru Ding, Dominik Aronsky, Nathan R Hoot, and Gabor D Kelen. The challenge of predicting demand for emergency department services. *Academic Emergency Medicine*, 15(4):337–346, 2008.

Jane McCusker, Alain Vadeboncoeur, Jean-Frédéric Lévesque, Antonio Ciampi, and Eric Belzile. Increases in emergency department occupancy are associated with adverse 30-day outcomes. *Academic Emergency Medicine*, 21(10):1092–1100, 2014.

Bernard J Morzuch and P Geoffrey Allen. Forecasting hospital emergency department arrivals. In *26th Annual Symposium on Forecasting, Santander, Spain*, pages 11–14, 2006.

National Centers for Environmental Information. Global historical climatology network (ghcn)-daily dataset, 2020. Available Online: `https://www.ncdc.noaa.gov/cdo-web/cart` (accessed on 08 December 2020).

John Neter, Michael H Kutner, Christopher J Nachtsheim, William Wasserman, et al. Applied linear statistical models. 1996.

E Lerzan Örmeci, Evrim Didem Güneş, and Derya Kunduzcu. A modeling framework for control of preventive services. *Manufacturing & Service Operations Management*, 18 (2):227–244, 2015.

Suleyman Özekici and Stanley R Pliska. Optimal scheduling of inspections: A delayed markov model with false positives and negatives. *Operations Research*, 39(2):261–273, 1991.

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

Ohad Perry and Ward Whitt. Responding to unexpected overloads in large-scale service systems. *Management Science*, 55(8):1353–1367, 2009.

William P Pierskalla and John A Voelker. A survey of maintenance models: the control and surveillance of deteriorating systems. *Naval Research Logistics Quarterly*, 23(3): 353–388, 1976.

Jesse M Pines, Robert J Batt, Joshua A Hilton, and Christian Terwiesch. The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Annals of emergency medicine*, 58(4):331–340, 2011.

Amber L Puha and Amy R Ward. Scheduling an overloaded multiclass many-server queue with impatient customers. In *Operations Research & Management Science in the Age of Analytics*, pages 189–217. INFORMS, 2019.

Zachariah Ramsey, Joseph S Palter, John Hardwick, Jordan Moskoff, Errick L Christian, and John Bailitz. Decreased nursing staffing adversely affects emergency department throughput metrics. *Western Journal of Emergency Medicine*, 19(3):496, 2018.

Alejandra Recio-Saucedo, Catherine Pope, Chiara Dall'Ora, Peter Griffiths, Jeremy Jones, Robert Crouch, and Jonathan Drennan. Safe staffing for nursing in emergency departments: evidence review. *Emergency Medicine Journal*, 32(11):888–894, 2015.

J. E. Reed and A. Ward. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33 (3):606–644, 2008.

J. Rumsfeld, K. Joynt, and T. Maddox. Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology*, 13:350–359, 2016.

Heinz Schättler and Urszula Ledzewicz. *Geometric optimal control: theory, methods and examples*, volume 38. Springer Science & Business Media, 2012.

Lisa M Schweigler, Jeffrey S Desmond, Melissa L McCarthy, Kyle J Bukowski, Edward L Ionides, and John G Younger. Forecasting models of emergency department crowding. *Academic Emergency Medicine*, 16(4):301–308, 2009.

Suresh P Sethi and Gerald L Thompson. *Optimal control theory: Applications to management science and economics*. Springer, 2000.

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

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

Alexander L Stolyar et al. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

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

Zhankun Sun, Nilay Tanık Argon, and Serhan Ziya. Patient triage and prioritization under austere conditions. *Management Science*, 64(10):4471–4489, 2017.

Dan Tandberg and Clifford Qualls. Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of emergency medicine*, 23(2):299–306, 1994.

Tolga Tezcan and JG Dai. Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*, 58(1):94–110, 2010.

Terry Therneau, Beth Atkinson, Brian Ripley, and Maintainer Brian Ripley. Package âĂŸrpartâĂŹ. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*, 2015.

Emmanuel Trélat. Optimal control and applications to aerospace: some results and challenges. *Journal of Optimization Theory and Applications*, 154(3):713–758, 2012.

Jan A Van Mieghem. Dynamic scheduling with convex delay costs: The generalized c| mu rule. *The Annals of Applied Probability*, pages 809–833, 1995.

Francis de Véricourt and Otis B Jennings. Nurse staffing in medical units: A queueing perspective. *Operations research*, 59(6):1320–1331, 2011.

Han-Yi Wang, Ghee Chew, C Kung, K Chung, and W Lee. The use of charlson comorbidity index for patients revisiting the emergency department within 72 hours. *Chang Gung medical journal*, 30(5):437, 2007.

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

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

W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

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

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

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

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

Adam Wierman. Fairness and scheduling in single server queues. *Surveys in Operations Research and Management Science*, 16(1):39–48, 2011.

Jingui Xie, Taozeng Zhu, An-Kuo Chao, and Shuaian Wang. Performance analysis of service systems with priority upgrades. *Annals of Operations Research*, 253(1):683–705, 2017.

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

Kuang Xu and Carri W Chan. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management*, 18 (3):314–331, 2016.

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

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

Galit B Yom-Tov, Yueming Xie, and Liron Yedidsion. An invitation control policy for proactive service systems: Balancing efficiency, value and service level. Technical report, working paper. ec1, 2018.

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

Xiaowei Zhang, L Jeff Hong, and Jiheng Zhang. Scaling and modeling of call center arrivals. In *Proceedings of the Winter Simulation Conference 2014*, pages 476–485. IEEE, 2014.

Zeyu Zheng, Harsha Honnappa, and Peter W Glynn. Approximating systems fed by poisson processes with rapidly changing arrival rates. *arXiv preprint arXiv:1807.06805*, 2018.

Lara M Zibners, Bema K Bonsu, John R Hayes, and Daniel M Cohen. Local weather effects on emergency department visits: a time series and regression analysis. *Pediatric emergency care*, 22(2):104–106, 2006.

Appendix A: Appendix for Chapter 1

A.1 Long-Run Regularity of the Fluid Model and Proof of Theorem 1

The key to proving Theorem 1 is to establish that the optimal solution to the long-run average fluid optimization problem, (F1), is a globally asymptotically stable equilibrium under the strict priority rule suggested by the modified $c\mu/\theta$-index. In this section, we first establish the long-run regularity of the fluid model under strict priority rules. The stability analysis of strict priority rules can be of independent interest, especially as such policies are often used in practice. Additionally, we identify an interesting bi-stability phenomenon for certain parameter regions under strict priority rules. We then use the long-run regularity results to prove Theorem 1.

A.1.1 System Stability under Strict Priority Rules

Due to the symmetry of the system, we provide the analysis for strict priority to Class 1 only, i.e., the analysis for strict priority to Class 2 follows identically by symmetry. Under P_1, when $q_1(t) > 0$, we will assign all capacity to Class 1. When $q_1(t) = 0$, we will assign to Class 1 the minimum amount of capacity necessary to maintain its queue at zero if there is enough capacity; otherwise, we will assign all the capacity to Class 1. In particular, the system dynamics are characterized as follows:

(i) If $q_1(t) > 0$,

$$\dot{q}_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t), \quad \dot{q}_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t); \tag{A.1}$$

(ii) If $q_1(t) = 0, q_2(t) > 0$,

$$\begin{aligned} \dot{q}_1(t) &= \lambda_1 - \mu_1 \left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \wedge s \right) + \gamma_2 q_2(t), \\ \dot{q}_2(t) &= \lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \right)^+ - \theta_2 q_2(t) - \gamma_2 q_2(t); \end{aligned} \tag{A.2}$$

(iii) If $q_1(t) = 0, q_2(t) = 0$,

$$\dot{q}_1(t) = \lambda_1 - \mu_1 \left(\frac{\lambda_1}{\mu_1} \wedge s \right), \quad \dot{q}_2(t) = \lambda_2 - \mu_2 \left(\left(s - \frac{\lambda_1}{\mu_1} \right)^+ \wedge \frac{\lambda_2}{\mu_2} \right). \tag{A.3}$$

Using a Lyapunov argument, Theorem 7 characterizes the long-run regularity of the fluid dynamical system under strict priority to Class 1.

Theorem 7. *Under Assumption 1, for the dynamical system* (A.1) - (A.3)*,*

Case I. *When* $\mu_1 > \frac{\gamma_2}{\theta_2+\gamma_2}\mu_2$*,*

Ia If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = 0.$$

Ib If $\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} \leq s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = \frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s\right)}{(\theta_2+\gamma_2)\mu_1 - \gamma_2\mu_2} > 0.$$

Ic If $s < \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$**.** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1\frac{\theta_2}{\theta_2+\gamma_2}} > 0, \quad q_2^e = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} > \frac{\lambda_2}{\theta_2+\gamma_2}.$$

Case II. *When* $\mu_1 < \frac{\gamma_2}{\theta_2+\gamma_2}\mu_2$*,*

IIa If $\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} < s$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = 0.$$

IIb If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$**,** *the system has two locally asymptotically stable equilibria*

$$q_{11}^e = 0, \quad q_{21}^e = 0$$

and

$$q^e_{12} = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} \geq 0, \quad q^e_{22} = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1 + \lambda_2 - s\mu_1)}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} \geq \frac{\lambda_2}{\theta_2+\gamma_2}.$$

IIc If $s = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has an equilibrium at* $(q^e_{11}, q^e_{21}) = (0,0)$ *and a locally asymptotically stable equilibrium at*

$$q^e_{12} = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} > 0, \quad q^e_{22} = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1 + \lambda_2 - s\mu_1)}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} > \frac{\lambda_2}{\theta_2+\gamma_2}.$$

IId If $s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q^e_{1} = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} > 0, \quad q^e_{2} = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1 + \lambda_2 - s\mu_1)}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} > \frac{\lambda_2}{\theta_2+\gamma_2}.$$

Remark 5. *We note that when* $\mu_1 = \gamma_2\mu_2/(\theta_2+\gamma_2)$*, the system can have uncountably many equilibrium points. In particular, for* $s = \lambda_1/\mu_1 + \lambda_2/\mu_2$*, any* (q^e_1, q^e_2) *satisfying* $q^e_1 = 0$ *and* $(\lambda_1 + \gamma_2 q^e_2)/\mu_1 < s$ *is an equilibrium point. We do not consider this parameter regime, i.e.* $\mu_1 = \gamma_2\mu_2/(\theta_2+\gamma_2)$*, in this chapter.*

PROOF: [Proof of Theorem 7] The stability analysis for P_1 divides the parameter regime into six cases. In each case, we construct a Lyapunov function to establish the asymptotic stability. As the proof for each case follows similar lines of analysis, we only present the proof for Case Ia which has a globally asymptotically stable equilibrium and Case IIb which has two locally asymptotically stable equilibria. The proofs for the rest of the cases given the appropriate Lyapunov functions follow similarly and are thus omitted.

The Lyapunov function we utilize to prove each case differs; they are summarized in the table below.

	Lyapunov function
Case Ia	Subcase 1: $\frac{1}{\mu_1}\|q_1 - q_1^e\| + \frac{1}{\mu_2}\|q_2 - q_2^e\|$; Subcase 2: $\|q_1 - q_1^e\| + \|q_2 - q_2^e\|$
Case Ib	Subcase 1: $\frac{1}{\mu_1}\|q_1 - q_1^e\| + \frac{1}{\mu_2}\|q_2 - q_2^e\|$; Subcase 2: $\|q_1 - q_1^e\| + \|q_2 - q_2^e\|$
Case Ic	$\|q_1 - q_1^e\| + \|q_2 - q_2^e\|$
Case IIa	$\|q_1 - q_1^e\| + \frac{\gamma}{\theta_2+\gamma}\|q_2 - q_2^e\|$
Case IIb	Local equilibrium $(0,0)$: $\frac{1}{\mu_1}\|q_1 - q_1^e\| + \frac{1}{\mu_2}\|q_2 - q_2^e\|$; Local equilibrium $\left(\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}}, \frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$: $\|q_1 - q_1^e\| + \|q_2 - q_2^e\|$
Case IIc	$\|q_1 - q_1^e\| + \|q_2 - q_2^e\|$
Case IId	$\|q_1 - q_1^e\| + \|q_2 - q_2^e\|$

Let V denote the Lyapunov function we constructed. To prove the asymptotic stability of an equilibrium point q^e, we need to verify that 1) $V(q^e) = 0$ and $V(q) \to \infty$ as $||q|| \to \infty$; 2) $\nabla_q V(q)^T f(q) < 0$ for $q \neq q^e$. In the case of local stability, the second condition is checked locally with q restricted to be in some neighborhood of q_e, i.e. $0 < ||q - q^e|| < \delta$ for some $\delta > 0$. As 1) is straightforward from our definition of the Lyapunov function, we focus on verifying 2) only.

Case I. $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2}$, i.e., $\frac{\gamma_2}{\mu_1} - \frac{\theta_2+\gamma_2}{\mu_2} < 0$.

Ia. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$.

Ia.Subcase 1. $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$.

Consider Lyapunov function of the form

$$V(q) = \frac{1}{\mu_1}|q_1 - q_1^e| + \frac{1}{\mu_2}|q_2 - q_2^e|,$$

where (q_1^e, q_2^e) is the corresponding equilibrium point $(0,0)$.

(i) If $q_1(t) > 0$,

$$\begin{aligned}
\dot{q}_1(t) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\
\dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).
\end{aligned}$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) \\
&\quad + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t) + \gamma_1 q_1(t)) \\
&= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_1}{\mu_2} - \frac{\theta_1 + \gamma_1}{\mu_1}\right) q_1(t) + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\
&< 0,
\end{aligned}$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$, and $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$.

(ii) If $q_1(t) = 0$, $q_2(t) > 0$,

(a) if $\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1} \geq s$,

$$\begin{aligned}
\dot{q}_1(t) &= \lambda_1 - \mu_1 s + \gamma_2 q_2(t) \\
\dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t).
\end{aligned}$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s + \gamma_2 q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t)) \\
&= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\
&< 0,
\end{aligned}$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$, and $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$.

(b) If $\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1} < s$,

$$\begin{aligned}
\dot{q}_1(t) &= \lambda_1 - \mu_1 \left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) + \gamma_2 q_2(t) = 0 \\
\dot{q}_2(t) &= \lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t).
\end{aligned}$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_2}\left(\lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t)\right) \\
&= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\
&< 0,
\end{aligned}$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\leq s$, and $\frac{\gamma_2}{\theta_2+\gamma_2}<\frac{\mu_1}{\mu_2}<\frac{\theta_1+\gamma_1}{\gamma_1}$.

Ia.Subcase 2. $\frac{\gamma_2}{\theta_2+\gamma_2}<\frac{\theta_1+\gamma_1}{\gamma_1}<\frac{\mu_1}{\mu_2}$.

Consider Lyapunov function of the form

$$V(q)=|q_1-q_1^e|+|q_2-q_2^e|,$$

where (q_1^e,q_2^e) is the corresponding equilibrium point $(0,0)$.

(i) If $q_1(t)>0$,

$$\begin{aligned}\dot{q}_1(t)&=\lambda_1-\mu_1 s-\theta_1 q_1(t)-\gamma_1 q_1(t)+\gamma_2 q_2(t)\\ \dot{q}_2(t)&=\lambda_2-\theta_2 q_2(t)-\gamma_2 q_2(t)+\gamma_1 q_1(t).\end{aligned}$$

$$\begin{aligned}\nabla_q V(q)^T f(q)&=(\lambda_1-\mu_1 s-\theta_1 q_1(t)-\gamma_1 q_1(t)+\gamma_2 q_2(t))+(\lambda_2-\theta_2 q_2(t)-\gamma_2 q_2(t)+\gamma_1 q_1(t))\\ &=\lambda_1-\mu_1 s+\lambda_2-\theta_1 q_1(t)-\theta_2 q_2(t)\\ &=\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_1}-s\right)-\theta_1 q_1(t)-\theta_2 q_2(t)\\ &<\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)-\theta_1 q_1(t)-\theta_2 q_2(t)\\ &<0,\end{aligned}$$

where the first inequality follows from the fact that $\mu_1>\mu_2$ (due to $\frac{\theta_1+\gamma_1}{\gamma_1}<\frac{\mu_1}{\mu_2}$), and the second inequality follows from the facts that $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\leq s$ and $q_1(t)>0$.

(ii) If $q_1(t)=0$, $q_2(t)>0$,

(a) If $\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1}\geq s$,

$$\begin{aligned}\dot{q}_1(t)&=\lambda_1-\mu_1 s+\gamma_2 q_2(t)\\ \dot{q}_2(t)&=\lambda_2-\theta_2 q_2(t)-\gamma_2 q_2(t).\end{aligned}$$

$$\begin{aligned}\nabla_q V(q)^T f(q)&=(\lambda_1-\mu_1 s+\gamma_2 q_2(t))+(\lambda_2-\theta_2 q_2(t)-\gamma_2 q_2(t))\\ &=\lambda_1-\mu_1 s+\lambda_2-\theta_2 q_2(t)\\ &=\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_1}-s\right)-\theta_2 q_2(t)\\ &<\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)-\theta_2 q_2(t)\\ &<0,\end{aligned}$$

where the first inequality follows from the fact that $\mu_1 > \mu_2$, and the second inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$ and $q_2(t) > 0$.

(b) If $\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} < s$,

$$\dot{q}_1(t) = \lambda_1 - \mu_1 \left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \right) + \gamma_2 q_2(t) = 0$$

$$\dot{q}_2(t) = \lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \right) - \theta_2 q_2(t) - \gamma_2 q_2(t).$$

$$\nabla_q V(q)^T f(q) = \mu_2 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2} \right) q_2(t) \right) < 0,$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$, and $\frac{\gamma_2}{\theta_2 + \gamma_2} < \frac{\theta_1 + \gamma_1}{\gamma_1} < \frac{\mu_1}{\mu_2}$.

IIb. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2 + \gamma} \frac{\lambda_2}{\mu_1}$.

To check for local stability, it is sufficient to find a Lyapunov function V that satisfies $\nabla_q V(q)^T f(q) < 0$ in an open neighborhood of the equilibrium point. We construct different Lyapunov functions for different equilibrium points.

(i) Local stability of $(q_1^e, q_2^e) = (0,0)$: Consider Lyapunov function of the form

$$V(q) = \frac{1}{\mu_1} |q_1 - q_1^e| + \frac{1}{\mu_2} |q_2 - q_1^e|.$$

Let $0 < \varepsilon < \frac{s\mu_1 - \lambda_1}{\gamma_2}$ be such that

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2} \right) \varepsilon < 0. \tag{A.4}$$

We know such ε exists because $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s < 0$ and $\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2} > 0$. Consider states (q_1, q_2) with $q_2 < \varepsilon$.

(a) If $q_1(t) > 0$,

$$\dot{q}_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$\dot{q}_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t) + \gamma_1 q_1(t)) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_1}{\mu_2} - \frac{\theta_1 + \gamma_1}{\mu_1}\right) q_1(t) + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\ &< \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the first inequality follows from the facts that $\frac{\mu_1}{\mu_2} < \frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$ and $q_1(t) > 0$, and the second inequality follows from (A.4) and the fact that $q_2(t) < \varepsilon$.

(b) If $q_1(t) = 0$, $q_2(t) > 0$ (the assumption $q_2(t) < \varepsilon$ implies that $\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1} < s$),

$$\begin{aligned}\dot{q}_1(t) &= \lambda_1 - \mu_1 \left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) + \gamma_2 q_2(t) = 0 \\ \dot{q}_2(t) &= \lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t).\end{aligned}$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_2}\left(\lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t)\right) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the inequality follows from (A.4) and the fact that $q_2(t) < \varepsilon$.

(ii) Local stability of $(q_1^e, q_2^e) = \left(\frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}}, \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$: Consider Lyapunov function of the form

$$V(q) = |q_1 - q_1^e| + |q_2 - q_2^e|.$$

Consider states q such that $q_1 > 0$ and $q_2 > 0$.

(a) If $q_1(t) \geq q_1^e$, $q_2(t) \geq q_2^e$ and $(q_1(t), q_2(t)) \neq (q_1^e, q_2^e)$,

$$\begin{aligned}\dot{q}_1(t) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\ \dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).\end{aligned}$$

$$
\begin{aligned}
\nabla_q V(q)^T f(q) &= (\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&= \lambda_1 - \mu_1 s + \lambda_2 - \theta_1 q_1(t) - \theta_2 q_2(t) \\
&< \lambda_1 - \mu_1 s + \lambda_2 - \theta_1 q_1^e - \theta_2 q_2^e \\
&= 0,
\end{aligned}
$$

where the inequality follows from the facts that $q_1(t) \geq q_1^e$, $q_2(t) \geq q_2^e$ and $(q_1(t), q_2(t)) \neq (q_1^e, q_2^e)$.

(b) If $q_1(t) \geq q_1^e$ and $q_2(t) < q_2^e$,

$$
\begin{aligned}
\dot{q}_1(t) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\
\dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).
\end{aligned}
$$

$$
\begin{aligned}
\nabla_q V(q)^T f(q) &= (\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) - (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&= \lambda_1 - \mu_1 s - \lambda_2 - (\theta_1 + 2\gamma_1) q_1(t) + (\theta_2 + 2\gamma_2) q_2(t) \\
&< \lambda_1 - \mu_1 s - \lambda_2 - (\theta_1 + 2\gamma_1) q_1^e + (\theta_2 + 2\gamma_2) q_2^e \\
&= 0,
\end{aligned}
$$

where the inequality follows from the facts that $q_1(t) \geq q_1^e$ and $q_2(t) < q_2^e$.

(c) If $q_1(t) < q_1^e$ and $q_2(t) \geq q_2^e$,

$$
\begin{aligned}
\dot{q}_1(t) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\
\dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).
\end{aligned}
$$

$$
\begin{aligned}
\nabla_q V(q)^T f(q) &= -(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&< -\lambda_1 + \mu_1 s + \lambda_2 + (\theta_1 + 2\gamma_1) q_1^e - (\theta_2 + 2\gamma_2) q_2^e \\
&= 0,
\end{aligned}
$$

where the inequality follows from the facts that $q_1(t) < q_1^e$ and $q_2(t) \geq q_2^e$.

(d) If $q_1(t) < q_1^e$ and $q_2(t) < q_2^e$,

$$
\begin{aligned}
\dot{q}_1(t) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\
\dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).
\end{aligned}
$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= -(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) - (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\ &= -\lambda_1 + \mu_1 s - \lambda_2 + \theta_1 q_1(t) + \theta_2 q_2(t) \\ &< -\lambda_1 + \mu_1 s - \lambda_2 + \theta_1 q_1^e + \theta_2 q_2^e \\ &= 0,\end{aligned}$$

where the inequality follows from the facts that $q_1(t) < q_1^e$ and $q_2(t) < q_2^e$. □

A.1.2 Proof of Theorem 1

PROOF: Consider the equivalent LP formulation (1.5) for the long-run average cost minimization problem (F1). For any given set of parameters, we first solve the LP (1.5) to obtain an optimal solution (z_1^{e*}, z_2^{e*}) which represents the optimal long-run average service capacity allocated to Class 1 and Class 2. We then show that (z_1^{e*}, z_2^{e*}), and the corresponding (q_1^{e*}, q_2^{e*}), is the globally asymptotically stable equilibrium under the modified $c\mu/\theta$-rule, which corresponds to P_1 or P_2 depending on which class has a higher modified $c\mu/\theta$-index. This step is based on the stability analysis for P_1 (or P_2 by symmetry) in Appendix A.1.1. Following similar parameter regimes examined in the stability analysis, we divide the analysis here into different cases. For each case, the tables below list the optimal LP solution (z_1^{e*}, z_2^{e*}), the corresponding (q_1^{e*}, q_2^{e*}), and the static control under which (q_1^{e*}, q_2^{e*}) is a globally asymptotically stable equilibrium.

Case I. $\frac{\mu_1}{\mu_2} < \frac{\gamma_2}{\theta_2+\gamma_2}$. In this case, the modified $c\mu/\theta$-rule prioritizes Class 2.

	(z_1^{e*}, z_2^{e*})	(q_1^{e*}, q_2^{e*})	Control
$\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} < s$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_1, P_2
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_2
$\frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2} \leq s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$	$\left(s - \frac{\lambda_2}{\mu_2} - \frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},\right.$ $\left.\frac{\lambda_2}{\mu_2} + \frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}\right)$	$\left(\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}, 0\right)$	P_2
$s < \frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}$	$(0,s)$	$\left(\frac{\lambda_1\theta_2+\gamma_2(\lambda_1+\lambda_2-s\mu_2)}{(\theta_1+\gamma_1)\theta_2+\gamma_2\theta_1}, \frac{\lambda_2+\frac{\gamma_1}{\theta_1+\gamma_1}\lambda_1-s\mu_2}{\theta_2+\gamma_2\frac{\theta_1}{\theta_1+\gamma_1}}\right)$	P_2

Case II (a). $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$, and $\frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2} \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$.

	Modified $c\mu/\theta$-rule prioritizes Class 1		
	(z_1^{e*}, z_2^{e*})	(q_1^{e*}, q_2^{e*})	Control
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_1, P_2
$\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} \leq s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$	$\left(\frac{\lambda_1}{\mu_1} + \frac{\gamma_2\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2}, s - \frac{\lambda_1}{\mu_1} - \frac{\gamma_2\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2}\right)$	$\left(0, \frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2}\right)$	P_1
$\frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2} \leq s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$	$(s,0)$	$\left(\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}}, \frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$	P_1
$s < \frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}$	$(s,0)$	$\left(\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}}, \frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$	P_1

Case II (b). $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$, and $\frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2} \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$.

	Modified $c\mu/\theta$-rule prioritizes Class 2		
	(z_1^{e*}, z_2^{e*})	(q_1^{e*}, q_2^{e*})	Control
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_1, P_2
$\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} \leq s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$	$\Bigg(s - \frac{\lambda_2}{\mu_2} - \frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},$ $\frac{\lambda_2}{\mu_2} + \frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}\Bigg)$	$\left(\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}, 0\right)$	P_2
$\frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2} \leq s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$	$\Bigg(s - \frac{\lambda_2}{\mu_2} - \frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},$ $\frac{\lambda_2}{\mu_2} + \frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}\Bigg)$	$\left(\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}, 0\right)$	P_2
$s < \frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}$	$(0, s)$	$\left(\frac{\lambda_1\theta_2+\gamma_2(\lambda_1+\lambda_2-s\mu_2)}{(\theta_1+\gamma_1)\theta_2+\gamma_2\theta_1}, \frac{\lambda_2+\frac{\gamma_1}{\theta_1+\gamma_1}\lambda_1-s\mu_2}{\theta_2+\gamma_2\frac{\theta_1}{\theta_1+\gamma_1}}\right)$	P_2

Case III. $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$, and $\frac{\lambda_2}{\mu_2} + \frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2} \geq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$.

This case follows from Case 2 by symmetry.

Case IV. $\frac{\mu_1}{\mu_2} > \frac{\theta_1+\gamma_1}{\gamma_1}$. In this case, the modified $c\mu/\theta$-rule prioritizes Class 1.

This case follows from Case 1 by symmetry. □

A.2 Proofs of the Results in Section 1.4

The proofs in this section are organized as follows. We start by showing that it is without loss of optimality to restrict our analysis to solutions without chattering behavior (Lemma 1). We then establish the optimal scheduling policy when we are close to the equilibrium (Proposition 1). Both proofs are based on solving the state trajectory q directly. Second, we use Pontryagin's Minimum Principle and Proposition 1 to prove Propositions 2 and 3. In particular, we provide more details about Pontryagin's Minimum Principle. We next prove the auxiliary lemmas (Lemmas 2–4), which are then used to prove Propositions 2 and 3. Note that we actually prove Proposition 3 first, because the proof of Proposition 2 utilizes the results established in Proposition 3. Putting the results of Propositions 1–3 together, we complete the proof of Theorem 2. Lastly, we characterize the policy curve in the special case where $\gamma_1 = 0$, $c_1\mu_1 < c_2\mu_2$, and $r_1 > r_2$ (Proposition 4).

A.2.1 Proof of Lemma 1

PROOF: We prove the lemma by first showing that the cost difference between a chattering trajectory and a properly constructed trajectory without chattering is negligible. This implies that any admissible control policy π that yields a chattering interval can be replaced by a cost-wise equivalent control $\tilde{\pi}$ that does not yield chattering state trajectories. Thus, it is without loss of optimality to consider state trajectories without chattering behavior for the transient optimal control problem (F2′).

Consider an interval $I_1 := [0, \varepsilon]$ where queue 1 is initiated at zero and receives no service capacity for an $\varepsilon > 0$ amount of time. During this interval, a queue accumulates in queue 1. Following I_1, $I_2 = (\varepsilon, \varepsilon + \varepsilon']$ is an interval of length $\varepsilon' > 0$, over which queue 1 receives full service capacity s and is eventually emptied at the end of I_2. Suppose q_2 is initiated at level $q_2(0) = q_{20}$, $q_{20} \in \mathbb{R}_+$. We compute the state trajectory and cost over $I_1 \cup I_2$, i.e. $[0, \varepsilon + \varepsilon']$.

Over the first interval I_1, the state trajectories evolve as

$$\begin{aligned} q_1(t) &= (q_{20}\gamma_2 + \lambda_1)t + o(\varepsilon), \quad t \in [0, \varepsilon] \\ q_2(t) &= q_{20} + (q_{20}(-\gamma_2 - \theta_2) + \lambda_2 - s\mu_2)t + o(\varepsilon), \quad t \in [0, \varepsilon]. \end{aligned}$$

Note that it is possible to ignore the boundary condition that $q_2(t) \geq 0$ for sufficiently small ε. At time ε, the end of time interval I_1, the length of q_1 and q_2 are

$$\begin{aligned} q_1(\varepsilon) &= (q_{20}\gamma_2 + \lambda_1)\varepsilon + o(\varepsilon) \\ q_2(\varepsilon) &= q_{20} + (q_{20}(-\gamma_2 - \theta_2) + \lambda_2 - s\mu_2)\varepsilon + o(\varepsilon). \end{aligned}$$

Using $(q_1(\varepsilon), q_2(\varepsilon))$ as the initial condition at the beginning of the interval I_2, we can characterize the trajectory of q_1 and q_2 over I_2 as

$$\begin{aligned} q_1(t) &= \varepsilon(q_{20}\gamma_2 + \lambda_1) + (t - \varepsilon)[\lambda_1 + \varepsilon(-\gamma_1 - \theta_1)(q_{20}\gamma_2 + \lambda_1) - s\mu_1 \\ &\qquad + \gamma_2(q_{20} - q_{20}\gamma_2\varepsilon - q_{20}\varepsilon\theta_2 + \varepsilon\lambda_2 - s\varepsilon\mu_2)] + o(\varepsilon), \quad t \in [\varepsilon, \varepsilon + \varepsilon'] \\ q_2(t) &= q_{20} - q_{20}\gamma_2\varepsilon - q_{20}\varepsilon\theta_2 + \varepsilon\lambda_2 - s\varepsilon\mu_2 + (t - \varepsilon)[\gamma_1\varepsilon(q_{20}\gamma_2 + \lambda_1) + \lambda_2 \\ &\qquad + (-\gamma_2 - \theta_2)(q_{20} - q_{20}\gamma_2\varepsilon - q_{20}\varepsilon\theta_2 + \varepsilon\lambda_2 - s\varepsilon\mu_2)] + o(\varepsilon), \quad t \in [\varepsilon, \varepsilon + \varepsilon']. \end{aligned}$$

In addition, the value of ε', the time it takes to empty queue 1 from initial level $q_1(\varepsilon)$, is

$$\varepsilon' = \frac{\varepsilon(q_{20}\gamma_2 + \lambda_1)}{-q_{20}\gamma_2 - \lambda_1 + s\mu_1} + o(\varepsilon).$$

The total holding cost over the two intervals I_1 and I_2 is given by

$$C = c_1 \int_0^{\varepsilon + \varepsilon'} q_1(t)dt + c_2 \int_0^{\varepsilon + \varepsilon'} q_2(t)dt.$$

In contrast, we now consider an interval with the same length, $\varepsilon + \varepsilon'$, and the same initial condition $(\tilde{q}_1(0), \tilde{q}_2(0)) = (0, q_{20})$. Now, instead of having q_1 increase from zero and then decrease to zero, we assign strict priority to Class 1 and maintain $\tilde{q}_1$ at zero. The rest of the service capacity is allocated to serve Class 2. Similarly, we characterize the corresponding state trajectory over this interval of length $\varepsilon+\varepsilon'$ as

$$
\begin{aligned}
&\tilde{q}_1(t)=0, \quad t\in\left[0,\varepsilon+\varepsilon'\right]\\
&\tilde{q}_2(t)=q_{20}+t\left[q_{20}(-\gamma_2\mu_1-\theta_2\mu_1+\gamma_2\mu_2)+\lambda_2\mu_1+\lambda_1\mu_2-s\mu_1\mu_2\right]/\mu_1+o(\varepsilon), \quad t\in\left[0,\varepsilon+\varepsilon'\right],
\end{aligned}
$$

and the total holding cost as

$$\tilde{C}=c_1\int_0^{\varepsilon+\varepsilon'}\tilde{q}_1(t)dt+c_2\int_0^{\varepsilon+\varepsilon'}\tilde{q}_2(t)dt.$$

Comparing C and $\tilde{C}$, we get

$$
\begin{aligned}
&C-\tilde{C}\\
&=\frac{\varepsilon^2}{2(q_{20}\gamma_2+\lambda_1-s\mu_1)^2}(q_{20}\gamma_2+\lambda_1)\left(c_2\varepsilon(q_{20}\gamma_2+\lambda_1)\left(q_{20}(\gamma_1\gamma_2+(\gamma_2+\theta_2)^2)+\gamma_1\lambda_1-(\gamma_2+\theta_2)\lambda_2\right)\right.\\
&\quad+c_2s\left((1+\varepsilon(\gamma_2+\theta_2))(q_{20}\gamma_2+\lambda_1)-s\mu_1\right)\mu_2-c_1\left(q_{20}^2\gamma_2^2\varepsilon(\gamma_1+\gamma_2+\theta_1+\theta_2)+\gamma_1\varepsilon\lambda_1^2+\varepsilon\theta_1\lambda_1^2\right.\\
&\quad\left.\left.-\gamma_2\varepsilon\lambda_1\lambda_2+s\lambda_1\mu_1-s^2\mu_1^2+s\gamma_2\varepsilon\lambda_1\mu_2+q_{20}\gamma_2\left(2\gamma_1\varepsilon\lambda_1+\varepsilon(2\theta_1+\theta_2)\lambda_1+s\mu_1+\gamma_2\varepsilon(\lambda_1-\lambda_2+s\mu_2)\right)\right)\right)\\
&=o(\varepsilon).
\end{aligned}
\tag{A.5}
$$

In addition, at the end of time $\varepsilon+\varepsilon'$, we have $q_1(\varepsilon+\varepsilon')=\tilde{q}_1(\varepsilon+\varepsilon')=0$, and

$$
\begin{aligned}
&q_2(\varepsilon+\varepsilon')-\tilde{q}_2(\varepsilon+\varepsilon')\\
&=-\frac{\varepsilon^2(q_{20}\gamma_2+\lambda_1)\left(q_{20}(\gamma_1\gamma_2+(\gamma_2+\theta_2)^2)+\gamma_1\lambda_1-(\gamma_2+\theta_2)(\lambda_2-s\mu_2)\right)}{q_{20}\gamma_2+\lambda_1-s\mu_1}\\
&=o(\varepsilon).
\end{aligned}
\tag{A.6}
$$

Importantly, (A.5) implies that the cost under the policy that has q_1 first increase and then decrease and the cost under strict priority rule to Class 1 which maintains q_1 at zero differ by $o(\varepsilon)$. From (A.6), the queue lengths at time $\varepsilon+\varepsilon'$ under the two policies also differ by $o(\varepsilon)$. Now for any interval of length L, suppose we divide it into $O(L/\varepsilon)$ small triangles (trajectories where q_1 first increases for ε units of time and then decreases to zero). Each has a cost difference $o(\varepsilon)$ from the cost under strict priority to Class 1. Then the overall cost difference between the two policies (chattering versus non-chattering) is $o(\varepsilon)O(L/\varepsilon)$, which goes to zero for fixed L as ε goes to zero. Note that any chattering interval consists of infinitely many such triangular trajectories with infinitesimally small intervals over which q_1 first increases above and then decreases to zero. This implies that any admissible control policy π that yields a chattering interval where q_1 fluctuates infinitesimally around zero can be replaced by a cost-wise equivalent control $\tilde{\pi}$ that maintains q_1 at zero over the same interval and agrees with π elsewhere. The same approach applies to any chattering interval of q_2 around zero – i.e., we can show that there exists a cost-wise equivalent control under which q_2 does not chatter (stays at zero). □

A.2.2 Proof of Proposition 1

PROOF: Let $(q_1(0), q_2(0)) = (\varepsilon, \varepsilon)$. Since the optimal control gives strict priority to one class at any given time, for $\varepsilon > 0$ sufficiently small, it is sufficient to compare the two strict priority rules; see Larrañaga (2015) for a similar observation. Under each priority rule, we characterize the fluid trajectory and calculate the cost. By comparing the costs under the strict priority rules, we note that when the system is initiated close enough to the origin, the optimal policy is to follow the $c\mu$-rule.

We first consider **strict priority to Class 1**. The time horizon is divided into two intervals with length t_1 and t_2 respectively. Class 1 first receives full service capacity and gets emptied at the end of the first interval. Over the second interval, Class 1 is maintained at zero queue and Class 2 is eventually emptied. The fluid trajectory over the first interval is characterized by

$$q_1(t) = \varepsilon + (-\gamma_1\varepsilon + \gamma_2\varepsilon - \varepsilon\theta_1 + \lambda_1 - s\mu_1)t + o(\varepsilon), \quad t \in [0, t_1]$$

$$q_2(t) = \varepsilon + (\gamma_1\varepsilon - \gamma_2\varepsilon - \varepsilon\theta_2 + \lambda_2)t + o(\varepsilon), \quad t \in [0, t_1],$$

and the value of t_1 is

$$t_1 = \frac{\varepsilon}{s\mu_1 - \lambda_1} + o(\varepsilon).$$

Taking the value of $(q_1(t_1), q_2(t_1))$ as the initial condition for the second interval, the fluid trajectory over the second interval is

$$
\begin{aligned}
q_1(t) &= 0, \quad t \in [t_1, t_1 + t_2] \\
q_2(t) &= \frac{1}{\mu_1(-\lambda_1 + s\mu_1)} \left[-\varepsilon\theta_2\mu_1 \left(-\lambda_1 + \lambda_2 + s\mu_1\right) + \gamma_2\varepsilon \left(\lambda_1 - \lambda_2 - s\mu_1\right) \left(\mu_1 - \mu_2\right) \right. \\
&\quad \left. - \left(\lambda_1 - s\mu_1\right) \left(\lambda_2\mu_1 + \lambda_1\mu_2 - s\mu_1\mu_2\right) \right] (t - t_1) + \varepsilon + \frac{\varepsilon\lambda_2}{-\lambda_1 + s\mu_1} + o(\varepsilon), \quad t \in [t_1, t_1 + t_2],
\end{aligned}
$$

and the value of t_2 is

$$
t_2 = \frac{\mu_1(-\lambda_1 + \lambda_2 + s\mu_1)\varepsilon}{(\lambda_1 - s\mu_1)(\lambda_2\mu_1 + \lambda_1\mu_2 - s\mu_1\mu_2)} + o(\varepsilon).
$$

The cumulative holding cost under P_1 over $[0, t_1 + t_2]$ is given by

$$
\begin{aligned}
C^{P_1} &= c_1 \int_0^{t_1} \left[\varepsilon + (-\gamma_1\varepsilon + \gamma_2\varepsilon - \varepsilon\theta_1 + \lambda_1 - s\mu_1)t\right] dt + c_2 \int_0^{t_1} \left[\varepsilon + (\gamma_1\varepsilon - \gamma_2\varepsilon - \varepsilon\theta_2 + \lambda_2)t\right] dt \\
&\quad + c_2 \int_{t_1}^{t_1+t_2} \left\{ \frac{1}{\mu_1(-\lambda_1 + s\mu_1)} \left[-\varepsilon\theta_2\mu_1 \left(-\lambda_1 + \lambda_2 + s\mu_1\right) + \gamma_2\varepsilon \left(\lambda_1 - \lambda_2 - s\mu_1\right) \left(\mu_1 - \mu_2\right) \right. \right. \\
&\quad \left. \left. - \left(\lambda_1 - s\mu_1\right) \left(\lambda_2\mu_1 + \lambda_1\mu_2 - s\mu_1\mu_2\right) \right] (t - t_1) + \varepsilon + \frac{\varepsilon\lambda_2}{-\lambda_1 + s\mu_1} \right\} dt + o(\varepsilon^2) \\
&= \frac{\varepsilon^2}{2(\lambda_1 - s\mu_1)} \left(-c_1 + \frac{c_2(\lambda_2\mu_2 - \lambda_1(\mu_1 + 2\mu_2) + s\mu_1(\mu_1 + 2\mu_2))}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2} \right) + o(\varepsilon^2).
\end{aligned}
$$

Next, we consider the **strict priority rule to Class 2**. Let C^{P_2} denote the total cost of clearing the fluid queue from initial backlog level $(q_1(0), q_2(0)) = (\varepsilon, \varepsilon)$. It follows by symmetry that

$$
C^{P_2} = \frac{\varepsilon^2}{2(\lambda_2 - s\mu_2)} \left(-c_2 + \frac{c_1(\lambda_1\mu_1 - \lambda_2(\mu_2 + 2\mu_1) + s\mu_2(\mu_2 + 2\mu_1))}{\lambda_1\mu_2 + (\lambda_2 - s\mu_2)\mu_1} \right) + o(\varepsilon^2).
$$

Comparing the total costs under P_1 and P_2, we get

$$
\begin{aligned}
C^{P_1} - C^{P_2} &= \frac{\varepsilon^2(c_1\mu_1 - c_2\mu_2)\left(\lambda_1^2 - \lambda_1(2\lambda_2 + s(\mu_1 - 2\mu_2)) + (\lambda_2 + 2s\mu_1)(\lambda_2 - s\mu_2)\right)}{2(\lambda_1 - s\mu_1)(-\lambda_2 + s\mu_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)} + o(\varepsilon^2) \\
&= \frac{\varepsilon^2(c_1\mu_1 - c_2\mu_2)\left(\lambda_1(\lambda_1 - s\mu_1) + (\lambda_2 - s\mu_2)(\lambda_2 + 2s\mu_1 - 2\lambda_1)\right)}{2(s\mu_1 - \lambda_1)(s\mu_2 - \lambda_2)(s\mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1)} + o(\varepsilon^2),
\end{aligned}
\tag{A.7}
$$

Note that as $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$, in (A.7), the denominator $2(s\mu_1 - \lambda_1)(s\mu_2 - \lambda_2)(s\mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1) > 0$, and in the numerator, $(\lambda_1(\lambda_1 - s\mu_1) + (\lambda_2 - s\mu_2)(\lambda_2 + 2s\mu_1 - 2\lambda_1)) < 0$. Thus, for ε sufficiently small, $C^{P_1} - C^{P_2} < 0$ if and only if $c_1\mu_1 > c_2\mu_2$, and vice versa. This indicates that if the system is initiated sufficiently close to the origin, then the $c\mu$-rule is optimal. □

A.2.3 Pontryagin's Minimum Principle

In this section, we provide more details about Pontryagin's Minimum Principle, which will be used in the proof of Lemma 2–4 and Propositions 2–3. Consider the transient optimization problem (F2′) (also presented below).

$$\begin{aligned} \min_{z} \quad & \int_0^\tau F\left(q(t)\right) dt \\ s.t. \quad & \dot{q}(t) = f\left(q(t), z(t)\right) \\ & g(q(t)) \leq 0 \\ & h(z(t)) \leq 0. \end{aligned} \tag{F2$'$ revisited}$$

The pure state constraint $g(q(t)) \leq 0$ is, in general, very hard to deal with as it does not explicitly involve the control $z(t)$ and can only be regulated indirectly via the ordinary differential equation $\dot{q}(t)$. To quantify how "implicitly" $g(q(t))$ depends on $z(t)$, define g_i^j, $j = 1, 2, ..., \ell$, $i = 1, 2$, recursively as

$$\begin{aligned} g_i^0(q(t), z(t)) &:= g_i(q(t)) \\ g_i^1(q(t), z(t)) &:= \nabla_q g_i^0(q(t), z(t))^T f\left(q(t), z(t)\right) \\ &\vdots \\ g_i^\ell(q(t), z(t)) &:= \nabla_q g_i^{\ell-1}(q(t), z(t))^T f\left(q(t), z(t)\right). \end{aligned}$$

If $\nabla_z g_i^j(q(t), z(t)) = 0$ for $0 \leq j \leq \ell - 1$, and $\nabla_z g_i^\ell(q(t), z(t)) \neq 0$, then the state constraint $g_i(q(t))$ is said to be of *order* ℓ. It is easy to see that for (F2′), each pure state constraint is of order 1.

We next introduce a full rank assumption, often referred to as *constraint qualification*, on $g(q(t))$ and $h(z(t))$. In particular, for $g(q(t))$ of order 1, the constraint qualification requires that the matrices

$$\begin{bmatrix} \frac{\partial g^1(q(t))}{\partial z} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \frac{\partial h(z(t))}{\partial z} & \operatorname{diag}\left(h(z(t))\right) \end{bmatrix}$$

have full rank for all $t \geq 0$. In the context of (F2′), we have

$$\operatorname{rank}\left[\frac{\partial g^1(q(t))}{\partial z}\right] = \operatorname{rank}\begin{bmatrix}\mu_1 & 0\\ 0 & \mu_2\end{bmatrix} = 2,$$

and

$$\operatorname{rank}\left[\frac{\partial h(z(t))}{\partial z} \quad \operatorname{diag}\left(h(z(t))\right)\right] = \operatorname{rank}\begin{bmatrix}1 & 1 & z_1(t)+z_2(t)-s & 0 & 0\\ -1 & 0 & 0 & -z_1(t) & 0\\ 0 & -1 & 0 & 0 & -z_2(t)\end{bmatrix} = 3,$$

as at least one of $z_1(t)$ and $z_2(t)$ is strictly positive at all times. Hence, (F2′) satisfies the constraint qualification.

Under the constraint qualification, Pontryagin's Minimum Principle contains a list of necessary conditions satisfied by any optimal solution to the control problem. The next theorem summarizes some of the necessary conditions we utilize in our development. We refer to the survey paper Hartl et al. (1995) for a comprehensive summary of developments regarding Pontryagin's Minimum Principle for optimal control problems with state constraints.

Theorem 8 (Pontryagin's Minimum Principle (Hartl et al. (1995); Sethi and Thompson (2000)))**.** *Assume that the constraint qualification holds. Let z^* be an optimal solution to* (F2′)*, q^* be the corresponding state trajectory, and τ^* be the optimal hitting time. Then, there exists a non-zero piecewise absolutely continuous adjoint vector $p^* : [0,\tau^*] \to \mathbb{R}^2$ with piecewise continuous derivatives, piecewise absolutely continuous Lagrangian multipliers $\eta^* : [0,\tau^*] \to \mathbb{R}^2$, $\xi^* : [0,\tau^*] \to \mathbb{R}^3$, and a vector $\omega^*(\beta_j) \in \mathbb{R}^2$ for each point β_j of discontinuity of p^* such that for almost every $t \in [0,\tau^*]$,*

1. *Ordinary Differential Equation condition:*

$$q^*(0) = q_0, \quad \dot{q}^*(t) = f\left(q^*(t), z^*(t)\right) \tag{ODE}$$

2. *Adjoint Vector condition:*

$$\dot{p}^*(t) = -\nabla_q L(q^*(t), z^*(t), p^*(t), \eta^*(t), \xi^*(t)) \tag{ADJ}$$

3. *Minimization condition:*

$$H(q^*(t), z^*(t), p^*(t)) = \min_z \{H(q^*(t), z(t), p^*(t))\} \tag{M}$$

4. *Hamiltonian condition:*

$$H(q^*(t), z^*(t), p^*(t)) = 0 \tag{H}$$

5. *Transversality condition:*

$$\nabla_z L(q^*(t), z^*(t), p^*(t), \eta^*(t), \xi^*(t)) = 0 \tag{T}$$

6. *Complementary condition:*

$$\begin{aligned} &\eta^*(t) \geq 0, \quad \eta^*(t)^T g(q^*(t)) = 0 \\ &\xi^*(t) \geq 0, \quad \xi^*(t)^T h(z^*(t)) = 0 \end{aligned} \tag{C}$$

7. *Jump condition: For any time in a boundary arc or a junction time, β, the adjoint vector p^*, and the Hamiltonian H may have a discontinuity, but they must satisfy the following jump conditions:*

$$\begin{aligned} &(J1): p^*(\beta-) = p^*(\beta+) + \omega_1^*(\beta)\nabla_q g_1(q^*(\beta)) + \omega_2^*(\beta)\nabla_q g_2(q^*(\beta)) \\ &(J2): H(q^*(\beta-), z(\beta-), p^*(\beta-)) \\ &\qquad\qquad = H(q^*(\beta+), z(\beta+), p^*(\beta+)) - \omega_1^*(\beta)\nabla_t g_1(q^*(\beta)) - \omega_2^*(\beta)\nabla_t g_2(q^*(\beta)) \\ &(J3): \omega^*(\beta) \geq 0, \quad \omega^*(\beta)^T g(q^*(\beta)) = 0. \end{aligned} \tag{J}$$

Next, we provide more explanations about the conditions in Pontryagin's Minimum Principle listed in Theorem 8 to complement the discussion in Section 1.4.3.

1. First, let

$$\begin{aligned} \zeta &:= \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 - \theta_2) + (\gamma_2 - \theta_1 + \theta_2)^2} \\ &= \sqrt{\gamma_2^2 + 2\gamma_2(\gamma_1 + \theta_2 - \theta_1) + (\gamma_1 - \theta_2 + \theta_1)^2}. \end{aligned} \tag{A.8}$$

Note that ζ is well-defined, because

$$\begin{aligned}\gamma_1^2+2\gamma_1(\gamma_2+\theta_1-\theta_2)+(\gamma_2-\theta_1+\theta_2)^2 &= \gamma_1^2+2\gamma_1(\gamma_2+\theta_1-\theta_2)+(-\gamma_2+\theta_1-\theta_2)^2\\ &\geq \gamma_1^2+2\gamma_1(-\gamma_2+\theta_1-\theta_2)+(-\gamma_2+\theta_1-\theta_2)^2\\ &= (\gamma_1-\gamma_2+\theta_1-\theta_2)^2 \geq 0.\end{aligned}$$

Solving the ordinary differential equations in (ADJ) for the dynamic of the adjoint vectors, we get

$$\begin{aligned}p_1^*(t) = \frac{1}{\zeta}e^{\frac{1}{2}t(\gamma_1+\gamma_2+\theta_1+\theta_2)}\Bigg\{&\zeta K_1(0)\cosh\left(\frac{t\zeta}{2}\right)+\sinh\left(\frac{t\zeta}{2}\right)\Bigg[(\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(0)\\ &-2\gamma_1 K_2(0)-2\gamma_1\int_0^t\frac{1}{2\zeta}e^{-\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)u}\Big(2c_1\gamma_2+c_2(\gamma_1-\gamma_2+\theta_1-\theta_2-\zeta)\\ &-e^{\zeta u}(2c_1\gamma_2+c_2(\gamma_1-\gamma_2+\theta_1-\theta_2+\zeta))+2(-1+e^{\zeta u})\gamma_2\eta_1^*(u)\\ &+(-\gamma_1+\gamma_2-\theta_1+\theta_2+\zeta+e^{\zeta u}(\gamma_1-\gamma_2+\theta_1-\theta_2+\zeta))\eta_2^*(u)\Big)du\Bigg]\\ &+\left(\zeta\cosh\left(\frac{t\zeta}{2}\right)+(\gamma_1-\gamma_2+\theta_1-\theta_2)\sinh\left(\frac{t\zeta}{2}\right)\right)\int_0^t\frac{1}{2\zeta}e^{-\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)u}\\ &\Big(-c_1(\gamma_1-\gamma_2+\theta_1-\theta_2+\zeta)+2c_2\gamma_1+(\gamma_1-\gamma_2+\theta_1-\theta_2+\zeta\\ &+e^{\zeta u}(-\gamma_1+\gamma_2-\theta_1+\theta_2+\zeta))\eta_1^*(u)-2\gamma_1\eta_2^*(u)\\ &-e^{\zeta u}(2c_2\gamma_1+c_1(-\gamma_1+\gamma_2-\theta_1+\theta_2+\zeta)-2\gamma_1\eta_2^*(u))\Big)du\Bigg\},\end{aligned}$$

where $K_1(0), K_2(0)$ are constants that depend on $p_1^*(0)$ and $p_2^*(0)$. The expression for $p_2^*(t)$ follows by symmetry.

The adjoint vector is connected to the value function under the optimal control. In particular, the value function $\Xi : \mathbb{R}_+^2 \to \mathbb{R}_+$ associated with (F2′) is defined by

$$\Xi(a_1,a_2) = \inf\left\{\int_0^\tau F(q(t))dt \,\middle|\, q_1(0)=a_1,\, q_2(0)=a_2,\, q \text{ is a feasible trajectory in (F2}')\right\}.$$

There exists an adjoint vector $p^*(t)$ such that $p^*(t) = \nabla_q\Xi(q^*(t))$ under the condition that $\nabla_q\Xi(q)$ is well defined (Frankowska, 2010). As the cost structure is linear and increasing in $q^*(t)$, it follows that $p^*(t) \geq 0$ for all $t \geq 0$.

2. Minimization condition (M) and the optimal assignment of service capacity in equations (1.9) - (1.10) reveal important properties of the optimal control structure. First, observe in (1.9) - (1.10) that on the interior arc when both states are strictly positive and the switching curve is non-zero, the optimal control is “bang-bang". Namely, it must be the case that one of the two classes is assigned full service capacity s. On the other hand, on the boundary arc when one of the states is at zero, the optimal control is of an “interior" type. Namely, both $z_1^*(t)$ and $z_2^*(t)$ stay strictly positive in the interior of the control region, i.e., $z_1^*(t), z_2^*(t) \in (0,s)$.

3. Consider time β, where $\beta < \tau^*$, as a time on a boundary arc or a junction time. If the adjoint vector p^* has a discontinuity at time β, then Jump condition (J) requires that

$$\begin{bmatrix} p_1^*(\beta-) \\ p_2^*(\beta-) \end{bmatrix} = \begin{bmatrix} p_1^*(\beta+) \\ p_2^*(\beta+) \end{bmatrix} + w_1^*(\beta) \begin{bmatrix} -1 \\ 0 \end{bmatrix} + w_2^*(\beta) \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} p_1^*(\beta+) - w_1^*(\beta) \\ p_2^*(\beta+) - w_2^*(\beta) \end{bmatrix},$$

and that

$$w_i^*(\beta) \geq 0, \quad w_i^*(\beta) g_i(q^*(\beta)) = 0, \quad i = 1,2.$$

Note that if $q_1^*(\beta) = 0$, then $q_2^*(\beta) > 0$ and thus $w_2^*(\beta) = 0$. The same holds true for q_2^*, namely, if $q_2^*(\beta) = 0$, then $q_1^*(\beta) > 0$ and thus $w_1^*(\beta) = 0$. Hence, only the adjoint vector associated with the queue that is at zero can have a jump, while the other adjoint vector remains continuous at time β.

In addition, since the pure state constraint $g(q)$ is time invariant, i.e., the function g does not have a time argument, we have $\nabla_t g(q^*(\beta)) = 0$. According to Jump condition (J), the Hamiltonian $H(q^*(t), z^*(t), p^*(t))$ is continuous over boundary arcs and at junction times.

4. Pontryagin’s Minimum Principle only requires the necessary conditions to be satisfied “almost everywhere". In particular, $q^*(t)$ and $p^*(t)$ can have discontinuities at

countably many points. For most problems studied in the literature, jumps only happen at junction times (Hartl et al., 1995). That said, in general, we cannot rule out the possibility of jumps on the boundary or interior arcs. In our analysis, we shall first assume that $p^*(t)$ is *continuous on interior arcs*. We then show that the continuity assumption indeed holds by verifying a sufficient version of Pontryagin's Minimum Principle for the optimal control problem (F2$'$).

We next introduce the sufficient version of Pontryagin's Minimum Principle. Since the terminal state in problem (F2$'$) is zero and $F(0) = 0$, (F2$'$) can be equivalently formulated as an optimal control problem without a terminal state constraint but rather over an infinite time horizon. The following sufficient conditions are adapted from Theorem 8.2 and Theorem 8.4 in (Hartl et al., 1995) for the equivalent version of (F2$'$) over an infinite time horizon.

Theorem 9 (Arrow-type sufficient condition)**.** *Let (q^*, z^*) be a feasible pair for an equivalent version of problem* (F2$'$) *with infinite time horizon. Assume that there exists a piecewise continuously differentiable function $p^*(t) : [0,\infty) \to \mathbb{R}^2$ and piecewise continuous functions $\eta^* : [0,\infty) \to \mathbb{R}^2$ and $\xi^* : [0,\infty) \to \mathbb{R}^3$, such that conditions (ODE), (ADJ), (M), (H), (T), (C) hold. Assume further that at all points β of discontinuity of p^*, there exists an $\omega^*(\beta) \in \mathbb{R}^2$ such that (J1) and (J3) in (J) hold. In addition, assume that the following limiting condition holds:*

$$\lim_{t\to\infty} p^*(t)^T \left(q(t) - q^*(t)\right) \geq 0 \quad \textit{for every other feasible state trajectory } q.$$

If the minimized Hamiltonian $H(q^*(t), z^*(t), p^*(t)) = \min_z\{H(q^*(t), z(t), p^*(t))\}$ is convex in $q^*(t)$ for all $(p^*(t), t)$, the pure state constraint $g(q(t))$ is quasiconvex in $q(t)$, and the control constraint $h(z(t))$ is quasiconvex in $z(t)$, then (q^*, z^*) is an optimal pair.

We first note that the solution we derive in this chapter indeed satisfies the sufficient conditions in Theorem 9 and is thus optimal. More specifically, first, we design the control by ensuring that conditions (ODE), (ADJ), (M), (T), and (C) are satisfied almost everywhere. In particular, in our proposed solution, the state trajectory $q^*(t)$ satisfies (ODE) at all the continuity points of the control $z^*(t)$. The adjoint vector $p^*(t)$ follows the ordinary differential equations in (ADJ) everywhere on the interior arcs. (M), (T), and (C) hold everywhere over the transient time horizon. Second, Jump condition (J) is guaranteed everywhere over boundary arcs and at junction times. Since $p^*(t)$ is continuous over interior arcs, conditions (J1) and (J3) in (J) indeed hold for all discontinuity points of $p^*(t)$. Third, for any feasible state trajectory $q(t)$ other than $q^*(t)$, $\lim_{t\to\infty} p^*(t)(q(t)-q^*(t)) \geq 0$ holds, because $p^*(t), q(t) \geq 0$ for all $t \geq 0$, and $\lim_{t\to\infty} q^*(t) = 0$. Lastly, following (1.9)–(1.10), the control $z^*(t)$ is linear in $q^*(t)$ for all $t \geq 0$. Hence, the minimized Hamiltonian $H(q^*(t), z^*(t), p^*(t))$ is linear in $q^*(t)$ for all $(p^*(t), t)$. The convexity conditions on $g(q(t))$ and $h(z(t))$ are also satisfied as $g(q(t))$ and $h(z(t))$ are linear in $q(t)$ and $z(t)$ respectively.

We are now ready to prove the results in Section 1.4.3 using Pontryagin's Minimum Principle.

A.2.4 Proof of Lemma 2

PROOF: The proof of Lemma 2 uses Transversality condition (T) and Complementarity condition (C) . Consider a boundary arc $[t_1, t_2]$ and a time epoch $t \in (t_1, t_2)$. First, by (1.9)–(1.10), the control over the boundary arc is of an "interior" type, and the amount of service capacity assigned to both classes $(z_1^*(t), z_2^*(t))$ is strictly positive. By Complementarity condition (C), the multipliers satisfy $\xi_2^*(t) = 0$ and $\xi_3^*(t) = 0$. Then, by Transversality condition (T), we have $\mu_1 p_1^*(t) = \mu_2 p_2^*(t) = \xi_1^*(t)$. Hence, the switching curve satisfies $\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t) = 0$ for $t \in (t_1, t_2)$. □

A.2.5 Proof of Lemma 3

PROOF: Recall that the switching curve is characterized by $\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t)$. Since $\psi(t) = 0$ on the boundary arcs and by our construction, $p^*(t)$ does not jump on the interior arcs, the switching curve $\psi(t)$ is continuous at all time $t \in [0, \tau^*]$ if $p^*(t)$ is continuous at the junction times. In the rest of the proof, we establish the continuity of $p^*(t)$ at the junction times.

Following Proposition 4.2 in Hartl et al. (1995) and Proposition 3.63 in Grass et al. (2008), for the optimal control problem (F2′) which has pure state constraints of order 1, the adjoint vector $p^*(t)$ is continuous at a junction time β, i.e., $\omega^*(\beta) = 0$, if the entry or exit is nontangential, i.e., $\dot{q}_i^*(\beta-) < 0$ or $\dot{q}_i^*(\beta+) > 0$, respectively. Namely, the nontangential condition requires that if β is an entry or contact point for q_i^*, then $\dot{q}_i^*(\beta-) < 0$. If β is an exit or contact point for q_i^*, then $\dot{q}_i^*(\beta+) > 0$. In what follows, we use this nontangential condition and/or Jump condition (J) to establish continuity of $p^*(t)$ at junction times. We prove the statement for the junction times associated with Class 1; the arguments for Class 2 follow by symmetry. The discussion is divided into three cases based on the relative level of service capacity s.

Case I. $\max\{\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}, \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}\} < s$**.**

(i) Let β be an entry or contact point for q_1^*.

In order to drive q_1^* to zero, full service capacity must be assigned to q_1^* right before β, i.e., $z_1^*(\beta-) = s$. Hence,

$$\dot{q}_1^*(\beta-) = \lambda_1 - \mu_1 z_1^*(\beta-) - \theta_1 q_1^*(\beta-) - \gamma_1 q_1^*(\beta-) + \gamma_2 q_2^*(\beta-) = \lambda_1 - \mu_1 s + \gamma_2 q_2^*(\beta-).$$

In addition, there exists some neighborhood $[\beta-\delta, \beta)$, $0 < \delta < \beta$, where $\dot{q}_1^*(t) < 0$ for all $t \in [\beta-\delta, \beta)$. This implies that

$$q_2^*(t) < (s\mu_1 - \lambda_1)/\gamma_2 \quad \text{for all } t \in [\beta-\delta, \beta).$$

We next show that $\dot{q}_1^*(\beta-) < 0$. Suppose by contradiction $\dot{q}_1^*(\beta-) = 0$, then it must be the case that $q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2$. On the other hand,

$$\dot{q}_2^*(t) = \lambda_2 - \mu_2 z_2^*(t) - (\theta_2+\gamma_2)q_2^*(t) + \gamma_1 q_1^*(t) \le \lambda_2 - (\theta_2+\gamma_2)q_2^*(t) + \gamma_1 q_1^*(t),$$

which is strictly negative if

$$q_2^*(t) > \lambda_2/(\theta_2+\gamma_2)+\gamma_1 q_1^*(t)/(\theta_2+\gamma_2).$$

For $s > \max\{\lambda_1/\mu_1+\lambda_2/\mu_2, \lambda_1/\mu_1+\lambda_2\gamma_2/((\gamma_2+\theta_2)\mu_1)\}$, it holds that

$$q_2^*(\beta) = (s\mu_1-\lambda_1)/\gamma_2 > \lambda_2/(\theta_2+\gamma_2).$$

Therefore, there exists some $\delta' > 0$, such that $\dot{q}_2^*(t) < 0$ and $q_2^*(t) > q_2^*(\beta)$ for $t \in (\beta - \delta', \beta)$. It follows that $\dot{q}_1^*(t) > 0$ for $t \in (\beta-\delta', \beta)$, which contradicts that $\dot{q}_1^*(t) < 0$ for all $t \in [\beta-\delta, \beta)$.

Therefore, $\dot{q}_1^*(\beta-) < 0$ at entry or contact point β.

(ii) Let β be an exit or contact point for q_1^*. Similar arguments as in Case I(i) apply, and we can show that $\dot{q}_1^*(\beta+) > 0$.

Since all the entry and exit trajectories are nontangential, the adjoint vectors $p^*(t)$ are continuous at the junction times associated with Class 1 in this case.

Case II. $s = \frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} > \frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$.

(i) Let β be an entry point for $q_1^*(t)$.

First, if $\dot{q}_1^*(\beta-) < 0$, then it follows from the nontangential condition that there is no jump in the adjoint vector $p^*(t)$ at time β.

Second, suppose for the sake of contradiction that $\dot{q}_1^*(\beta-) = 0$. It then follows that

$$q_2^*(\beta) = (s\mu_1-\lambda_1)/\gamma_2 = \lambda_2/(\theta_2+\gamma_2).$$

Note that the point $(0, \lambda_2/(\theta_2+\gamma_2))$ is a locally asymptotically stable equilibrium point for the joint queue length process under priority to Class 1, while $(0,0)$ is the equilibrium under priority to Class 2. Hence, priority must be switched from Class 1 to Class 2 at time β. This implies that β cannot be an entry point for $q_1^*(t)$, a contradiction.

Therefore, $\dot{q}_1^*(\beta-) < 0$ at entry point β for q_1^*.

(ii) Let β be an exit point for q_1^*.

First, if $\dot{q}_1^*(\beta+) > 0$, then it follows from the nontangential condition that there is no jump in the adjoint vector $p^*(t)$ at time β.

Second, suppose for the sake of contradiction that $\dot{q}_1^*(\beta+) = 0$. Then,

$$q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 = \lambda_2/(\theta_2 + \gamma_2).$$

Following the same reasoning as in Case II(i), since the point $(0, \lambda_2/(\theta_2 + \gamma_2))$ is a locally asymptotically stable equilibrium point for the joint queue length process, priority must be switched from Class 1 to Class 2 at time β. This implies that $z_1^*(\beta+) = 0$ and

$$\dot{q}_1^*(\beta+) = \lambda_1 - \mu_1 z_1^*(\beta+) - \gamma_1 q_1^*(\beta+) + \gamma_2 q_2^*(\beta+) > 0,$$

a contradiction.

Therefore, $\dot{q}_1^*(\beta+) > 0$ at exit point β for q_1^*.

(iii) Let β be a contact point for q_1^*.

First, if $\dot{q}_1^*(\beta-) < 0$ and $\dot{q}_1^*(\beta+) > 0$, then $p^*(t)$ does not have any jump at time β due to the nontangential condition.

Second, if $\dot{q}_1^*(\beta-) = 0$, then following the same arguments as in Case II(i) and Case II(ii), it holds that $q_2^*(\beta) = \lambda_2/(\theta_2 + \gamma_2)$ and priority is switched from Class 1 to Class 2 at time β. In this case, Jump condition (J) requires the adjoint vector $p^*(t)$ to have no jump at time β. To see this, suppose for the sake of contradiction that $p^*(t)$ jumps at β. Then, Jump condition (J) characterizes that $p_1^*(\beta+) = p_1^*(\beta-) + w_1^*(\beta)$, for some $w_1^*(\beta) > 0$. Recall that the switching curve is defined as $\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t)$. Since Class 1 is prioritized right before β, it holds that $\psi(\beta-) \geq 0$. If $p_1^*(t)$ has a jump with strictly positive size $w_1^*(\beta)$ at time β, then $\psi(\beta+) > 0$. However, this implies that priority cannot be switched to Class 2 at time β, which is a contradiction.

Third, the case where $\dot{q}_1^*(\beta+) = 0$ can be ruled out by exactly the same arguments in Case II(ii).

In the cases where β is an entry or exit point, we show that the trajectories are nontangential. Hence the adjoint vectors $p^*(t)$ are continuous at these junction times associated with Class 1. In the case where β is a contact point, we have established the continuity of the adjoint vectors $p^*(t)$ at β by either showing that the trajectories are nontangetial or using Jump condition (J) (in the case where priority is switched from Class 1 to Class 2 at β)

Case III. $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s < \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$**.**

(i) Let β be an entry point for q_1^*.

First, if $\dot{q}_1^*(\beta-) < 0$, then $p_1^*(t)$ does not jump at β due to the nontangential condition.

Second, suppose for the sake of contradiction that $\dot{q}_1^*(\beta-) = 0$. Then, $q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 < \lambda_2/(\theta_2+\gamma_2)$. Recall that the dynamic of q_2^* follows $\dot{q}_2^*(t) = \lambda_2 - \mu_2 z_2^*(t) - (\theta_2+\gamma_2)q_2^*(t) + \gamma_1 q_1^*(t)$. Because priority is kept at Class 1 over the boundary arc following β, there exists some $\delta > 0$ such that $\dot{q}_2^*(t) > 0$ for $t \in [\beta, \beta+\delta)$. This implies that $\dot{q}_1^*(t) > 0$ for $t \in (\beta, \beta+\delta)$, contradicting the fact that β is an entry point for $q_1^*(t)$.

Therefore, $\dot{q}_1^*(\beta-) < 0$ at entry point β for q_1^*.

(ii) Let β be an exit point for q_1^*.

First, if $\dot{q}_1^*(\beta+) > 0$, then $p_1^*(t)$ does not jump at β due to the nontangential condition.

Second, suppose for the sake of contradiction that $\dot{q}_1^*(\beta+) = 0$. Then, priority must be kept at Class 1 at time β and over some interval $[\beta, \beta+\delta_1)$, $\delta_1 > 0$; otherwise, $\dot{q}_1^*(\beta+) > 0$. In addition, we have, $q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 < \lambda_2/(\theta_2+\gamma_2)$. It then follows from the dynamic of q_2^* that there further exists some δ_2, $0 < \delta_2 < \delta_1$, such that $z_1^*(t) = s$, $\dot{q}_1^*(t) > 0$ and $\dot{q}_2^*(t) > 0$ for $t \in (\beta, \beta+\delta_2)$. Since $p_1^*(t) \geq 0$, $p_2^*(t) \geq$

0 and $q_2^*(t) > 0$ for $t \in (\beta, \beta + \delta_2)$, we have $H(q^*(t), z^*(t), p^*(t)) = p_1^*(t)\dot{q}_1^*(t) + p_2^*(t)\dot{q}_2^*(t) + c_1 q_1^*(t) + c_2 q_2^*(t) > 0$ for $t \in (\beta, \beta + \delta_2)$. However, the Hamiltonian condition (H) requires that $H(q^*(t), z^*(t), p^*(t)) = 0$ almost everywhere, which gives a contradiction.

Therefore, $\dot{q}_1^*(\beta+) > 0$ at exit point β for q_1^*.

(iii) Let β be a contact point for q_1^*.

First, if $\dot{q}_1^*(\beta-) < 0$ and $\dot{q}_1^*(\beta+) > 0$, then $p^*(t)$ does not jump at β due to the nontangential condition.

Second, for $\dot{q}_1^*(\beta-) = 0$, we first note that if priority is switched from Class 1 to Class 2 at time β, then Jump condition (J) requires that $p^*(t)$ does not jump at β due to the same reasoning as in Case II(iii). Next, suppose for the sake of contradiction that $\dot{q}_1^*(\beta-) = 0$ and priority is kept at Class 1 over some interval $[\beta, \beta + \delta_1)$, $\delta_1 > 0$. Then, following the same arguments as in Case III(ii), there exists some δ_2, $0 < \delta_2 < \delta_1$, such that $z_1^*(t) = s$, $\dot{q}_1^*(t) > 0$ and $\dot{q}_2^*(t) > 0$ for $t \in (\beta, \beta + \delta_2)$, which violates the Hamiltonian condition (H), and thus gives a contradiction.

Third, the case where $\dot{q}_1^*(\beta+) = 0$ is ruled out by the same arguments as in Case III.(ii).

In the cases where β is an entry or exit point, we show that the trajectories are nontangential. Hence the adjoint vectors $p^*(t)$ are continuous at these junction times associated with Class 1. In the case where β is a contact point, we have established the continuity of $p^*(t)$ at β by either showing that the trajectories are nontangetial or using Jump condition (J).

Taking Cases I, II, III together, we have shown that the adjoint vectors $p^*(t)$ are continuous at all the junction times. This further implies that the switching curve $\psi(t)$ is continuous at all $t \in [0, \tau^*]$. □

A.2.6 Proof of Lemma 4

PROOF: By Lemma 1, we restrict to trajectories without chattering behavior. For any entry or contact point τ_j, there exists a nontrivial interval $(0,\alpha_j)$ such that for $t\in(0,\alpha_j)$, $q_1^*(\tau_j-t)$ and $q_2^*(\tau_j-t)$ are both strictly positive. Thus, the multiplier η^* is equal to zero over any interior arc. Recall from (A.8) that

$$\zeta=\sqrt{\gamma_1^2+2\gamma_1(\gamma_2+\theta_1-\theta_2)+(\gamma_2-\theta_1+\theta_2)^2}=\sqrt{\gamma_2^2+2\gamma_2(\gamma_1+\theta_2-\theta_1)+(\gamma_1-\theta_2+\theta_1)^2}.$$

We get from (ADJ) that for $t\in(0,\alpha_j)$,

$$\begin{aligned}
p_1^*(\tau_j-t)=&\frac{c_1}{\theta_1+\gamma_1\frac{\theta_2}{\gamma_2+\theta_2}}+\frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}}+e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2)(\tau_j-t)}\Bigg(K_1(\tau_j)\cosh\left[\frac{1}{2}\zeta(\tau_j-t)\right]\\
&+\frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j)-2\gamma_1K_2(\tau_j)\right)\sinh\left[\frac{1}{2}\zeta(\tau_j-t)\right]\Bigg)\\
=&\frac{c_1}{\theta_1+\gamma_1\frac{\theta_2}{\gamma_2+\theta_2}}+\frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}}+e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2)(\tau_j-t)}\\
&\Bigg[\frac{1}{2}\left(K_1(\tau_j)+\frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j)-2\gamma_1K_2(\tau_j)\right)\right)e^{\frac{1}{2}\zeta(\tau_j-t)}\\
&-\frac{1}{2}\left(K_1(\tau_j)+\frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j)-2\gamma_1K_2(\tau_j)\right)\right)e^{-\frac{1}{2}\zeta(\tau_j-t)}\Bigg]\\
=&\frac{c_1}{\theta_1+\gamma_1\frac{\theta_2}{\gamma_2+\theta_2}}+\frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}}\\
&+\frac{1}{2}\left(K_1(\tau_j)+\frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j)-2\gamma_1K_2(\tau_j)\right)\right)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j-t)}\\
&-\frac{1}{2}\left(K_1(\tau_j)+\frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j)-2\gamma_1K_2(\tau_j)\right)\right)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j-t)}.
\end{aligned}$$

where $K_1(\tau_j),K_2(\tau_j)$ are constants that depend on $p_1^*(\tau_j)$ and $p_2^*(\tau_j)$.

Let

$$A_1(\tau_j):=\frac{1}{2}\left(K_1(\tau_j)+\frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j)-2\gamma_1K_2(\tau_j)\right)\right). \tag{A.9}$$

It is immediate that

$$\begin{aligned}
p_1^*(\tau_j-t)=&\frac{c_1}{\theta_1+\gamma_1\frac{\theta_2}{\gamma_2+\theta_2}}+\frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}}+A_1(\tau_j)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j-t)}\\
&-A_1(\tau_j)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j-t)}.
\end{aligned}\tag{A.10}$$

By symmetry, for

$$A_2(\tau_j) := \frac{1}{2}\left(K_2(\tau_j) + \frac{1}{\zeta}\left((\gamma_2 - \gamma_1 + \theta_2 - \theta_1)K_2(\tau_j) - 2\gamma_2 K_1(\tau_j)\right)\right), \tag{A.11}$$

we have

$$\begin{aligned} p_2^*(\tau_j - t) = {} & \frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2+\gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} + A_2(\tau_j) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j - t)} \\ & - A_2(\tau_j) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j - t)}. \end{aligned} \tag{A.12}$$

The backward switching curve from time τ_j over the interval $(0, \alpha_j)$ is given by

$$\begin{aligned} \psi(\tau_j - t) = {} & \left(\frac{c_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2 \frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}}\right)\mu_1 - \left(\frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2+\gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}}\right)\mu_2 \\ & + \left(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j - t)} \\ & - \left(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j - t)}. \end{aligned}$$

Lastly, we note that $\gamma_1 + \gamma_2 + \theta_1 + \theta_2 - \zeta > 0$. This is because under Assumption 1, at least one of θ_1 and θ_2 is strictly positive, and then,

$$\begin{aligned} \zeta &= \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 - \theta_2) + (\gamma_2 - \theta_1 + \theta_2)^2} \\ &< \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 + \theta_2) + (\gamma_2 + \theta_1 + \theta_2)^2} \\ &= \gamma_1 + \gamma_2 + \theta_1 + \theta_2. \end{aligned}$$

The statement follows from defining

$$v_1 := \frac{1}{2}(\gamma_1 + \gamma_2 + \theta_1 + \theta_2 + \zeta), \quad v_2 := \frac{1}{2}(\gamma_1 + \gamma_2 + \theta_1 + \theta_2 - \zeta).$$

□

A.2.7 Proof of Proposition 3

PROOF: The proof utilizes Proposition 1 and the possible shapes of the switching curve, $\psi(\tau_N - t)$, characterized in Lemma 4. It is divided into two cases, depending on the relationship between the $c\mu$-index and the modified $c\mu/\theta$-index.

Case I. First, we consider the parameter regime where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, namely,

$$(c_1\mu_1 - c_2\mu_2)(r_1 - r_2) > 0, \quad \text{for } r_1, r_2 \text{ in (1.6) and (1.7).}$$

For the moment, suppose Class 1 has a higher $c\mu$-index and modified $c\mu/\theta$-index.

By Proposition 1, when the state is in an ε-neighborhood of the origin, it is optimal to assign strict priority to Class 1. Recall that τ_N is the last entry or contact point (forward in time) when one of the states hits zero. It follows that τ_N must be the last epoch forward in time when q_1^* hits zero, and q_1^* is then maintained at zero after τ_N, i.e., $q_1^*(t) = 0$ for $t \in [\tau_N, \tau^*]$. By Lemma 4, the switching curve right before τ_N satisfies for some $\alpha_N < \tau_N$,

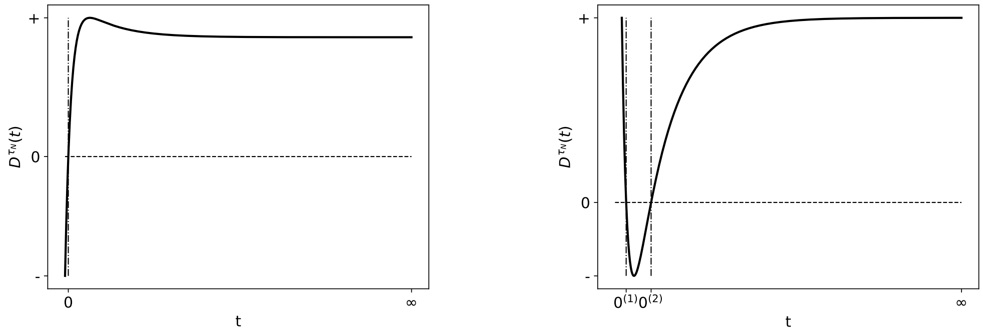
$$\begin{aligned}\psi(\tau_N - t) = r_1 - r_2 &+ (\mu_1 A_1(\tau_N) - \mu_2 A_2(\tau_N))\, e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_N - t)} \\ &- (\mu_1 A_1(\tau_N) - \mu_2 A_2(\tau_N))\, e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_N - t)}, \quad t \in (0, \tau_N - \alpha_N),\end{aligned}$$

where $A_1(\tau_N)$ and $A_2(\tau_N)$ are constants in $\mathbb{R}$. Furthermore, $D^{\tau_N}(t)$, the pseudo switching curve backward from τ_N, satisfies

$$\lim_{t\to\infty} D^{\tau_N}(t) = r_1 - r_2 > 0.$$

The structure of $D^{\tau_N}(t)$ regulates that it can have at most two zeros. With $\lim_{t\to\infty} D^{\tau_N}(t) > 0$, the two possible function shapes $D^{\tau_N}(t)$ can take are demonstrated in Figure A.1, with one root in Figure A.1a and two roots in Figure A.1b. Figure A.1 is comprehensive in the sense that any $D^{\tau_N}(t)$ function shares the same behavior in crossing zeros and in the limit as $t \to \infty$. In particular, if $D^{\tau_N}(t)$ has one zero as in Figure A.1a, then it must be that $D^{\tau_N}(t)$ is increasing at the zero point and eventually converges to $r_1 - r_2$. Once $D^{\tau_N}(t)$ crosses zero, it will never decrease to zero again. Likewise, if $D^{\tau_N}(t)$ has two zeros as in Figure A.1b, then it must be that $D^{\tau_N}(t)$ has negative slope at the first zero, has positive slope at the second zero, and eventually converges to $r_1 - r_2$. Once $D^{\tau_N}(t)$ crosses the second zero point, it will never decrease to zero again. We comment that if the values of $A_1(\tau_N)$ are $A_2(\tau_N)$ are known, then there is no ambiguity in the trajectory of $D^{\tau_N}(t)$, and thus no notion of “possible" function shapes. Nevertheless, due to the degrees of freedom inherent to Pontryagin’s Minimum Principle, it is hard to characterize these coefficients exactly. Therefore, the idea is to infer the structure of the optimal control from the interaction of the coefficients without explicitly characterizing their values.

Figure A.1: Possible trajectory of $D^{\tau_1}(t)$ with $c_1\mu_1 > c_2\mu_2$
(modified $c_1\mu_1/\theta_1 >$ modified $c_2\mu_2/\theta_2$)



(a) One zero point

(b) Two zero points

We first note that the interval $[\tau_N, \tau^*]$ is a boundary arc over which q_1^* is maintained at zero. It follows that $\psi(t) = 0$ for $t \in (\tau_N, \tau^*)$ (Lemma 2), and $\psi(t)$ is continuous in time so that $\psi(\tau_N) = 0$ (Lemma 3). Furthermore, since the optimal control is “bang-bang" right before τ_N, in order to drive q_1^* to zero at time τ_N, strict priority must be given to Class 1 in some non-trivial neighborhood before τ_N. Namely, there exists $\varepsilon_{\tau_N} > 0$ such that $\psi(t) > 0$ for $t \in (\tau_N - \varepsilon_{\tau_N}, \tau_N)$. For $D^{\tau_N}(t)$, this implies that $D^{\tau_N}(0) = 0$ and $D^{\tau_N}(t) > 0$ for $t \in (0, \varepsilon_{\tau_N})$. Thus for the possible structures in Figure A.1, if $D^{\tau_N}(t)$ has one zero (Figure A.1a), then $D^{\tau_N}(0)$ is at this unique zero point. If $D^{\tau_N}(t)$ has two zeros (Figure A.1b), then $D^{\tau_N}(0)$ is at the second zero. This implies that as long as the dynamic of the switching curve $\psi(\tau_N - t)$ follows that of $D^{\tau_N}(t)$, $\psi(\tau_N - t) > 0$. It is important to note that the trajectory of $\psi(\tau_N - t)$ agrees with $D^{\tau_N}(t)$ for t in some non-degenerative interval $(0, \tau_N - \alpha_N)$.

Next, taking the derivative of $D^{\tau_N}(t)$ with respect to t, it is easy to see that $dD^{\tau_N}(t)$ can have at most one root. Since $D^{\tau_N}(0) = 0$ and $D^{\tau_N}(t) > 0$ for $t \in (0, \varepsilon_{\tau_N})$, it holds that for any interval $[0, \ell)$, $\ell > 0$, either $D^{\tau_N}(t)$ is strictly increasing over $[0, \ell)$ or $D^{\tau_N}(\ell) >$

$\lim_{t\to\infty} D^{\tau_N}(t) - \delta$ for some $\delta > 0$ (which can be arbitrarily small). In either case, $D^{\tau_N}(\ell) > \delta'$ for some $\delta' > 0$. If $\eta_1^*(\tau_N - t) = 0$ and $\eta_2^*(\tau_N - t) = 0$ for $t \in [0,\ell)$, then the same holds true for the backward switching curve $\psi(\tau_1 - t)$ over the interval $t \in [0,\ell)$. To this end, it is only possible for $\psi(\tau_N - t)$ to deviate from the dynamic of $D^{\tau_N}(t)$ if $\eta_2^*(\tau_N - \beta)$ becomes strictly positive at some time $0 < \beta \le t$. (Naturally, $\beta \le \alpha_N$.) Now, suppose there exists such $\beta > 0$, i.e., $\eta_2^*(\tau_N - \beta) > 0$. Note that $\eta_1^*(\tau_N - t) = 0$ and $\eta_2^*(\tau_N - t) = 0$ for all $t \in [0,\beta)$. As $D^{\tau_N}(\beta) > \delta'$ for some $\delta' > 0$ and $\eta_2^*(\tau_N - \beta) > 0$, it follows that $\psi(\tau_N - \beta) \ge \delta' > 0$. However, $\eta_2^*(\tau_N - \beta)$ becomes positive only if $q_2^*(\tau_N - \beta) = 0$, which implies that strict priority is given to Class 2 right before time $(\tau_N - \beta)$, i.e., $\psi((\tau_N - \beta)-) \le 0$. However, due to the continuity of the switching curve, this contradicts the fact that $\psi(\tau_N - \beta) \ge \delta' > 0$. Therefore, for all $t \in (0, \tau_N]$, $\psi(\tau_N - t)$ follows the dynamic of $D^{\tau_N}(t)$ and remains strictly positive. We then conclude that strict priority to Class 1 is optimal throughout the transient time horizon.

The proof for the case where Class 2 has a higher $c\mu$-index and higher modified $c\mu/\theta$-index follows similarly. In this case, strict priority to Class 2 is optimal throughout the transient time horizon.

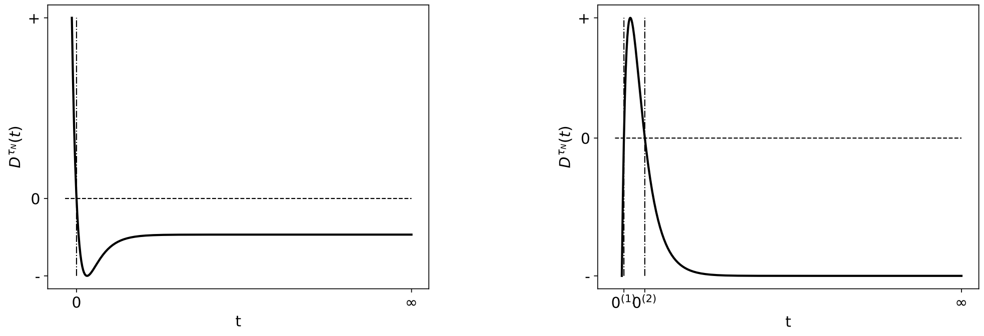
Case II. We consider the case where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, namely,

$$(c_1\mu_1 - c_2\mu_2)(r_1 - r_2) < 0, \quad \text{for } r_1, r_2 \text{ in (1.6) and (1.7).}$$

For the moment, suppose Class 1 has a higher $c\mu$-index and Class 2 has a higher modified $c\mu/\theta$-index. Following similar lines of arguments as in Case I, the backward switching curve $\psi(\tau_N - t)$ follows the dynamic of $D^{\tau_N}(t)$ for some non-trivial time interval $t \in (0, \alpha_N)$. Again, the structure of $D^{\tau_N}(t)$ guarantees that it can have at most two zeros. With Class 2 having a higher modified $c\mu/\theta$-index, the two possible shapes for $D^{\tau_N}(t)$ are demonstrated in Figure A.2, with Figure A.2a crossing zero once and Figure A.2b crossing zero twice. In particular, if $D^{\tau_N}(t)$ has one zero as in Figure A.2a, then it must be that $D^{\tau_N}(t)$ is decreasing at the zero point and eventually converges to $r_1 - r_2 < 0$. Once $D^{\tau_N}(t)$

crosses zero, it will never increase to zero again. Likewise, if $D^{\tau_N}(t)$ has two zeros as in Figure A.2b, then it must be that $D^{\tau_N}(t)$ has positive slope at the first zero, has negative slope at the second zero, and eventually converges to $r_1 - r_2$. Once $D^{\tau_N}(t)$ crosses the second zero point, it will never increase to zero again.

Figure A.2: Possible trajectory of $D^{\tau_1}(t)$ with $c_1\mu_1 > c_2\mu_2$
(modified $c_1\mu_1/\theta_1 <$ modified $c_2\mu_2/\theta_2$)



(a) One zero point

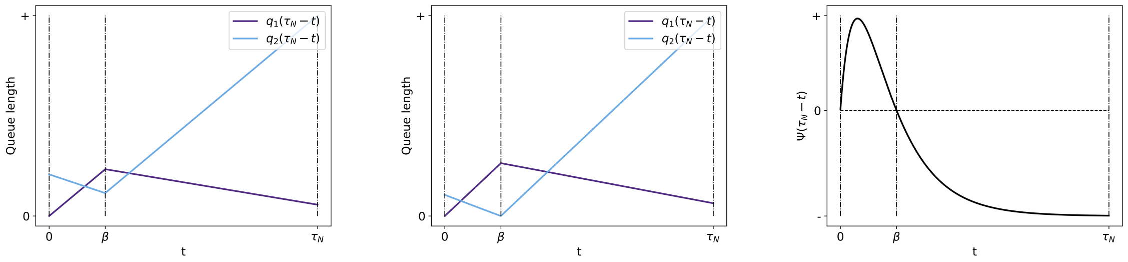
(b) Two zero points

By Proposition 1, for $c_1\mu_1 > c_2\mu_2$, it is optimal to give strict priority to Class 1 when the system state is close enough to the origin. Therefore, τ_N is the last time before τ when q_1^* hits zero. In order to empty q_1^*, strict priority must be given to Class 1 for some non-trivial time interval right before τ_N. This implies that there exits $\varepsilon_{\tau_N} > 0$ such that $D_{\tau_N}(0) = 0$ and $D_{\tau_N}(t) > 0$ for $t \in (0, \varepsilon_{\tau_N})$. In this case, we can rule out Figure A.2a. $D_{\tau_N}(0)$ must be at the first zero in Figure A.2b. Now, let time $\beta > 0$ denote the second zero in Figure A.2b, i.e., $D^{\tau_N}(\beta) = 0$. Then, one of the following three scenarios holds.

Scenario 1. $\tau_N \leq \beta$. The backward switching curve $\psi(\tau_N - t)$ agrees with $D^{\tau_N}(t)$ for all $t \in [0, \tau_N]$. Because $\psi(\tau_N - t) > 0$ for all $t \in (0, \tau_N)$, strict priority is given to Class 1 throughout the transient time horizon.

Scenario 2. $\tau_N > \beta$. The backward switching curve $\psi(\tau_N - t)$ follows $D^{\tau_N}(t)$ for $t \in [0, \beta)$. Both $q_1^*(\tau_N - t)$ and $q_2^*(\tau_N - t)$ stay strictly positive over $t \in (0, \beta)$. At time $t = \beta$, priority is switched from Class 1 to Class 2 (backward in time). In this scenario, we consider the cases where either both queues are strictly positive at $t = \beta$ as in Figure A.3a, or β is a contact point as in Figure A.3b. In either cases, the multipliers $\eta_1^*(\tau_N - t)$ and $\eta_2^*(\tau_N - t)$ stay at zero (or become positive only at one point). Then the backward switching curve $\psi(\tau_N - t)$ further follows $D^{\tau_N}(t)$ for some non-trivial interval, $(\beta, \beta+\delta)$ for some $\delta > 0$. Following similar arguments as in Case I, once crossing zero at $t = \beta$, the backward switching curve $\psi(\tau_N - t)$ remains strictly negative afterwards as shown in Figure A.3c. In this case, the optimal control (forward in time) switches priority once from Class 2 to Class 1.

Figure A.3: Backward state trajectory and switching curve in Scenario 2



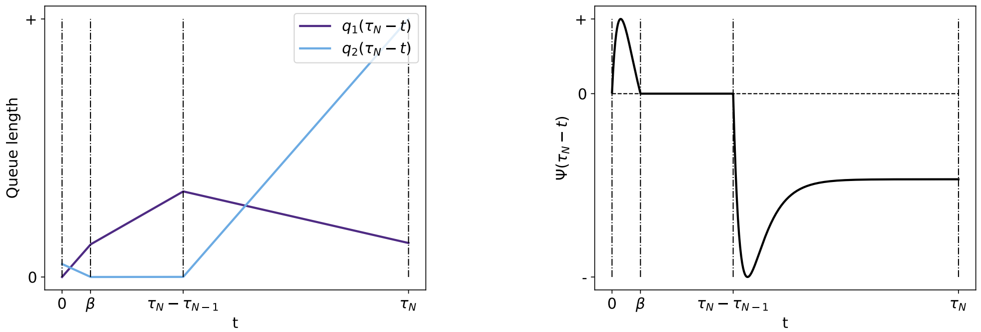
(a) Non-junction time β **(b)** Contact point β **(c)** Backward switching curve

Scenario 3. $\tau_N > \beta$. The backward switching curve $\psi(\tau_N - t)$ follows $D^{\tau_N}(t)$ for all $t \in [0, \beta)$. Both $q_1^*(\tau_N - t)$ and $q_2^*(\tau_N - t)$ stay strictly positive over $t \in (0, \beta)$. Different from the Scenario 2, β is an exit point (forward in time) for the trajectory of q_2^*; see Figure A.4a. Correspondingly, the entry point is τ_{N-1}. At time τ_{N-1}, the switching curve $\psi(\tau_{N-1}) = 0$. Now, we repeat the structural derivation for the backward switching curve starting from τ_{N-1}, namely, for the function $\psi(\tau_{N-1} - t)$. In order to drive q_2^* to zero at time τ_{N-1}, strict priority must be assigned to q_2^* for some amount of time right before τ_{N-1}. As such, there exits $\varepsilon_{\tau_{N-1}} > 0$ such that $D^{\tau_{N-1}}(0) = 0$ and $D^{\tau_{N-1}}(t) < 0$ for $t \in (0, \varepsilon_{\tau_{N-1}})$. Again, following similar arguments as in Case I, we can show that once crossing zero at τ_{N-1}, the switching curve $\psi(\tau_{N-1} - t)$ remains strictly negative for $t \in (0, \tau_{N-1})$. In this case, the optimal control (forward in time) switches priority once from Class 2 to Class 1. The structure of the backward switching curve in this case is illustrated in Figure A.4b.

In all the three scenarios above, the optimal control either assigns strict priority to Class 1 throughout, or switches priority once from Class 2 to Class 1.

When Class 2 has a higher $c\mu$-index and Class 1 has a higher modified $c\mu/\theta$-index, the proof holds in a similar fashion. In this case, the optimal control either invariantly assigns

Figure A.4: Backward state trajectory and switching curve in Scenario 3



(a) Exit time β

(b) Backward switching curve

strict priority to Class 2, or switches once from prioritizing Class 1 to Class 2. □

A.2.8 Proof of Proposition 2

PROOF: First, as shown in Proposition 3, if the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, then the modified $c\mu/\theta$-rule (the $c\mu$-rule) is optimal throughout the transient time horizon and the claim follows.

Next, consider the case where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, namely,

$$(c_1\mu_1 - c_2\mu_2)(r_1 - r_2) < 0, \quad \text{for } r_1, r_2 \text{ in (1.6) and (1.7).}$$

By Propositions 1 and 3, when the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, the optimal control follows the $c\mu$-rule near the origin and switches priority at most once along the trajectory. However, it remains to be shown whether or not the optimal control will ever switch priority. Namely, the work left is to prove that there exists a set of initial conditions from which the optimal trajectories switch priority from one class to the other. In this proof, we establish the existence of such initial values and provide a partial characterization of the states at which the system will follow the modified $c\mu/\theta$-rule.

For the moment, we consider the case where the $c\mu$-rule prioritizes Class 1 and the modified $c\mu/\theta$-rule prioritizes Class 2. We first note that by the definition of τ_1, both queues are strictly positive for $t < \tau_1$. Thus, the multipliers $\eta_1^*(t) = \eta_2^*(t) = 0$ for $t < \tau_1$.

By Lemma 4, the backward switching curve before τ_1 is characterized as follows

$$\begin{aligned}\psi(\tau_1 - t) = r_1 - r_2 &+ (\mu_1 A_1(\tau_1) - \mu_2 A_2(\tau_1))\, e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_1-t)} \\ &- (\mu_1 A_1(\tau_1) - \mu_2 A_2(\tau_1))\, e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_1-t)},\end{aligned} \tag{A.13}$$

where $A_1(\tau_1), A_2(\tau_1)$ are constants in $\mathbb{R}$, and $0 \leq \zeta < \gamma_1 + \gamma_2 + \theta_1 + \theta_2$.

Note that due to class-transition, when one queue gets emptied, the other queue cannot be arbitrarily large. In the case where Class 1 gets emptied at τ_1, it holds that $q_1^*(\tau_1) = 0$, and for any $\varepsilon > 0$,

$$q_2^*(\tau_1) < (s\mu_1 - \lambda_1)/\gamma_2 + \varepsilon.$$

Similarly, in the other case where Class 2 gets emptied at τ_1, it holds that $q_2^*(\tau_1) = 0$ and for any $\varepsilon > 0$, we have $q_1^*(\tau_1) < (s\mu_2 - \lambda_2)/\gamma_1 + \varepsilon$. Since $q_1^*(\tau_1)$ and $q_2^*(\tau_1)$ are uniformly bounded for any initialization, using the fact that $p^*(t) = \nabla_q \Xi(q^*(t))$, it holds that $p_1^*(\tau_1)$ and $p_2^*(\tau_1)$ are bounded for any initialization.

Now, from the form of $A_1(\tau_1)$ and $A_2(\tau_1)$ in the proof of Lemma 4, in particular, (A.9) and (A.11), we see that $A_1(\tau_1)$ and $A_2(\tau_1)$ are bounded if $p_1^*(\tau_1)$ and $p_2^*(\tau_1)$ are bounded, uniformly for any initialization.

Lastly, if the system is initialized with a large queue, τ_1, the time to empty queue 1 for the first time forward in time, is large. As t approaches τ_1 in (A.13), the sign of the backward switching curve will eventually be governed by $r_1 - r_2$. In other words, for M sufficiently large, the modified $c\mu/\theta$-rule is optimal at time t if $q_1(t) + q_2(t) > M$.

The arguments for the other case where the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1 follow by symmetry. □

A.2.9 Proof of Theorem 2

PROOF: The statement of Theorem 2 follows directly from Propositions 1, 2, and 3. □

A.2.10 Proof of Proposition 4

PROOF: For $c_1\mu_1 < c_2\mu_2$ and $r_1 > r_2$, Theorem 2 indicates that a one-time switch in priority from Class 1 to Class 2 will take place if the system is initialized far enough from the origin. To derive the policy curve at which (state) the switching takes place, we apply the Hamiltonian condition (H). In particular, let (a_1, a_2) be a state where priority is just switched from Class 1 to Class 2, i.e., (a_1, a_2) is on the policy curve, where $a_1 \geq 0$ and $a_2 > 0$. We denote the time of the switching by t_1. We also denote $t_2 > t_1$ as the time Class 2 gets emptied and $t_3 = \tau^* > t_2$ as the time Class 1 gets emptied.

Starting from time t_1, the dynamic of the adjoint vector for $p^*(t)$ is specified by (ADJ) as

$$\begin{aligned}
p_1^*(t) &= K_1(t_1)e^{t\theta_1} + e^{t\theta_1}\int_0^t e^{-s\theta_1}\left(-c_1 + \eta_1^*(s)\right)ds \\
p_2^*(t) &= K_2(t_1)e^{t(\theta_2+\gamma_2)} + \frac{K_1\gamma_2}{\gamma_2 - \theta_1 + \theta_2}\left(e^{t\theta_1} - e^{t(\theta_2+\gamma_2)}\right) \\
&\quad + \frac{\gamma_2}{\gamma_2 - \theta_1 + \theta_2} e^{t\theta_1}\int_0^t e^{-s\theta_1}\left(-c_1 + \eta_1^*(s)\right)ds \\
&\quad - \frac{\gamma_2}{\gamma_2 - \theta_1 + \theta_2} e^{t(\theta_2+\gamma_2)}\int_0^t e^{s(-\gamma_2-\theta_2)}\left(-c_1 + \eta_1^*(s)\right)ds \\
&\quad + e^{t(\theta_2+\gamma_2)}\int_0^t e^{s(-\gamma_2-\theta_2)}\left(-c_2 + \eta_2^*(s)\right)ds,
\end{aligned} \tag{A.14}$$

where $K_1(t_1)$ and $K_2(t_1)$ are constants that depends on $p_1^*(t_1)$ and $p_2^*(t_1)$. Since there is no other switch in priority (Proposition 3) after t_1, $q_1(t) > 0$ for $t \in (t_1, t_3)$, and $q_2(t) > 0$ for $t \in (t_1, t_2)$. Then, (A.14) reduces to

$$\begin{aligned}
p_1^*(t) &= \frac{c_1}{\theta_1} + e^{\theta_1 t}K_1(t_1) \quad \text{for } t \in [t_1, t_3] \\
p_2^*(t) &= \frac{c_2}{\theta_2 + \gamma_2} + \frac{c_1\gamma_2}{\theta_1(\theta_2 + \gamma_2)} + \frac{e^{\theta_1 t}\gamma_2 K_1(t_1) + e^{(\theta_2+\gamma_2)t}\left(-\gamma_2 K_1(t_1) + (\gamma_2 - \theta_1 + \theta_2)K_2(t_1)\right)}{\gamma_2 - \theta_j + \theta_2},
\end{aligned} \tag{A.15}$$

for $t \in [t_1, t_2]$.

The rest of the analysis is divided into three time intervals. For each one of the three intervals, we characterize the state trajectory $q^*(t)$ and the adjoint vector $p^*(t)$. Then, plugging the values of $q^*(t)$ and $p^*(t)$ into the Hamiltonian and utilizing the Hamiltonian condition (H), we are able to characterize the constants $K_1(t_1), K_2(t_1)$ in (A.15) as well as the policy curve. These steps will become self-explanatory as the proof proceeds.

Case I. q_1^* is strictly positive or has just reached zero at time t_1. In this case, full service capacity s is assigned to Class 1 at time t_1-.

Interval 1: At time t_1-, we assign s servers to Class 1 and 0 servers to Class 2.

$$\begin{aligned}
q_1^*(t_1-) &= a_1 \\
q_2^*(t_1-) &= a_2 \\
H(q^*(t_1-), z^*(t_1-), p^*(t_1-)) &= c_1a_1 + c_2a_2 + (a_2\gamma_2 - a_1\theta_1 + \lambda_1 - s\mu_1)\left(\frac{c_1}{\theta_1} + K_1(t_1)\right) \\
&\quad + (-a_2(\theta_2+\gamma_2) + \lambda_2)\left(\frac{c_1\gamma_2 + c_2\theta_1}{\gamma_2\theta_1 + \theta_1\theta_2} + K_2(t_1)\right).
\end{aligned}$$

Interval 2: Over $[t_1, t_2)$, we assign 0 server to Class 1 and s servers to Class 2, and Class 2 gets emptied at time t_2.

$$\begin{aligned}
q_1^*(t) &= -\frac{1}{\theta_1(\theta_2+\gamma_2)(\gamma_2-\theta_1+\theta_2)} e^{-(t-t_1)(\gamma_2+\theta_1+\theta_2)} \\
&\quad \Big(e^{(t-t_1)\theta_1}\gamma_2\theta_1(a_2(\theta_2+\gamma_2) - \lambda_2 + s\mu_2) \\
&\quad - e^{(t-t_1)(\theta_2+\gamma_2)}(\theta_2+\gamma_2)(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) \\
&\quad - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\
&\quad - e^{(t-t_1)(\gamma_2+\theta_1+\theta_2)}(\gamma_2-\theta_1+\theta_2)(\theta_2\lambda_1 + \gamma_2(\lambda_1+\lambda_2-s\mu_2))\Big), \\
q_2^*(t) &= \frac{1}{\theta_2+\gamma_2} e^{-(t-t_1)(\theta_2+\gamma_2)}\left(a_2(\theta_2+\gamma_2) + (-1 + e^{(t-t_1)(\theta_2+\gamma_2)})(\lambda_2 - s\mu_2)\right), \\
t_2 - t_1 &= \frac{1}{\theta_2+\gamma_2}\log\left(\frac{-a_2\gamma_2 - a_2\theta_2 + \lambda_2 - s\mu_2}{\lambda_2 - s\mu_2}\right) \\
H(q^*(t), z^*(t), p^*(t)) &= \frac{1}{\theta_1(\theta_2+\gamma_2)}\big\{c_1\theta_2\lambda_1 + c_2\theta_1(\lambda_2 - s\mu_2) + c_1\gamma_2(\lambda_1+\lambda_2-s\mu_2) \\
&\quad - \theta_1(\theta_2+\gamma_2)\big[a_1\theta_1K_1(t_1) - \lambda_1K_1(t_1) + a_2\theta_2K_2(t_1) - \lambda_2K_2(t_1) \\
&\quad + s\mu_2K_2(t_1) + a_2\gamma_2(-K_1(t_1) + K_2(t_1))\big]\big\}.
\end{aligned}$$

Putting the analysis for Interval 1 and Interval 2 together, we can solve for $K_1(t_1)$ and $K_2(t_1)$ from the system of equations $H(q^*(t_1), z^*(t_1), p^*(t_1)) = 0$ and $H(q^*(t), z^*(t), p^*(t)) =$

0 for $t \in [t_1, t_2)$. In particular,

$$\begin{aligned}
K_1(t_1) &= \frac{c_1(-a_2(\theta_2+\gamma_2)+\lambda_2)\mu_1 + c_2 a_2 \theta_1 \mu_2 + c_1(a_2\gamma_2 + \lambda_1 - s\mu_1)\mu_2}{\theta_1(a_2(\theta_2+\gamma_2)-\lambda_2)\mu_1 + \theta_1(-a_2\gamma_2 + a_1\theta_1 - \lambda_1 + s\mu_1)\mu_2} \\
K_2(t_1) &= -\frac{c_1\gamma_2 + c_2\theta_1}{\gamma_2\theta_1 + \theta_1\theta_2} + \frac{(c_1 a_1 + c_2 a_2)\mu_1}{a_2(\theta_2+\gamma_2)\mu_1 - \lambda_2\mu_1 - a_2\gamma_2\mu_2 + (a_1\theta_1 - \lambda_1 + s\mu_1)\mu_2}.
\end{aligned} \tag{A.16}$$

Interval 3: Over $[t_2, t_3]$, we assign enough servers to maintain Class 2 at zero and the rest of the service capacity to Class 1. Class 1 gets emptied at time t_3.

$$\begin{aligned}
q_1^*(t) = {} & \frac{e^{-(t-t_2)\theta_1}}{\theta_1(-\gamma_2+\theta_1-\theta_2)} \Bigg(-(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\
& \left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}} - \frac{1}{\mu_2}((\lambda_2 - s\mu_2)(-(\gamma_2-\theta_1+\theta_2)\mu_1 + \gamma_2\mu_2) \\
& + e^{(t-t_2)\theta_1}(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)) \Bigg),
\end{aligned}$$

$$\begin{aligned}
q_2^*(t) &= 0, \\
t_3 - t_2 &= \frac{1}{\theta_1}\log\Bigg(\frac{1}{(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}\Bigg((\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2) \\
& - \mu_2(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\
& \left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg)\Bigg).
\end{aligned}$$

Note that $[t_2, t_3)$ is a boundary arc for q_2^* and an interior arc for q_1^*. As $\dot{q}_2^*(t) = 0$, we have

$$H(q^*(t), z^*(t), p^*(t)) = p_1^*(t)\dot{q}_1^*(t) + p_2^*(t)\dot{q}_2^*(t) + c_1 q_1^*(t) + c_2 q_2^*(t) = p_1^*(t)\dot{q}_1^*(t) + c_1 q_1^*(t).$$

Then, plugging the expression of $q_1^*(t)$, (A.15), into $H(q^*(t), z^*(t), p^*(t))$, we get

$$\begin{aligned}
& H(q^*(t), z^*(t), p^*(t)) \\
& = \frac{K_1(t_1)}{\mu_2(\gamma_2-\theta_1+\theta_2)}\Bigg((\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2)\left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma_2}} \\
& \quad - \mu_2\big(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2\big)\Bigg) \\
& \quad + \frac{c_1(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\theta_1\mu_2}.
\end{aligned}$$

Plugging the value of $K_1(t_1)$, (A.16), into the equality $H(q^*(t), z^*(t), p^*(t)) = 0$ for $t \in [t_2, t_3)$ establishes the relationship (a_1, a_2) must satisfy. This gives the policy curve in Proposition 4.

Case II. q_1^* is equal to zero at time t_1 and has been maintained at zero over interval $[t_1 - \varepsilon, t_1]$ for some $\varepsilon > 0$. In this case, the right amount of service capacity is assigned to Class 1 at time t_1- to maintain q_1^* at zero.

Interval 1: At time t_1-, we assign $(\lambda_1 + \gamma_2 q_2^*(t_1-))/\mu_1$ servers to Class 1 and the rest of the servers to Class 2.

$$
\begin{aligned}
& q_1^*(t_1-) = 0 \\
& q_2^*(t_1-) = a_2 \\
& H(q^*(t_1-), z^*(t_1-), p^*(t_1-)) \\
& = c_2 a_2 + \left(-a_2(\gamma_2 + \theta_2) + \lambda_2 - s\mu_2 + \frac{(a_2\gamma_2 + \lambda_1)\mu_2}{\mu_1} \right) \left(\frac{c_1\gamma_2 + c_2\theta_1}{\gamma_2\theta_1 + \theta_1\theta_2} + K_2(t_1) \right)
\end{aligned}
$$

Interval 2: Over $[t_1, t_2)$, we assign 0 servers to Class 1 and s servers to Class 2, and Class 2 gets emptied at time t_2.

$$
\begin{aligned}
q_1^*(t) = & -\frac{1}{\theta_1(\theta_2 + \gamma_2)(\gamma_2 - \theta_1 + \theta_2)} e^{-(t-t_1)(\gamma_2 + \theta_1 + \theta_2)} \\
& \Big(e^{(t-t_1)\theta_1} \gamma_2 \theta_1 (a_2(\theta_2 + \gamma_2) - \lambda_2 + s\mu_2) \\
& - e^{(t-t_1)(\theta_2+\gamma_2)} (\theta_2 + \gamma_2)(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\
& - e^{(t-t_1)(\gamma_2+\theta_1+\theta_2)} (\gamma_2 - \theta_1 + \theta_2)(\theta_2\lambda_1 + \gamma_2(\lambda_1 + \lambda_2 - s\mu_2)) \Big), \\
q_2^*(t) = & \frac{1}{\theta_2 + \gamma_2} e^{-(t-t_1)(\theta_2+\gamma_2)} \left(a_2(\theta_2 + \gamma_2) + (-1 + e^{(t-t_1)(\theta_2+\gamma_2)})(\lambda_2 - s\mu_2) \right), \\
t_2 - t_1 = & \frac{1}{\theta_2 + \gamma_2} \log \left(\frac{-a_2\gamma_2 - a_2\theta_2 + \lambda_2 - s\mu_2}{\lambda_2 - s\mu_2} \right) \\
H(q^*(t), z^*(t), p^*(t)) = & \frac{1}{\theta_1(\theta_2 + \gamma_2)} \{ c_1\theta_2\lambda_1 + c_2\theta_1(\lambda_2 - s\mu_2) + c_1\gamma_2(\lambda_1 + \lambda_2 - s\mu_2) \\
& - \theta_1(\theta_2 + \gamma_2) \big[-\lambda_1 K_1(t_1) + a_2\theta_2 K_2(t_1) - \lambda_2 K_2(t_1) \\
& + s\mu_2 K_2(t_1) + a_2\gamma_2(-K_1(t_1) + K_2(t_1)) \big] \}.
\end{aligned}
$$

Putting the analysis for Interval 1 and Interval 2 together, we can solve for $K_1(t_1)$ and $K_2(t_1)$ from the system of equations $H(q^*(t_1), z^*(t_1), p^*(t_1)) = 0$ and $H(q^*(t), z^*(t), p^*(t)) = 0$ for $t \in [t_1, t_2)$. In particular, we get

$$\begin{aligned}
K_1(t_1) &= \frac{c_1(-a_2(\theta_2+\gamma_2)+\lambda_2)\mu_1 + c_2 a_2 \theta_1 \mu_2 + c_1(a_2\gamma_2 + \lambda_1 - s\mu_1)\mu_2}{\theta_1(a_2(\theta_2+\gamma_2)-\lambda_2)\mu_1 + \theta_1(-a_2\gamma_2 - \lambda_1 + s\mu_1)\mu_2} \\
K_2(t_1) &= -\frac{c_1\gamma_2 + c_2\theta_1}{\gamma_2\theta_1 + \theta_1\theta_2} + \frac{c_2 a_2 \mu_1}{a_2(\theta_2+\gamma_2)\mu_1 - \lambda_2\mu_1 - a_2\gamma_2\mu_2 + (-\lambda_1 + s\mu_1)\mu_2}.
\end{aligned} \tag{A.17}$$

Interval 3: Over $[t_2, t_3]$, we assign enough servers to maintain Class 2 at zero and the rest of the service capacity to Class 1. Class 1 gets emptied at time t_3.

$$\begin{aligned}
q_1^*(t) = {} & \frac{e^{-(t-t_2)\theta_1}}{\theta_1(-\gamma_2+\theta_1-\theta_2)} \Bigg(-(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\
& \left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2 + s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}} - \frac{1}{\mu_2}((\lambda_2 - s\mu_2)(-(\gamma_2-\theta_1+\theta_2)\mu_1 + \gamma_2\mu_2) \\
& + e^{(t-t_2)\theta_1}(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2))\Bigg),
\end{aligned}$$

$$q_2^*(t) = 0,$$

$$\begin{aligned}
t_3 - t_2 = {} & \frac{1}{\theta_1}\log\Bigg(\frac{1}{(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}\Bigg((\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2) \\
& - \mu_2\,(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2)\left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2 + s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg)\Bigg).
\end{aligned}$$

Note that $[t_2, t_3)$ is a boundary arc for q_2^* and an interior arc for q_1^*. As $\dot{q}_2^*(t) = 0$, we have

$$H(q^*(t), z^*(t), p^*(t)) = p_1^*(t)\dot{q}_1^*(t) + p_2^*(t)\dot{q}_2^*(t) + c_1 q_1^*(t) + c_2 q_2^*(t) = p_1^*(t)\dot{q}_1^*(t) + c_1 q_1^*(t).$$

Then, plugging the expression of $q_1^*(t)$, (A.15) into $H(q^*(t), z^*(t), p^*(t))$, we get

$$\begin{aligned}
H(q^*(t), z^*(t), p^*(t)) = {} & \frac{K_1(t_1)}{\mu_2(\gamma_2-\theta_1+\theta_2)}\Bigg(-\mu_2\big(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2\big) \\
& + (\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2)\left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2 + s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg) \\
& + \frac{c_1(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\theta_1\mu_2}.
\end{aligned}$$

Plugging the value of $K_1(t_1)$, (A.17), into the equality $H(q^*(t), z^*(t), p^*(t)) = 0$ for $t \in [t_2, t_3)$ establishes the relationship a_2 must satisfy in order for priority to be switched from P_1 to P_2 given that q_2^* is at level a_2 and q_1^* has been maintained at zero for some amount of time. It is easy to see that setting $H(q^*(t), z^*(t), p^*(t)) = 0$ in Case 2 retrieves the point $(0, a_2)$ on the switching curve established in Case 1.

It is important to note that the switching point $(0, a_2)$ analyzed in Case 2 assumes that q_1^* has been maintained at zero before priority is switched. On the other hand, the switching point $(0, a_2)$ on the policy curve derived in Case 1 assumes that q_1^* just hits zero when priority is switched from P_1 to P_2. It is well expected that the switching points in the two cases coincide with each other. Our proof rigorously verifies this. □

A.3 Proof of Theorem 3

PROOF: We dissect the transient optimization problem over the entire time horizon $[0, T + \tau^*]$ into a two-stage optimal control problem. The first-stage problem (1.12) is over the time interval $[0, T)$. The second problem (1.13) is over the time interval $[T, T + \tau^*]$ and its initial condition is equal to the terminal state in problem (1.12). We also note that (1.13) over $[T, T + \tau^*]$ is equivalent to (F2$'$) over $[0, \tau^*]$ with the appropriate initial condition. In what follows, to distinguish problems (1.12) and (F2$'$), we will append superscripts [1] and [2] to the queue length processes, dual vectors, etc., associated with problems (1.12) and (F2$'$), respectively. For example, we will write the time horizon for (1.12) as $[0^{[1]}, T^{[1]})$ and the time horizon for (F2$'$) as $[0^{[2]}, \tau^{*[2]}]$, where $0^{[1]} = 0, T^{[1]} = T, 0^{[2]} = 0$, and $\tau^{*[2]} = \tau^*$.

We first note that for the second-stage problem (F2$'$) over $[0^{[2]}, \tau^{*[2]}]$, it follows directly from Theorem 2 that the optimal scheduling policy follows the $c\mu$-rule when the states are sufficiently small, and follows the modified $c\mu/\theta$-rule when the states are sufficiently large. The work left is to show that the optimal scheduling policy switches priority at most once over the entire transient time horizon $[0, T + \tau^*]$. To do this, we establish an analogous version of Proposition 3 below.

Claim A. Under Assumptions 1 and 4, for the transient optimal control problem (1.12) and (F2$'$), if the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, then the optimal transient scheduling policy does not switch priority. If the two index rules prioritize different classes, then the optimal transient scheduling policy switches priority at most once over the transient time horizon $[0, T + \tau^*]$.

To establish Claim A, we observe that problem (1.12) over the initial period $[0^{[1]}, T^{[1]})$ is an optimal control problem with fixed time, free terminal state, terminal cost, and no state constraints. For this type of problems, the following version of Pontryagin's Minimum Principle applies.

Lemma 5 (Theorem 3.4 in Grass et al. (2008))**.** *Under Assumption 4, let $z^{*[1]}$ be an optimal solution to* (1.12)*, and $q^{*[1]}$ be the corresponding state trajectory. There exists a continuous and piecewise continuously differentiable adjoint vector $p^{*[1]} : [0^{[1]}, T^{[1]}] \to \mathbb{R}^2_+$ satisfying for all $t \in [0^{[1]}, T^{[1]}]$:*

1. *Ordinary Differential Equation condition (ODE):*

$$q^{*[1]}(0) = q_0, \quad \dot{q}^{*[1]}(t) = f^{[1]}\left(q^{*[1]}(t), z^{*[1]}(t), t\right)$$

2. *Adjoint Vector condition (ADJ):*

$$\dot{p}^{*[1]}(t) = -\nabla_q H^{[1]}(q^{*[1]}(t), z^{*[1]}(t), p^{*[1]}(t), t)$$

3. *Minimization condition (M):*

$$H^{[1]}(q^{*[1]}(t), z^{*[1]}(t), p^{*[1]}(t), t) = \min_z \{H^{[1]}(q^{*[1]}(t), z^{[1]}(t), p^{*[1]}(t), t)\}$$

4. *Transversality condition (T):*

$$p^{*[1]}(T^{[1]}) = \nabla_q \Xi(q^{*[1]}(T^{[1]})). \tag{A.18}$$

Note that as we allow for time-varying arrival rates on $[0^{[1]}, T^{[1]}]$, $f^{[1]}$ and $H^{[1]}$ have an explicit time component. We draw several connections between the two versions of Pontryagin's Minimum Principles in Lemma 5 and Theorem 8. First, the construction of the Hamiltonian and conditions (ODE) and (M) are essentially the same in the two versions, except that fluid dynamic $f^{[1]}$ and the Hamiltonian $H^{[1]}$ in Lemma 5 are time-dependent through $\lambda^{[1]}(t)$. Second, Minimization condition (M) in Lemma 5 specifies the dynamics of the adjoint vector for problem (1.12)

$$\dot{p}_1^{*[1]}(t) = (\theta_1+\gamma_1)p_1^{*[1]}(t) - \gamma_1 p_2^{*[1]}(t) - c_1, \quad \dot{p}_2^{*[1]}(t) = (\theta_2+\gamma_2)p_2^{*[1]}(t) - \gamma_2 p_1^{*[1]}(t) - c_2, \tag{A.19}$$

while the Minimization condition (M) in Theorem 8 gives that for problem (F2′)

$$\begin{aligned} \dot{p}_1^{*[2]}(t) &= (\theta_1+\gamma_1)p_1^{*[2]}(t) - \gamma_1 p_2^{*[2]}(t) - c_1 + \eta_1^{*[2]}(t), \\ \dot{p}_2^{*[2]}(t) &= (\theta_2+\gamma_2)p_2^{*[2]}(t) - \gamma_2 p_1^{*[2]}(t) - c_2 + \eta_2^{*[2]}(t). \end{aligned} \tag{A.20}$$

Comparing (A.19) with (A.20), we note that the adjoint vectors for problems (1.12) and (F2′) follow the same dynamic when the state constraints in (F2′) are not active, namely, when both queues are strictly positive. Third, Transversality condition (A.18) holds exclusively for the first-stage problem (1.12) which has fixed terminal time and no terminal (state) constraint.

To this end, for the second-stage problem (F2′), let $\tau_1^{[2]}$ denote the first time one of the two queues hits zero. The pseudo switching curve associated with $\tau_1^{[2]}$ is given by

$$\begin{aligned} D^{\tau_1^{[2]}}(t) = r_1 - r_2 &+ \left(\mu_1 A_1^{[2]}(\tau_1^{[2]}) - \mu_2 A_2^{[2]}(\tau_1^{[2]})\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_1^{[2]}-t)} \\ &- \left(\mu_1 A_1^{[2]}(\tau_1^{[2]}) - \mu_2 A_2^{[2]}(\tau_1^{[2]})\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_1^{[2]}-t)}, \quad \text{for all } t \geq 0. \end{aligned}$$

Since both queues are strictly positive for $t \in [0^{[2]}, \tau_1^{[2]})$, the multiplies $\eta_1^{*[2]}(t) = \eta_2^{*[2]}(t) = 0$ for $t \in [0^{[2]}, \tau_1^{[2]})$. It follows that the switching curve for problem (F2′) backward from time $\tau_1^{[2]}$ agrees with $D^{\tau_1^{[2]}}(t)$, namely,

$$\psi^{[2]}(\tau_1^{[2]} - t) = D^{\tau_1^{[2]}}(t) \quad \text{for all } t \in (0^{[2]}, \tau_1^{[2]}].$$

Now, recall that $\Xi(q_0)$ is the value function from state q_0 in the second-stage problem (F2′). Thus, it follows from Transversality condition (A.18) in Lemma 5 that

$$p^{*[1]}(T^{[1]}) = \nabla_q \Xi(q^{*[1]}(T^{[1]})) = \nabla_q \Xi(q^{*[2]}(0^{[2]})) = p^{*[2]}(0^{[2]}). \tag{A.21}$$

By (A.21), together with the fact that the adjoint vectors for problems (1.12) and (F2′) follow the same dynamic when both queues are strictly positive, it is easy to see that the backward switching curve for the first-stage problem $\psi^{[1]}$ is connected to the pseudo switching curve for the second-stage problem $D^{\tau_1^{[2]}}$ via

$$\psi^{[1]}(T^{[1]}-t) = D^{\tau_1^{[2]}}(\tau_1^{[2]}+t), \quad \text{for all } t \in [0^{[1]}, T^{[1]}]. \tag{A.22}$$

It follows from (A.22) that analyzing the first-stage backward switching curve $\psi^{[1]}(T^{[1]}-t)$ is equivalent to analyzing the second-stage pseudo switching curve $D^{\tau_1^{[2]}}(\tau_1^{[2]}+t)$ extended beyond the beginning epoch of the second-stage problem for another $T^{[1]}$ time units. It is then straightforward to see that the arguments in the proof of Proposition 3 extend to the first-stage problem (1.12) and Claim A follows. □

A.4 The Special Cases with No Class-Transition and Abandonment

The special case where $\gamma_1 = \gamma_2 = \theta_1 = \theta_2 = 0$ is not covered in Theorem 2, as Assumption 1 does not hold in this case. However, the same lines of argument, utilizing the Pontryagin's Minimum Principle, can be use in this case to establish the optimality of the $c\mu$-rule. Indeed, the proof is more concise here and nicely illustrates the main idea behind our proof strategy.

Corollary 2. *If $\gamma_1 = \gamma_2 = \theta_1 = \theta_2 = 0$, and $s > \lambda_1/\mu_1 + \lambda_2/\lambda_2$, the $c\mu$-rule is optimal for the transient fluid optimal control problem (F2′).*

PROOF: Suppose without loss of generality that $c_1\mu_1 > c_2\mu_2$. The queue length process evolves as

$$\dot{q}_1(t) = \lambda_1 - \mu_1 z_1(t) \quad \text{and} \quad \dot{q}_2(t) = \lambda_2 - \mu_2 z_2(t).$$

The Hamiltonian is

$$\begin{aligned} H(q(t), z(t), p(t)) &= p_1(t)\dot{q}_1(t) + p_2(t)\dot{q}_2(t) + c_1 q_1(t) + c_2 q_2(t) \\ &= p_1(t)\left(\lambda_1 - \mu_1 z_1(t)\right) + p_2(t)\left(\lambda_2 - \mu_2 z_2(t)\right) + c_1 q_1(t) + c_2 q_2(t). \end{aligned}$$

The augmented Halmiltonian takes the form

$$\begin{aligned}L(q(t),z(t),p(t),\eta(t),\xi(t)) &= H(x,s,p)+\eta(t)^T g(q(t))+\xi(t)^T h(z(t))\\ &= p_1(t)\,(\lambda_1-\mu_1 z_1(t))+p_2(t)\,(\lambda_2-\mu_2 z_2(t))+c_1q_1(t)+c_2q_2(t)\\ &\quad +\eta_1(t)(-q_1(t))+\eta_2(t)(-q_2(t))+\xi_1(t)(z_1(t)+z_2(t)-s)\\ &\quad +\xi_2(t)(-z_1(t))+\xi_3(t)(-z_2(t)).\end{aligned}$$

Since $\dot{p}^*(t) = -\nabla_q L(q^*(t),z^*(t),p^*(t),\eta^*(t),\xi^*(t))$, we have

$$\dot{p}_1^*(t) = -c_1+\eta_1^*(t) \quad \text{and} \quad \dot{p}_2^*(t) = -c_2+\eta_2^*(t). \tag{A.23}$$

Hence,

$$p_1^*(t) = -c_1 t+\int_0^t \eta_1^*(s)ds+K_1 \quad \text{and} \quad p_2^*(t) = -c_2 t+\int_0^t \eta_2^*(s)ds+K_2,$$

where K_1 and K_2 are constants that depend on $p^*(0)$.

The switching curve is

$$\psi(t) = \mu_1 p_1^*(t)-\mu_2 p_2^*(t).$$

Proposition 1 still holds in this case. Hence, when the queue length process is arbitrarily close to the origin, the $c\mu$-rule is optimal and Class 1 should be given strict priority. Let τ_N be the last time epoch (forward in time) when $q_1^*(t)$ hits zero, i.e.,

$$\tau_N = \sup\left\{t : q_1^*(t) = 0,\, q_1^*(t-\varepsilon) > 0 \text{ for some } \varepsilon > 0\right\}.$$

Following the same lines of arguments as in Lemmas 2 and 3, we have the switching curve $\psi(t) = 0$ for $t \geq \tau_N$.

We next characterize the optimal control before τ_N. To this end, observe that by construction, both queues are strictly positive before τ_N. Therefore, there exists a non-trivial period $[0,\beta], \beta < \tau_N$, such that for $t \in [0,\beta]$, the backward switching curve is characterized by

$$\psi(\tau_N-t) = \psi(\tau_N)+(c_1\mu_1-c_2\mu_2)t+\left(\mu_2\int_{\tau_N-t}^{\tau_N}\eta_2^*(s)ds-\mu_1\int_{\tau_N-t}^{\tau_N}\eta_1^*(s)ds\right) = (c_1\mu_1-c_2\mu_2)t. \tag{A.24}$$

Since $c_1\mu_1 > c_2\mu_2$, the significance of (A.24) is that strict priority must be assigned to Class 1 during this period. Moreover, as no queue has the possibility to hit zero over this period, the characterization of the switching curve (A.24) indeed holds for all $t \in [0, \tau_N]$. Namely, strict priority to Class 1 is optimal throughout $[0, \tau^*]$. □

A.4.1 Full Characterization of the Dual Vectors When $\gamma_1 = \gamma_2 = \theta_1 = \theta_2 = 0$

When establishing the optimal scheduling policy, we use Pontryagin's Minimum Principle to derive structural properties of the dual vectors $(p^*(t),\ \eta^*(t),\ \xi^*(t))$ without characterizing their expressions explicitly. The latter step can be prohibitively hard for systems with convoluted dynamics, as is the case for our model with both abandonment and class-transition. On the other hand, for simplified systems without abandonment or class-transition, we can provide a full characterization of the dual vectors. We next illustrate the derivation.

By Corollary 3, the $c\mu$-rule is optimal at all time for systems without abandonment and without class-transition. Suppose without loss of generality that the $c\mu$-rule prioritizes Class 1, i.e., $c_1\mu_1 > c_2\mu_2$. In this case, the value function associated with state (a_1, a_2) is equal to the cost of emptying the system under P_1 when the system is initialized at (a_1, a_2). We can then calculate the value function by solving the state trajectory and the cost directly. Specifically, the value function takes the form

$$\Xi(a_1,a_2) = \frac{1}{2(\lambda_1 - s\mu_1)}\left(-c_1 a_1^2 + \frac{c_2\left(a_2^2\mu_1(-\lambda_1 + s\mu_1) + a_1^2\lambda_2\mu_2 - 2a_1a_2(\lambda_1 - s\mu_1)\mu_2\right)}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2}\right).$$

For a fixed initial condition, q_0, let $q^*(t)$ denote the (optimal) state trajectory under P_1, which can be solved directly. Along the optimal state trajectory, τ_1 is the time epoch when q_1^* first gets emptied. q_1^* is then maintained at zero after time τ_1, until q_2^* reaches zero at time τ^*.

Using the fact that there exists an adjoint vector $p^*(t) = \nabla_q \Xi(q_1^*(t), q_2^*(t))$, we have

$$\begin{aligned}
p_1^*(t) &= \frac{1}{\lambda_1 - s\mu_1}\left(-c_1 q_1^*(t) + \frac{c_2(-q_2^*(t)\lambda_1 + q_1^*(t)\lambda_2 + sq_2^*(t)\mu_1)\mu_2}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2}\right), \quad t \in [0, \tau^*] \\
p_2^*(t) &= \frac{c_2(q_2^*(t)\mu_1 + q_1^*(t)\mu_2)}{-\lambda_2\mu_1 - \lambda_1\mu_2 + s\mu_1\mu_2}, \quad t \in [0, \tau^*].
\end{aligned} \tag{A.25}$$

The switching curve is then given by

$$\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t), \quad t \in [0, \tau^*],$$

where $p^*(t)$ is calculated explicitly in (A.25).

In addition, it follows from (A.23) that at all regular points of $p_i^*(t)$ where $p_i^*(t)$ is differentiable with respect to t, $\eta_i^*(t) = \dot{p}_i^*(t) + c_i$, $i = 1, 2$. In this case,

$$\begin{aligned}
\eta_1^* &= \begin{cases} 0, \quad t \in [0, \tau_1] \\ c_1 - c_2\mu_2/\mu_1, \quad t \in [\tau_1, \tau^*] \end{cases} \\
\eta_2^* &= 0, \quad t \in [0, \tau^*].
\end{aligned}$$

Lastly, we can infer from Transversality condition (T) and Complementarity condition (C) that

$$\begin{aligned}
\xi_1^*(t) &= \mu_1 p_1^*(t), \quad t \in [0, \tau^*] \\
\xi_2^*(t) &= 0, \quad t \in [0, \tau^*] \\
\xi_3^*(t) &= \begin{cases} \mu_1 p_1^*(t) - \mu_2 p_2^*(t), \quad t \in [0, \tau_1] \\ 0, \quad t \in [\tau_1, \tau^*]. \end{cases}
\end{aligned}$$

We comment that similar analysis to delineate the dual vectors is not replicable for the general system with both abandonment and class-transition. We shall illustrate the difficulty for a simplified system with one-way class-transition, namely, $\gamma_1 = 0$. Consider the scenario where the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1 ($\gamma_1 = 0$). With the policy curve explicitly characterized in Proposition 4, one can potentially calculate the value function (by calculating the optimal state trajectory starting from any state) and derive the dual vectors as above. However, due to the intertwined system dynamics introduced by class-transition, we have not found a way to fully characterize the optimal state trajectory analytically, particularly in the segment where strict priority is given to Class 1. In the other scenario where the $c\mu$-rule prioritizes Class 1 and the modified $c\mu/\theta$-rule prioritizes Class 2 ($\gamma_1 = 0$), the analysis is hindered by not being able to characterize the policy curve as well as the optimal state trajectory.

A.5 MDP Solutions in Section 1.4.4

In this section, we provide details about how we solve the transient scheduling problem (S2) to derive the MDP policy in Figure 1.14. In addition, we elaborate on the initialization for the simulation experiments in Table 1.1.

We use the uniformization approach with truncation to solve the MDP (S2). Let $\Lambda := \lambda_1 + \lambda_2 + (\mu_1 + \mu_2)s + (\theta_1 + \theta_2 + \gamma_1 + \gamma_2)X_{max}$, where $s = 3$ in the small system we consider, and the maximum number in system after truncation is $X_{max} = 40$. To truncate the infinite state space Markov process, the transition rates are modified such that the number-in-system does not exceed X_{max} for each class. In our setting, if $X_1 = 40$, the arrival rate to Class 1 is set to $\lambda_1 = 0$, and the deterioration rate from Class 2 to Class 1 is set to $\gamma_2 = 0$. Similar treatment is applied to Class 2 when $X_2 = 40$.

Define the set of feasible server allocations as

$$\mathscr{Z}(X_1, X_2) := \{(Z_1, Z_2) \in \mathbb{Z}_+ \times \mathbb{Z}_+ : Z_1 \leq X_1, Z_2 \leq X_2, Z_1 + Z_2 \leq s\}.$$

The bellman operator for the MDP takes the form

$$
\begin{aligned}
\Xi(X_1,X_2) = \frac{1}{\Lambda}\Bigg[& c_1(X_1-Z_1)+c_2(X_2-Z_2) \\
& + \min_{(Z_1,Z_2)\in\mathscr{X}(X_1,X_2)} \Bigg\{ \lambda_1\Xi(X_1+1,X_2)+\lambda_2\Xi(X_1,X_2+1) \\
& + (Z_1\mu_1+\theta_1(X_1-Z_1))\Xi(X_1-1,X_2)+(Z_2\mu_2+\theta_2(X_2-Z_2))\Xi(X_1,X_2-1) \\
& + \gamma_1(X_1-Z_1)\Xi(X_1-1,X_2+1)+\gamma_2(X_2-Z_2)\Xi(X_1+1,X_2-1) \\
& + (\Lambda-\lambda_1-\lambda_2-Z_1\mu_1-\theta_1(X_1-Z_1)-Z_2\mu_2-\theta_2(X_2-Z_2)-\gamma_1(X_1-Z_1) \\
& - \gamma_2(X_2-Z_2))\Xi(X_1,X_2)\Bigg\}\Bigg] \quad \text{if } X_1+X_2>s,
\end{aligned}
$$

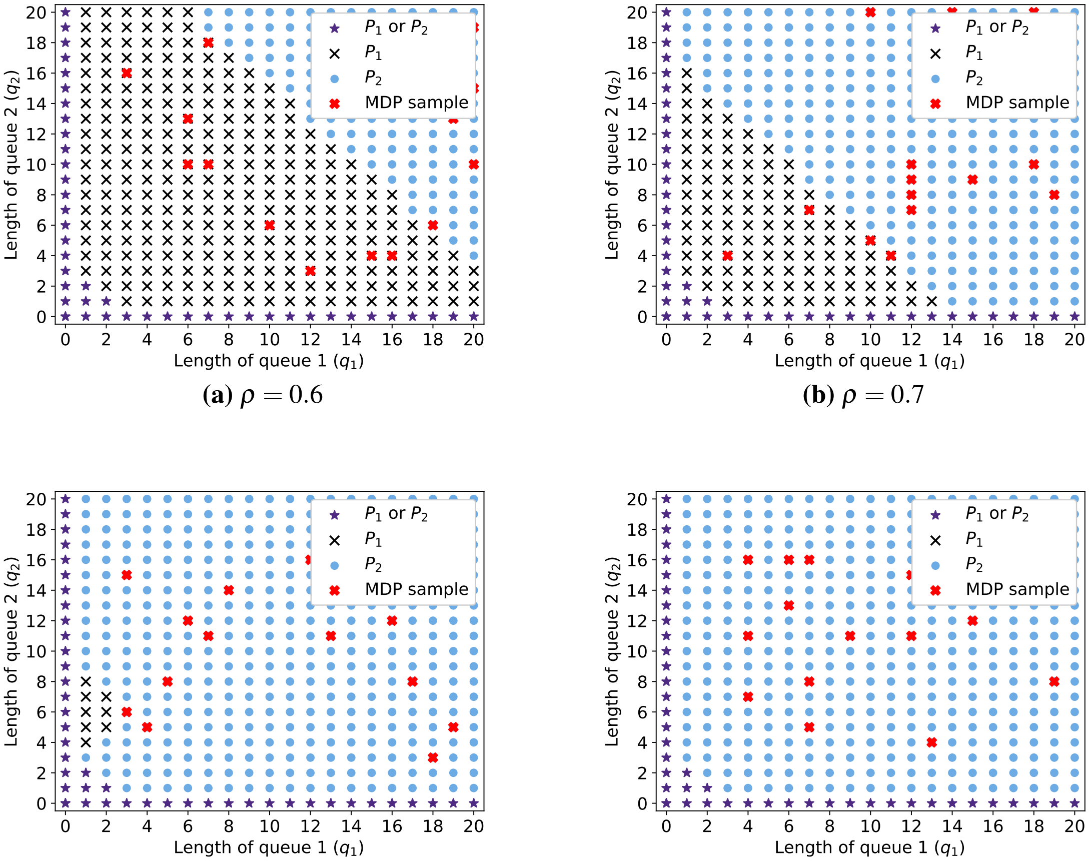
and

$$\Xi(X_1,X_2)=0 \quad \text{if } X_1+X_2\le s. \tag{A.26}$$

Note that (A.26) reflects the terminal cost 0 when the system reaches 0 queue (absorbing states) in the transient control problem (S2).

In Table 1.1, when simulating the system dynamics under different policies, we select $J=15$ initial conditions by sampling X_1 and X_2 independently and uniformly from 3 to 20. Since the small system in consideration has 3 servers, the lower bound is set so that there is positive queue at initialization under any server allocation. Figure A.5 illustrates the selected initial points as red crosses.

Figure A.5: Initialization (red crosses) for the simulation in Table 1.1 and the corresponding optimal MDP solutions



(a) $\rho = 0.6$

(b) $\rho = 0.7$

(c) $\rho = 0.8$

(d) $\rho = 0.9$

Appendix B: Appendix for Chapter 2

B.1 Explicit Representation of the SARIMA and ARIMAX Models

To express the SARIMA model explicitly, we let B be the backward shift operator, where

$$B^j y_t = y_{t-j}, \quad j = 0, \pm 1, \cdots .$$

In the equation above and hereafter, the subscript t is a time index for each shift. We define the related operators

$$\begin{aligned}
\phi(B) &= 1 - \phi_1 B - \phi_1 B^2 - \cdots - \phi_p B^p \\
\Phi(B) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \\
\theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \\
\Theta(B) &= 1 + \Theta_1 B + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \\
\Delta &= 1 - B \\
\Delta_s &= 1 - B^s,
\end{aligned}$$

where $\phi(B)$ is the non-seasonal AR polynomial, $\Phi(B)$ is the seasonal AR polynomial, $\theta(B)$ is the non-seasonal MA polynomial, $\Theta(B)$ is the seasonal MA polynomial, Δ is the non-seasonal difference operator, and Δ_s is the seasonal difference operator. A SARIMA(p,d,q)(P,D,Q)s model can be formally written as

$$\phi(B)\Phi(B)\Delta^d \Delta_s^D y_t = \theta(B)\Theta(B)\varepsilon_t,$$

where ε_t is a noise term that follows a normal distribution with mean 0 and standard deviation σ.

The ARIMAX(p,d,q) model combines the SARIMA(p,d,q)(P,D,Q)s model (where the seasonal hyperparameters (P, D, Q, s) are set to 0) and a linear regression model with external regressors. Let x_t be the vector of covariates in the linear regression model, and x_t^T be its transpose. Let β be the vector of coefficients for the covariates. Then a ARIMAX(p,d,q) model can be formally represented as

$$\phi(B)\Delta^d y_t = x_t^T \beta + \theta(B)\varepsilon_t.$$

For our selected ARIMAX(1,1,4) model, the above representation reduces to

$$\begin{aligned} y_t &= (x_t^*)^T \beta^* + (1+\phi_1^*)y_{t-1} - \phi_1^* y_{t-2} + \varepsilon_t^* + \theta_1^* \varepsilon_{t-1}^* + \theta_2^* \varepsilon_{t-2}^* + \theta_3^* \varepsilon_{t-3}^* + \theta_4^* \varepsilon_{t-4}^* \\ &= (x_t^*)^T \beta^* + 0.0128 y_{t-1} + 0.9872 y_{t-2} + \varepsilon_t^* + \theta_1^* \varepsilon_{t-1}^* + \theta_2^* \varepsilon_{t-2}^* + \theta_3^* \varepsilon_{t-3}^* + \theta_4^* \varepsilon_{t-4}^*, \end{aligned}$$

where x_t^* is the vector of covariates in the embedded linear regression model, and β^* is the associated vector of estimated coefficients, whose value, together with the other estimated parameters denoted with an asterisk in the superscript, is provided in Table 3. Note that y_{t-1} is the arrival count during the previous shift, and y_{t-2} is the arrival count the shift before the previous shift. Their estimated coefficients suggest that y_{t-1} and y_{t-2} are positively correlated with y_t, the arrival count during the current shift. Specifically, the higher patient count was during the previous two shifts, the more likely that the current shift sees a larger patient volume.

Appendix C: Appendix for Chapter 3

C.1 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

$$\mathscr{C}_\pi(\ell) := c_1 N_1 + c_2 N_2(N_1,\ell) + (h + a\gamma)\mathbb{E}\left[Q(N_1 + N_2(N_1,\ell),\ell)\right]. \tag{C.1}$$

We also write

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

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

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 6. *For the multi-server queue with abandonment,*

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

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

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

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

from which (C.2) follows.

Case 2: $\mu < \gamma$**.** 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

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

We can apply the same arguments as in Case 1 to the auxiliary system, and infer (C.2).

Case 3: $\mu>\gamma$. 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

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

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,

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

We can apply the same arguments as in Case 1 to the auxiliary system, and (C.2) follows.

□

Appendices C.2–C.6 contain the proofs of the main results. In Appendix C.2, we prove Proposition 5 which specifies the nontrivial cost parameter regime for the staffing problem. In Appendix C.3, 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.3.1 (for $\alpha > 1/2$) and C.3.2 (for $\alpha \leq 1/2$). In Appendix C.3.3, 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 5) follows (see Appendix C.3.4). In Appendix C.4, we prove Theorem 4 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 C.5, we prove Theorem 6, 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 C.6, we analyze the two-stage staffing problem with surge-stage prediction errors. The results for small prediction errors (Proposition 6) are proved in Appendix C.6.1 and the results for moderate to large prediction errors (Proposition 7) are proved in Appendix C.6.2

C.2 Proof of Proposition 5

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

$$B_1(N_1,N_2(N_1,\ell),\ell) \leq N_1 \quad \text{and} \quad B_2(N_1,N_2(N_1,\ell),\ell) \leq N_2(N_1,\ell). \tag{C.3}$$

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, (C.3) holds regardless of the assignment policy.

Proof of (I). Following (C.1), the total cost satisfies

$$\begin{aligned}
&\mathscr{C}_\pi(\ell) = c_1N_1 + c_2N_2(N_1,\ell) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\
&\geq c_1\mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + c_2\mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] \\
&\quad + \left(\frac{h\mu}{\gamma}+a\mu\right)\frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\
&\geq \min\left\{c_1,c_2,\frac{h\mu}{\gamma}+a\mu\right\}\Bigg(\mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + \mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] \\
&\quad + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right]\Bigg) \\
&= \left(\frac{h\mu}{\gamma}+a\mu\right)\frac{\ell}{\mu} \\
&= \left(\frac{h}{\gamma}+a\right)\ell,
\end{aligned} \tag{C.4}$$

where the second to last equality in (C.4) follows from the steady-state balance equation:

$$\begin{aligned}
\ell &= \mu\mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + \mu\mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \gamma\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right] \\
\frac{\ell}{\mu} &= \mathbb{E}\left[B_1(N_1,N_2(N_1,\ell),\ell)\right] + \mathbb{E}\left[B_2(N_1,N_2(N_1,\ell),\ell)\right] + \frac{\gamma}{\mu}\mathbb{E}\left[Q(N_1+N_2(N_1,\ell),\ell)\right].
\end{aligned} \tag{C.5}$$

Moreover, the cost lower bound in (C.4) 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

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

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 $\mathscr{C}_{\pi_0}(\ell) = (h+a\gamma)\,\ell/\gamma$. Hence, π_0 achieves the cost lower bound, and is optimal to the optimization problem (3.2).

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 (C.1) that $\mathscr{C}_{\pi'}(\Lambda) \le \mathscr{C}_{\pi}(\Lambda)$. 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 (C.1), the total cost satisfies

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

where the last inequality follows by observing from (C.5) 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 $N_2^*(N_1, \Lambda) = 0$. □

C.3 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.3–C.5, 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., $\mathscr{C}_u^\lambda$) 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

$$\hat{\mathscr{C}}_u^\lambda(\Lambda) := \frac{\mathscr{C}_u^\lambda(\Lambda) - c_1\lambda/\mu}{(\lambda/\mu)^{\max\{\alpha, 1/2\}}}, \quad \text{and} \quad \hat{\mathscr{C}}_u^\lambda := \mathbb{E}\left[\hat{\mathscr{C}}_u^\lambda(\Lambda)\right]. \tag{C.6}$$

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

$$\mathscr{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda) := c_2 N_2^\lambda(N_1^\lambda, \ell^\lambda) + (h + a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right]. \tag{C.7}$$

Replacing the realized arrival rate ℓ^λ with Λ^λ in (C.7), we define

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

where the expectation operator on the right-hand side is with respect to the queue process. Note that $\mathscr{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)$ is a constant while $\mathscr{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)$ 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 5 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 3.1.3 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

$$|D_2^\lambda(N_1^\lambda, \Lambda^\lambda)|/(\lambda/\mu)^{\max\{\alpha,1/2\}} < Y \quad \text{for all } \lambda > 0. \tag{C.8}$$

We remark that (C.8) 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.3.1 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

$$\begin{aligned}
&N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda)\\
&= \lambda/\mu + (\beta_1 + \beta_2(\beta_1, x))\,(\lambda/\mu)^\alpha + o((\lambda/\mu)^\alpha)\\
&= \frac{\lambda + x\lambda^\alpha\mu^{1-\alpha}}{\mu} + \frac{(\lambda/\mu)^\alpha\,(\beta_1 + \beta_2(\beta_1, x) - x)}{\left(\lambda/\mu + (\lambda/\mu)^\alpha x\right)^\alpha}\left(\frac{\lambda + \lambda^\alpha\mu^{1-\alpha}x}{\mu}\right)^\alpha + o((\lambda/\mu)^\alpha)\\
&= \ell^\lambda/\mu + (\beta_1 + \beta_2(\beta_1, x) - x)\left(\ell^\lambda/\mu\right)^\alpha + o((\ell^\lambda/\mu)^\alpha).
\end{aligned} \tag{C.9}$$

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 $\mathscr{R}^\lambda$. The lemma is adapted from Theorem 4.1 and Theorem 4.2 in Maman (2009).

Lemma 7. *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,

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

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 (C.9), 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 7 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 (C.9), 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,

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

and

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

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. □

Lemma 8. *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*

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

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

$$\hat{r}(\beta_1, \beta_2(\beta_1, x), x) := \begin{cases} c_2\beta_2(\beta_1, x) & \textit{if } \beta_1 + \beta_2(\beta_1, x) \geq x \\ c_2\beta_2(\beta_1, x) + (h\mu/\gamma + a\mu)(x - \beta_1 - \beta_2(\beta_1, x)) & \textit{if } \beta_1 + \beta_2(\beta_1, x) < x. \end{cases} \tag{C.10}$$

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

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

By Lemma 7 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,

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

Therefore,

$$\lim_{\lambda\to\infty}\frac{1}{\sqrt{\lambda/\mu}}\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] = 0. \tag{C.11}$$

(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,

$$\lim_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^\alpha}\mathbb{E}\left[Q(N_1^\lambda, N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] = \frac{\mu}{\gamma}(x - \beta_1 - \beta_2(\beta_1, x)). \tag{C.12}$$

Lastly, when $\beta_1+\beta_2(\beta_1,x)=x$, we get from Lemma 6 that

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

Then,

$$\lim_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^\alpha}\mathbb{E}\left[Q(N_1^\lambda+N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]=0. \tag{C.13}$$

The statement of the lemma then follows from (C.11), (C.12), and (C.13). □

Based on Lemma 8, let β_1^* and $\beta_2^*(\beta_1,X)$ be the optimal solution to

$$\min_{\beta_1\in\mathbb{R}}\left\{c_1\beta_1+\mathbb{E}\left[\min_{\beta_2(\beta_1,X)\in\mathbb{R}_+}\hat{r}(\beta_1,\beta_2(\beta_1,X),X)\right]\right\},\qquad \text{for } \hat{z} \text{ defined in (C.10).}$$

It is straightforward to derive that

$$\beta_1^*=\arg\min_{\beta\in\mathbb{R}} c_1\beta+c_2\mathbb{E}\left[(X-\beta)^+\right]=\bar{F}_X^{-1}(c_1/c_2),\quad \text{and}\quad \beta_2^*(\beta_1,X)=(X-\beta_1)^+. \tag{C.14}$$

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 (C.14). Note that $u_{2,UH}$ is exactly the two-stage newsvendor solution in Definition 2.

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

Lemma 9. *Assume that $\alpha>1/2$. Under the two-stage uncertainty hedging rule defined in* (C.14) *(equivalently, the two-stage newsvendor solution), we have*

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

for $\hat{z}$ defined in (C.10).

PROOF: It follows from Lemma 8 that

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

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

$$\lim_{\lambda\to\infty}\mathbb{E}\left[\hat{\mathscr{C}}^{\lambda}(\Lambda^{\lambda})\right]=\mathbb{E}\left[\lim_{\lambda\to\infty}\hat{\mathscr{C}}^{\lambda}(\Lambda^{\lambda})\right] \tag{C.15}$$

To this end, we utilize the dominated convergence theorem.

Note that

$$\begin{aligned}\hat{\mathscr{C}}^{\lambda}(\Lambda^{\lambda}) =& c_1\beta_1^*+c_2\beta_2^*(\beta_1^*,X)+\frac{1}{(\lambda/\mu)^{\alpha}}\left(D_1^{\lambda}+D_2^{\lambda}(N_1^{\lambda},\Lambda^{\lambda})\right)\\ &+\frac{1}{(\lambda/\mu)^{\alpha}}(h+a\gamma)\mathbb{E}\left[Q(N_1^{\lambda}+N_2^{\lambda}(N_1^{\lambda},\Lambda^{\lambda}),\Lambda^{\lambda})|\Lambda^{\lambda}\right].\end{aligned} \tag{C.16}$$

For the first two terms on the right-hand side of (C.16), 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 (C.16), note that D_1^{λ} is a constant that is $o((\lambda/\mu)^{\alpha})$. This, together with (C.8), 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 (C.16), we utilize Lemma 6 to get that

$$\begin{aligned}&\mathbb{E}\left[Q(N_1^{\lambda}+N_2^{\lambda}(N_1^{\lambda},\Lambda^{\lambda}),\Lambda^{\lambda})|\Lambda^{\lambda}\right]\\ &\le\max\{\mu/\gamma,1\}\left(\left(\Lambda^{\lambda}/\mu-N_1^{\lambda}-N_2^{\lambda}(N_1^{\lambda},\Lambda^{\lambda})\right)^+ +\sqrt{4\pi/\mu}\sqrt{\Lambda^{\lambda}}+1/\log 2\right)\\ &\le\max\{\mu/\gamma,1\}\left(\left(\Lambda^{\lambda}/\mu-N_1^{\lambda}\right)^+ +\sqrt{4\pi/\mu}\sqrt{\Lambda^{\lambda}}+1/\log 2\right)\\ &=\max\{\mu/\gamma,1\}\left(\left((X-\beta_1^*)\,(\lambda/\mu)^{\alpha}-D_1^{\lambda}\right)^+ +\sqrt{4\pi/\mu}\sqrt{\lambda/\mu+X\lambda^{\alpha}\mu^{1-\alpha}}+1/\log 2\right)\\ &\le\max\{\mu/\gamma,1\}\left((|X|+|\beta_1^*|)\,(\lambda/\mu)^{\alpha}+|D_1^{\lambda}|+\sqrt{4\pi/\mu}\sqrt{\lambda/\mu}+\sqrt{4\pi/\mu}\sqrt{|X|\lambda^{\alpha}\mu^{1-\alpha}}+1/\log 2\right).\end{aligned} \tag{C.17}$$

In (C.17), $D_1^{\lambda}=o((\lambda/\mu)^{\alpha})$ 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, $\mathbb{E}\left[\sqrt{|X|}\right] \leq \sqrt{\mathbb{E}\left[|X|\right]} < \infty$. Therefore, there exists some random variable Y with $\mathbb{E}[Y] < \infty$, such that

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

Therefore, $|\hat{\mathscr{C}}^{\lambda}(\Lambda^{\lambda})|$ in (C.16) is uniformly bounded by an integrable random variable, and (C.15) is justified. □

C.3.2 Two-Stage Uncertainty Hedging Rule for $\alpha \leq 1/2$

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 10. *Assume that $\alpha \leq 1/2$. 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*

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

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

$$\hat{r}(\beta_1,\beta_2(\beta_1,x),x) := c_2\beta_2(\beta_1,x) + \\ \left(\frac{h\mu}{\gamma}+a\mu\right)\frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\left(\beta_1+\beta_2(\beta_1,x)-x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)-\left(\beta_1+\beta_2(\beta_1,x)-x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left(\left(\beta_1+\beta_2(\beta_1,x)-x\mathbb{1}_{\{\alpha=1/2\}}\right)\sqrt{\frac{\mu}{\gamma}}\right)}{H\left(-\left(\beta_1+\beta_2(\beta_1,x)-x\mathbb{1}_{\{\alpha=1/2\}}\right)\right)}}. \tag{C.18}$$

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

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

From the steady-state flow balance equation

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

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. □

Based on Lemma 10, let β_1^* and $\beta_2^*(\beta_1, X)$ be the optimal solution to

$$
\min_{\beta_1 \in \mathbb{R}} \left\{c_1\beta_1 + \mathbb{E}\left[\min_{\beta_2(\beta_1, X) \in \mathbb{R}_+} \hat{r}(\beta_1, \beta_2(\beta_1, X), X)\right]\right\}, \qquad \text{for } \hat{z} \text{ defined in (C.18).} \qquad \text{(C.19)}
$$

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 6. *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* (C.18) *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* (C.19)*. 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 5), 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 11. *Assume that $\alpha \le 1/2$. Under the two-stage uncertainty hedging rule defined in* (C.19)*, 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 (C.18) as

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

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

$$\tilde{\beta}^* := \operatorname*{arg\,min}_{\tilde{\beta} \geq \beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}} \quad c_2\tilde{\beta} + g(\tilde{\beta}). \tag{C.20}$$

Note that by construction, we have

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

Corresponding to (C.20), let

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

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 6.

We discuss the following cases:

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

$$\beta_2^*(\beta_1, x) = \beta^\dagger - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}. \tag{C.21}$$

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

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

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

$$\beta_2^*(\beta_1, x) \leq M - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}. \tag{C.22}$$

Case 2(i.b): $\tilde{\beta}^* > M$. In this case, (C.20) can be rewritten as

$$\tilde{\beta}^* = \operatorname*{arg\,min}_{\tilde{\beta} \geq M} \quad c_2\tilde{\beta} + g(\tilde{\beta}).$$

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

$$c_2 M + g(M) < c_2 y + g(y). \tag{C.23}$$

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

$$\beta_2^*(\beta_1, x) \leq 2M - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}}. \tag{C.24}$$

Case 2(ii): If $\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}} > M$, then by definition of M, (C.23) holds for all $y > 2(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}})$. Hence, $\tilde{\beta}^* \leq 2(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}})$, and

$$\beta_2^*(\beta_1, x) \leq 2(\beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}) - \beta_1 + x\mathbb{1}_{\{\alpha=1/2\}} = \beta_1 - x\mathbb{1}_{\{\alpha=1/2\}}. \tag{C.25}$$

In summary, by (C.21), (C.22), (C.24), and (C.25), we get that

$$\beta_2^*(\beta_1, x) \leq |\beta^\dagger| + 2M + |\beta_1| + |x|. \tag{C.26}$$

Let $\tilde{X} := |\beta^\dagger| + 2M + |\beta_1| + |X|$. The statement follows from (C.26) and $\mathbb{E}[|X|] < \infty$. □

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

Lemma 12. *Assume that $\alpha \leq 1/2$. Under the two-stage uncertainty hedging rule defined in* (C.19)*, we have*

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

for $\hat{z}$ defined in (C.18).

PROOF: It follows from Lemma 10 that

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

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

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\hat{\mathscr{C}}^\lambda(\Lambda^\lambda)\right] = \mathbb{E}\left[\lim_{\lambda\to\infty} \hat{\mathscr{C}}^\lambda(\Lambda^\lambda)\right] \tag{C.27}$$

To this end, we utilize the dominated convergence theorem.

We start by writing

$$\begin{aligned}\hat{\mathscr{C}}^{\lambda}(\Lambda^{\lambda}) =& c_1\beta_1^* + c_2\beta_2^*(\beta_1^*, X) + \frac{1}{\sqrt{\lambda/\mu}}\left(D_1^{\lambda} + D_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda})\right) \\ &+ \frac{1}{\sqrt{\lambda/\mu}}(h + a\gamma)\mathbb{E}\left[Q(N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda})|\Lambda^{\lambda}\right] \\ =& c_1\beta_1^* + c_2\beta_2^*(\beta_1^*, X) + \frac{1}{\sqrt{\lambda/\mu}}\left(D_1^{\lambda} + D_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda})\right) \\ &+ \frac{\Lambda^{\lambda}}{\sqrt{\lambda/\mu}}\,(h/\gamma + a)\,\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right),\end{aligned} \tag{C.28}$$

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

$$\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) \le \mathbb{P}\left(BL, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right). \tag{C.29}$$

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

$$\mathbb{P}\left(BL, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) \le \mathbb{P}\left(BL, N_1^{\lambda}, \lambda + |X|\lambda^{\alpha}\mu^{1-\alpha}\right). \tag{C.30}$$

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),

$$L(N_1^{\lambda}, \lambda + |X|\lambda^{\alpha}\mu^{1-\alpha}) \le L(N_1^{\lambda} - 1, \lambda) + L(1, |X|\lambda^{\alpha}\mu^{1-\alpha}). \tag{C.31}$$

Combining (C.29)–(C.31), we have

$$\begin{aligned}\Lambda^{\lambda}\mathbb{P}\left(AB, N_1^{\lambda} + N_2^{\lambda}(N_1^{\lambda}, \Lambda^{\lambda}), \Lambda^{\lambda}\right) &\le \Lambda^{\lambda}\mathbb{P}\left(BL, N_1^{\lambda}, \lambda + |X|\lambda^{\alpha}\mu^{1-\alpha}\right) \\ &\le \lambda\mathbb{P}\left(BL, N_1^{\lambda} - 1, \lambda\right) + |X|\lambda^{\alpha}\mu^{1-\alpha}\mathbb{P}\left(BL, 1, |X|\lambda^{\alpha}\mu^{1-\alpha}\right) \\ &\le \lambda\mathbb{P}\left(BL, N_1^{\lambda} - 1, \lambda\right) + |X|\lambda^{\alpha}\mu^{1-\alpha}.\end{aligned} \tag{C.32}$$

Dividing both sides of (C.32) by $\sqrt{\lambda/\mu}$, we get that

$$\frac{\Lambda^\lambda}{\sqrt{\lambda/\mu}}\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda\right) \leq \frac{\lambda}{\sqrt{\lambda/\mu}}\mathbb{P}\left(BL, N_1^\lambda - 1, \lambda\right) + |X|\frac{\lambda^\alpha \mu^{1-\alpha}}{\sqrt{\lambda/\mu}}, \tag{C.33}$$

where the first term on the right-hand side of (C.33) is a constant. By equation (17) in Whitt (1984),

$$\lim_{\lambda\to\infty} \frac{\lambda}{\sqrt{\lambda/\mu}}\,\mathbb{P}\left(BL, N_1^\lambda - 1, \lambda\right) = \mu\frac{\phi(\beta_1^*)}{\Phi(\beta_1^*)}. \tag{C.34}$$

Furthermore,

$$\lim_{\lambda\to\infty} |X|\frac{\lambda^\alpha \mu^{1-\alpha}}{\sqrt{\lambda/\mu}} = \begin{cases} \mu|X| & \text{if } \alpha = 1/2 \\ 0 & \text{if } \alpha < 1/2. \end{cases} \tag{C.35}$$

By (C.33)–(C.35), we have for λ large enough,

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

This, together with Lemma 11, 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 $|\hat{\mathscr{C}}^\lambda(\Lambda^\lambda)|$ in (C.28) is uniformly bounded by an integrable random variable, and the interchange of limit and expectation in (C.27) is justified.

□

C.3.3 Optimality Gap of $u_{2,UH}$

In Appendices C.3.1 and C.3.2, 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 (C.14), 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 (C.19). Let $\mathscr{C}_{2,UH}^\lambda$ be the expected total cost defined under the two-stage uncertainty hedging rules. Recall that $\mathscr{C}_{2,*}^\lambda$ is the optimal total cost for the two-stage optimization problem (3.2). The next lemma quantifies the optimality gap of the proposed policy to the exact two-stage optimum.

Lemma 13. *For $\alpha \in (0,1)$, we have $\mathscr{C}^{\lambda}_{2,UH} - \mathscr{C}^{\lambda}_{2,*} = o(\lambda^{\max\{1/2,\alpha\}})$.*

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

$$\liminf_{\lambda\to\infty} \hat{\mathscr{C}}^{\lambda}_{u} \geq \lim_{\lambda\to\infty} \hat{\mathscr{C}}^{\lambda}_{2,UH}. \tag{C.36}$$

Note that the limit on the right-hand side of (C.36) is well defined because of Lemma 9 and Lemma 12.

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

$$\liminf_{\lambda\to\infty} \frac{N^{\lambda}_1 - \lambda/\mu}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} > -\infty. \tag{C.37}$$

To see this, for any sequence realized arrival rate ℓ^{λ}, recall from the proof of Proposition 5 that $B_1(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})$ and $B_2(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})$ denote the steady-state number of busy servers among the base and surge staff, respectively. It follows that

$$\begin{aligned}
&\mathbb{E}\left[\mathscr{R}^{\lambda}(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right] \\
&= c_2 N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}) + (h + a\gamma)\,\mathbb{E}\left[Q(N^{\lambda}_1 + N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right] \\
&\geq \frac{c_2}{\mu}\mu\mathbb{E}\left[B_2(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right] + \left(\frac{h}{\gamma} + a\right)\gamma\mathbb{E}\left[Q(N^{\lambda}_1 + N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right] \\
&\geq \min\left\{\frac{c_2}{\mu}, \frac{h}{\gamma} + a\right\}\left(\mu\mathbb{E}\left[B_2(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right] + \gamma\mathbb{E}\left[Q(N^{\lambda}_1 + N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right]\right) \\
&= \min\left\{\frac{c_2}{\mu}, \frac{h}{\gamma} + a\right\}\left(\ell^{\lambda} - \mu\mathbb{E}\left[B_1(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \ell^{\lambda}), \ell^{\lambda})\right]\right) \\
&\geq \min\left\{\frac{c_2}{\mu}, \frac{h}{\gamma} + a\right\}\left(\ell^{\lambda} - \mu N^{\lambda}_1\right) \\
&= c_2\left(\frac{\ell^{\lambda}}{\mu} - N^{\lambda}_1\right).
\end{aligned}$$

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

$$\mathbb{E}\left[\mathscr{R}^{\lambda}(N^{\lambda}_1, N^{\lambda}_2(N^{\lambda}_1, \Lambda^{\lambda}), \Lambda^{\lambda})\right] \geq c_2\left(\frac{\lambda}{\mu} - N^{\lambda}_1\right).$$

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

$$\begin{aligned}\hat{\mathscr{C}}_u^\lambda &= c_1 \frac{N_1^\lambda - \lambda/\mu}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} + \frac{\mathbb{E}\left[\mathscr{R}^\lambda(N_1^\lambda, N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right]}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} \\ &\geq c_1 \frac{N_1^\lambda - \lambda/\mu}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} + c_2 \frac{\lambda/\mu - N_1^\lambda}{(\lambda/\mu)^{\max\{1/2,\alpha\}}} \\ &= (c_2 - c_1)\frac{\lambda/\mu - N_1^\lambda}{(\lambda/\mu)^{\max\{1/2,\alpha\}}}.\end{aligned} \tag{C.38}$$

If (C.37) does not hold, then it follows from (C.38) and Assumption 5 that $\liminf_{\lambda\to\infty}\hat{\mathscr{C}}_u^\lambda = \infty$. For the purpose of characterizing (near-)optimal staffing rules, we assume without loss of generality that $\liminf_{\lambda\to\infty}\hat{\mathscr{C}}_u^\lambda < \infty$.

Now, consider a subsequence of systems indexed by λ_i on which the lim inf in (C.36) is obtained, namely,

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

Along this subsequence,

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

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,

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

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

$$\frac{N_1^{\lambda_{i_j}} - \lambda_{i_j}/\mu}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}} \to \beta_1 \in \mathbb{R} \quad \text{as } \lambda_{i_j}\to\infty. \tag{C.39}$$

It follows from (C.39) that

$$\begin{aligned}\lim_{\lambda_{i_j}\to\infty}\hat{\mathscr{C}}_u^{\lambda_{i_j}} &\geq \lim_{\lambda_{i_j}\to\infty}\frac{c_1\left(N_1^{\lambda_{i_j}}-\lambda_{i_j}/\mu\right)}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}+\liminf_{\lambda_{i_j}\to\infty}\frac{\mathbb{E}\left[\mathscr{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\Lambda^{\lambda_{i_j}}),\Lambda^{\lambda_{i_j}})\right]}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\\ &= c_1\beta_1+\liminf_{\lambda_{i_j}\to\infty}\frac{\mathbb{E}\left[\mathscr{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\Lambda^{\lambda_{i_j}}),\Lambda^{\lambda_{i_j}})\right]}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\\ &\geq c_1\beta_1+\mathbb{E}\left[\liminf_{\lambda_{i_j}\to\infty}\frac{\mathscr{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\Lambda^{\lambda_{i_j}}),\Lambda^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\right],\end{aligned} \tag{C.40}$$

where the last inequality follows from Fatou's lemma.

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

$$\liminf_{\lambda_{i_j}\to\infty}\frac{\mathscr{R}^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},N_2^{\lambda_{i_j}}(N_1^{\lambda_{i_j}},\ell^{\lambda_{i_j}}),\ell^{\lambda_{i_j}})}{\left(\lambda_{i_j}/\mu\right)^{\max\{1/2,\alpha\}}}\geq \hat{r}(\beta_1,\beta_2^*(\beta_1,x),x). \tag{C.41}$$

In (C.41), when $\alpha>1/2$, $\hat{z}$ is defined in (C.10) and $\beta_2^*(\beta_1,X)$ is defined in (C.14). In the other case where $\alpha\leq 1/2$, $\hat{z}$ is defined in (C.18) and $\beta_2^*(\beta_1,X)$ is defined in (C.19). To see that (C.41) 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), (C.41) follows from Lemma 8, Lemma 10, and the definition of $\beta_2^*(\beta_1,x)$.

For case (ii), we have

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

and (C.41) 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 (C.41).

Now, it follows from (C.40) and (C.41) that

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

Furthermore, since β_1^* is constructed such that

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

it follows that

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

where the last equality follows from Lemma 9 and Lemma 12. Since the subsequence indexed by λ_{i_j} is arbitrary, we have established (C.36).

Next, we apply (C.36) to the sequence of exact optimal two-stage staffing rules, i.e., $u_{2,*}$, and get that

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

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

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

Thus,

$$\lim_{\lambda\to\infty} \hat{\mathscr{C}}_{2,*}^{\lambda} = \lim_{\lambda\to\infty} \hat{\mathscr{C}}_{2,UH}^{\lambda}. \tag{C.42}$$

The statement follows from (C.42). □

The following corollary is a direct consequence from the proof of Lemma 13.

Corollary 3. *For $\alpha \in (0,1)$, let β_1^* and $\beta_2^*(\beta_1^*, X)$ be defined in* (C.14) *when $\alpha > 1/2$, and defined in* (C.19) *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*

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

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

Corollary 3 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 13, 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.3.4 Proof of Theorem 5

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 13. □

C.4 Proof of Theorem 4

The proof of Theorem 4 builds on the performance quantification of $u_2(\beta_1, \beta_2(\beta_1, X))$ and $u_{2,UH}$ introduced in Appendix C.3. For the sequence of systems indexed by λ, recall that $\mathscr{C}^{\lambda}_{1,*}$ is the optimal total cost for the single-stage optimization problem (3.4), and $\mathscr{C}^{\lambda}_{2,*}$ is the optimal total cost for the two-stage optimization problem (3.2). We establish Theorem 4 for different values of α.

C.4.1 Benefit of Surge Staffing When $\alpha < 1/2$

Lemma 14. *If $\alpha < 1/2$, then $\mathscr{C}^{\lambda}_{1,*} - \mathscr{C}^{\lambda}_{2,*} = o(\sqrt{\lambda})$.*

PROOF: We start by determining the parameters β_1^* and $\beta_2^*(\beta_1^*, X)$ defined in (C.19) 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 (C.18) becomes

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

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 $\beta_2^*(\beta_1^*,x) := 0$ 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,

$$\mathscr{C}_{2,UH}^{\lambda} - \mathscr{C}_{1,*}^{\lambda} \geq 0 \quad \text{for all } \lambda > 0. \tag{C.43}$$

In addition, we establish in Lemma 13 that

$$\mathscr{C}_{2,UH}^{\lambda} - \mathscr{C}_{2,*}^{\lambda} = o(\sqrt{\lambda}). \tag{C.44}$$

The statement follows from (C.43) and (C.44). □

C.4.2 Benefit of Surge Staffing When $\alpha = 1/2$

Lemma 15. *If $\alpha = 1/2$, then $\mathscr{C}_{1,*}^{\lambda} - \mathscr{C}_{2,*}^{\lambda} = O(\sqrt{\lambda})$.*

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 $\mathscr{C}_{2,\dagger}^{\lambda}$ 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 12 that

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

Since $(\beta_1^{\dagger},\beta_2^{\dagger}(\beta_1^{\dagger},x))$ is not necessarily optimal for the optimization problem in (C.19), we have

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

It then follows from Lemma 12 that

$$\mathscr{C}_{2,\dagger}^{\lambda} - \mathscr{C}_{2,UH}^{\lambda} = O(\sqrt{\lambda}). \tag{C.45}$$

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

$$\mathscr{C}_{2,\dagger}^{\lambda} - \mathscr{C}_{1,*}^{\lambda} = O(\sqrt{\lambda}). \tag{C.46}$$

Lastly, by Lemma 13, we have

$$\mathscr{C}^{\lambda}_{2,UH} - \mathscr{C}^{\lambda}_{2,*} = o(\sqrt{\lambda}). \tag{C.47}$$

The statement follows from (C.45)–(C.47). □

Figure C.1 below illustrates the performance gap between the employed policies in the proof of Lemma 15.

Figure C.1: Cost saving for $\alpha = 1/2$

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

C.4.3 Benefit of Surge Staffing When $\alpha > 1/2$

Lemma 16. *If $\alpha > 1/2$, then $\mathscr{C}^{\lambda}_{1,*} - \mathscr{C}^{\lambda}_{2,*} = \Theta(\lambda^{\alpha})$.*

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

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

Moreover, Lemma 9 establishes that

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

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 9 show that

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

Therefore, if

$$\arg\min_{\beta\in\mathbb{R}} c_1\beta+\left(\frac{h\mu}{\gamma}+a\mu\right)\mathbb{E}\left[(X-\beta)^+\right]>\arg\min_{\beta\in\mathbb{R}} c_1\beta+c_2\mathbb{E}\left[(X-\beta)^+\right], \tag{C.48}$$

then

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

so that

$$\mathscr{C}^{\lambda}_{1,NV}-\mathscr{C}^{\lambda}_{2,NV}=\Theta(\lambda^{\alpha}). \tag{C.49}$$

Note that a sufficient condition for (C.48) 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

$$\mathscr{C}^{\lambda}_{1,NV}-\mathscr{C}^{\lambda}_{1,*}=O(\lambda^{1-\alpha})=o(\sqrt{\lambda}). \tag{C.50}$$

By Lemma 13, we also have

$$\mathscr{C}^{\lambda}_{2,NV}-\mathscr{C}^{\lambda}_{2,*}=o(\lambda^{\alpha}). \tag{C.51}$$

The statement follows from (C.49)–(C.51). □

Figure C.2 below illustrates the performance gap between the employed policies in the proof of Lemma 16.

Figure C.2: 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 4 follows from Lemmas 14–16.

C.5 Proof of Theorem 6

Before we prove Theorem 6, 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

$$N_1^\lambda = \lambda/\mu + \beta_1^*(\lambda/\mu)^\alpha + D_1^\lambda, \quad N_2^\lambda(N_1^\lambda, \Lambda^\lambda) = \beta_2^*(\beta_1^*, X)(\lambda/\mu)^\alpha + D_2^\lambda(N_1^\lambda, \Lambda^\lambda), \tag{C.52}$$

for $D_1^\lambda = o((\lambda/\mu)^\alpha)$, and $D_2^\lambda(N_1^\lambda, \Lambda^\lambda) = o_{UI}((\lambda/\mu)^\alpha)$. Let u be a policy of the form (C.52). 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

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

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

Lemma 17. *If $\mathscr{C}_u^\lambda < \mathscr{C}_{\tilde{u}}^\lambda$, then $\mathscr{C}_{\tilde{u}}^\lambda - \mathscr{C}_u^\lambda = o(\sqrt{\lambda})$.*

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

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

where the second to last equality follows from $\beta_1^* = \bar{F}_X^{-1}(c_1/c_2)$ 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 $X \geq \beta_1^*$. Thus,

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

and $\mathscr{C}_u^\lambda \geq \mathscr{C}_{\tilde{u}}^\lambda$.

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

$$\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X \geq \beta_1^*\}}\right] = \mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X \geq \beta_1^*\}}\right]. \quad \text{(C.53)}$$

In the other case where $X < \beta_1^*$, it follows from (C.11) in the proof of Lemma 8 that

$$\begin{aligned} &\lim_{\lambda \to \infty} \frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X < \beta_1^*\}} | \Lambda^\lambda\right] \\ &= \lim_{\lambda \to \infty} \frac{1}{(\lambda/\mu)^{1/2}} \mathbb{E}\left[Q(\tilde{N}_1^\lambda + \tilde{N}_2^\lambda(\tilde{N}_1^\lambda, \Lambda^\lambda), \Lambda^\lambda)\mathbb{1}_{\{X < \beta_1^*\}} | \Lambda^\lambda\right] = 0. \end{aligned} \quad \text{(C.54)}$$

The above equality and subsequent inequalities involving random variables hold in a path-by-path sense. Furthermore, recall from Lemma 6 that

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

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

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

Moreover, the same derivation applies to $\tilde{u}$. Thus, we can apply the dominated convergence theorem to (C.54) and get that

$$\begin{aligned}
&\lim_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\mathbb{E}\left[Q(N_1^\lambda+N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right]\\
&=\frac{1}{(\lambda/\mu)^{1/2}}\mathbb{E}\left[Q(\tilde{N}_1^\lambda+\tilde{N}_2^\lambda(\tilde{N}_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] \qquad\qquad (C.55)\\
&=0.
\end{aligned}$$

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

$$\begin{aligned}
\mathscr{C}_u^\lambda &= \mathscr{S}_u^\lambda+(h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda+N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right]\\
&= \mathscr{S}_u^\lambda+(h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda+N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] \qquad (C.56)\\
&\quad +(h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda+N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\mathbb{1}_{\{X\geq\beta_1^*\}}\right].
\end{aligned}$$

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

$$\begin{aligned}
\mathscr{C}_{\tilde{u}}^\lambda &= \mathscr{S}_{\tilde{u}}^\lambda+(h+a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda+\tilde{N}_2^\lambda(\tilde{N}_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right]\\
&= \mathscr{S}_{\tilde{u}}^\lambda+(h+a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda+\tilde{N}_2^\lambda(\tilde{N}_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\mathbb{1}_{\{X<\beta_1^*\}}\right] \qquad (C.57)\\
&\quad +(h+a\gamma)\,\mathbb{E}\left[Q(\tilde{N}_1^\lambda+\tilde{N}_2^\lambda(\tilde{N}_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\mathbb{1}_{\{X\geq\beta_1^*\}}\right].
\end{aligned}$$

Then,

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

where the first equality follows from (C.53), (C.56) and (C.57), and the second equality follows from (C.55). □

Recall from Section 3.4.2 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

$$\bar{\mathscr{C}}_u^\lambda:=\frac{1}{(\lambda/\mu)^{1/2}}\left(\mathscr{C}_u^\lambda-c_1\frac{\lambda}{\mu}-c_1\beta_1^*\left(\frac{\lambda}{\mu}\right)^\alpha-c_2\mathbb{E}\left[(X-\beta_1^*)^+\right]\left(\frac{\lambda}{\mu}\right)^\alpha\right). \qquad (C.58)$$

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

$$\psi(x) := \begin{cases} 0 & \text{if } x < \beta_1^* \\ c_2\eta^* + \left(\frac{h\mu}{\gamma} + a\mu\right) \frac{\sqrt{\frac{\gamma}{\mu}}\left[H\left(\eta^*\sqrt{\frac{\mu}{\gamma}}\right) - \eta^*\sqrt{\frac{\mu}{\gamma}}\right]}{1+\sqrt{\frac{\gamma}{\mu}}\frac{H\left(\eta^*\sqrt{\frac{\mu}{\gamma}}\right)}{H(-\eta^*)}} & \text{if } x \geq \beta_1^*. \end{cases} \tag{C.59}$$

Lemma 18. *We have*

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

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

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{\lambda/\mu}} c_2\left(N_2^\lambda(N_1^\lambda, \Lambda^\lambda) - (X - \beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha + D_1^\lambda \mathbb{1}_{\{X > \beta_1^*\}}\right) = \bar{n}(X), \tag{C.60}$$

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

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\frac{1}{\sqrt{\lambda/\mu}} c_2\left(N_2^\lambda(N_1^\lambda, \Lambda^\lambda) - (X - \beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha + D_1^\lambda \mathbb{1}_{\{X > \beta_1^*\}}\right)\right] = \mathbb{E}\left[\bar{n}(X)\right]. \tag{C.61}$$

To see (C.61), note that when $X < \beta_1^*$,

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

When $X > \beta_1^*$,

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

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 (C.61) can then be justified by (C.60) and the dominated convergence theorem.

For queue length, it follows from (C.11) in the proof of Lemma 8 (for the case where $X < \beta_1^*$), and the same analysis as in the proof of Lemma 11 (for the case where $X > \beta_1^*$) that

$$\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}}(h + a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right] = \bar{q}(X), \tag{C.62}$$

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

$$\lim_{\lambda\to\infty} \frac{1}{(\lambda/\mu)^{1/2}}(h + a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)\right] = \mathbb{E}\left[\bar{q}(X)\right]. \tag{C.63}$$

To see (C.63), it follows from Lemma 6 that

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

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 (C.63) is justified by (C.62) and the dominated convergence theorem.

Then, for $\bar{\mathscr{C}}_u^\lambda$ defined in (C.58) and ψ defined in (C.59),

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

from which the statement follows. □

We now present the proof of Theorem 6.

PROOF: [Proof of Theorem 6] It follows from (C.36) in the proof of Lemma 13 that for all $u \in U$,

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

where $\beta_1^* = \bar{F}_X^{-1}(c_1/c_2)$. Thus, for a sequence of policies $u \in U$, we consider $\bar{\mathscr{C}}_u^\lambda$ defined in (C.58). We next show that for all $u \in U$,

$$\liminf_{\lambda\to\infty} \bar{\mathscr{C}}_u^\lambda \geq \lim_{\lambda\to\infty} \bar{\mathscr{C}}_{2,QED}^\lambda, \tag{C.64}$$

where the limit on the right-hand side of (C.64) is rigorously established in Lemma 18. Similar to the proof of Lemma 13, for the purpose of characterizing (near-)optimal staffing rules, we assume without loss of generality that $\limsup_{\lambda\to\infty}\bar{\mathscr{C}}_u^\lambda < \infty$.

First, by Corollary 3, 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 $D_1^\lambda = o((\lambda/\mu)^\alpha)$ and $D_2^\lambda(N_1^\lambda,\Lambda^\lambda) = o_{UI}((\lambda/\mu)^\alpha)$, i.e., the two-stage newsvendor solutions.

In addition, Lemma 17 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

$$\begin{aligned}\bar{\mathscr{C}}_u^\lambda &= \frac{1}{(\lambda/\mu)^{1/2}}\mathbb{E}\Big[c_2N_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right] \\ &\quad - c_2(X-\beta_1^*)^+\left(\frac{\lambda}{\mu}\right)^\alpha\Big] \\ &= \frac{1}{(\lambda/\mu)^{1/2}}\mathbb{E}\left[c_2D_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right]\right].\end{aligned}$$

By Fatou's lemma,

$$\begin{aligned}&\liminf_{\lambda\to\infty}\bar{\mathscr{C}}_u^\lambda \\ &\geq \mathbb{E}\left[\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2D_2^\lambda(N_1^\lambda,\Lambda^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\Lambda^\lambda),\Lambda^\lambda)|\Lambda^\lambda\right]\right)\right].\end{aligned} \tag{C.65}$$

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

$$\liminf_{\lambda\to\infty}\frac{1}{(\lambda/\mu)^{1/2}}\left(c_2D_2^\lambda(N_1^\lambda,\ell^\lambda) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda,\ell^\lambda),\ell^\lambda)\right]\right) \geq \psi(x), \tag{C.66}$$

where ψ is defined in (C.59). 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 $\mathbb{E}\left[Q^\lambda(N_1^\lambda + N_2^\lambda(N_1^\lambda, \ell^\lambda), \ell^\lambda)\right] \geq 0$,

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

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

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

where recall from the proof of Proposition 5 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,

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

For case (iii), it follows from (C.11) in the proof of Lemma 8 (for the case where $x < \beta_1^*$), and the same analysis as in the proof of Lemma 11 (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, (C.66) follows from the definition of η^* in (3.11).

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 (C.66).

So far we have established (C.66). This, together with (C.65) and Lemma 18, gives that

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

which establishes (C.64).

In this last step, note that by (C.64), we have

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

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

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

Therefore,

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

which implies that $\mathscr{C}_{2,QED}^{\lambda} - \mathscr{C}_{2,*}^{\lambda} = o(\sqrt{\lambda})$. □

C.6 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.

C.6.1 Small Prediction Error: Proof of Proposition 6

PROOF: Statement (I) follows exactly the same lines of analysis as the proof of Theorem 4 for $\alpha > 1/2$. Statement (II) follows exactly the same lines of analysis as the proof of Theorem 6. Lastly, following the same lines of analysis as the proof of Theorem 6, we can show that $\mathscr{C}_{2,ERR}^{e,\lambda} - \mathscr{C}_{2,*}^{o,\lambda} = o(\sqrt{\lambda})$. This, together with statement (II), implies statement (III). To elaborate on the generalization, we explain why the proof of Proposition 6 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 (3.11). The rest of the analysis is generalized similarly. □

C.6.2 Moderate to Large Prediction Error: Proof of Proposition 7

PROOF: We first show that there exists an optimal solution to (3.18). In particular, consider the inner-problem in (3.18):

$$\min_{N_2^\lambda(N_1^\lambda,Y)} \left\{ c_2 N_2^\lambda(N_1^\lambda,Y) + (h\mu/\gamma + a\mu)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda,Y)\right)^+ |Y\right]\right\}. \quad \text{(C.67)}$$

Note that (C.67) 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

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

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

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

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 (3.18).

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 4 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 $\hat{\mathscr{C}}_u^\lambda$ in (C.6), we define

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

Similar lines of arguments as in the proof of Lemma 9 establish that

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

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 9 show that

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

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

By Assumption 5 and the continuity of Y, it can be verified that $\lim_{\lambda\to\infty}\hat{\mathscr{C}}_{\tilde{u}}^{e,\lambda} > \lim_{\lambda\to\infty}\hat{\mathscr{C}}_u^{e,\lambda}$. Thus,

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

Moreover, similar derivation as in the proof of Lemma 13 gives that

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

The statement follows.

Case 2: $\nu = \alpha$. 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

$$\min_{\beta_1}\left\{c_1\beta_1 + \mathbb{E}\left[\min_{\beta_2(\beta_1,Y)\in\mathbb{R}_+}\left\{c_2\beta_2(\beta_1,Y) + (h\mu/\gamma + a\mu)\,\mathbb{E}\left[(Y+Z-\beta_1-\beta_2(\beta_1,Y))^+\,\middle|Y\right]\right\}\right]\right\}. \tag{C.68}$$

We first show that an optimal solution to (C.68) exists. Consider the inner-problem in (C.68):

$$\min_{\beta_2(\beta_1,Y)\in\mathbb{R}_+} c_2\beta_2(\beta_1,Y) + (h\mu/\gamma + a\mu)\,\mathbb{E}\left[(Y+Z-\beta_1-\beta_2(\beta_1,Y))^+\,\middle|Y\right]. \tag{C.69}$$

Note that (C.69) 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

$$\beta_2^*(\beta_1, Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + Y - \beta_1\right)^+. \tag{C.70}$$

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

$$\begin{aligned}\frac{\partial}{\partial\beta_1}h(\beta_1) = c_1 &- c_2\mathbb{P}\left(Y > \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + \beta_1\right)\\ &- \left(\frac{h\mu}{\gamma} + a\mu\right)\mathbb{P}\left(Y \le \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + \beta_1, Y + Z > \beta_1\right).\end{aligned} \tag{C.71}$$

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

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

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

Following similar lines of arguments as in the proof of Lemma 9 and Lemma 13, we get that

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

and

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

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

$$\tilde{\beta} := \arg\min_{\beta\in\mathbb{R}} c_1\beta + \left(\frac{h\mu}{\gamma} + a\mu\right)\mathbb{E}\left[(Y + Z - \beta)^+\right] = \bar{F}_{Y+Z}^{-1}\left(\frac{c_1}{h\mu/\gamma + a\mu}\right). \tag{C.73}$$

Similar derivation as in the proof of Lemma 9 gives that

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

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

$$\mathscr{C}_{\tilde{u}}^{e,\lambda} - \mathscr{C}_{1,*}^{e,\lambda} = O(\lambda^{1-\alpha}). \tag{C.74}$$

If Assumption 6 holds, then

$$\beta_2^*(\beta_1^*,Y) = \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu}\right) + Y - \beta_1^*\right)^+ > 0 \quad \text{with probability } p > 0. \tag{C.75}$$

To see (C.75), 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 (C.71) that $\beta_1^* = \tilde{\beta}$, for $\tilde{\beta}$ defined in (C.73). However, plugging in the value of $\tilde{\beta}$ in (C.70) 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 6, implies that $\beta_2^*(\beta_1^*,Y) > 0$ with probability $p > 0$, a contradiction. Thus, (C.75) holds. It follows from (C.75) that $\lim_{\lambda\to\infty}\hat{\mathscr{C}}_{\tilde{u}}^{e,\lambda} > \lim_{\lambda\to\infty}\hat{\mathscr{C}}_u^{e,\lambda}$, so that

$$\mathscr{C}_{\tilde{u}}^{e,\lambda} - \mathscr{C}_u^{e,\lambda} = \Theta(\lambda^\alpha). \tag{C.76}$$

In the other case where Assumption 6 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 (C.68), and

$$\mathscr{C}_{\tilde{u}}^{e,\lambda} - \mathscr{C}_{u}^{e,\lambda} = o(\lambda^{\alpha}). \tag{C.77}$$

The statement follows from (C.72), (C.74), (C.76), and (C.77).

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

Case 1: $\mu = \gamma$. 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

$$\begin{aligned}
&\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \\
&\leq \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y, Z\right] \\
&\leq \sqrt{\frac{4\pi}{\mu}}\sqrt{\Lambda^\lambda}\exp\left(-\frac{\mu}{4\Lambda^\lambda}\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^2\right) \\
&\quad + \left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ + \frac{1}{\log 2}.
\end{aligned} \tag{C.78}$$

Taking expectation of (C.78) conditional on Y gives

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right] \\
&\leq \mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right] \\
&\leq \mathbb{E}\left[\left(\frac{\Lambda^\lambda}{\mu} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right] + \mathbb{E}\left[\sqrt{\frac{4\pi}{\mu}}\sqrt{\Lambda^\lambda}\middle| Y\right] + \frac{1}{\log 2}.
\end{aligned} \tag{C.79}$$

It follows from (C.79) that

$$\begin{aligned}
&c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h+a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \Big| Y\right]\right] \\
&\leq c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h+a\gamma)\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
&\leq c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h+a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \Big| Y\right]\right] \\
&\quad + \mathbb{E}\left[\sqrt{4\pi/\mu}\sqrt{\Lambda^\lambda}\right] + 1/\log 2 \\
&\leq c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h+a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \Big| Y\right]\right] \\
&\quad + \sqrt{4\pi/\mu}\sqrt{\lambda} + \sqrt{4\pi/\mu}\sqrt{\lambda^\alpha \mu^{1-\alpha}\mathbb{E}[|Y|]} + \sqrt{4\pi/\mu}\sqrt{\lambda^\nu \mu^{1-\nu}\mathbb{E}[|Z|]} + 1/\log 2,
\end{aligned} \tag{C.80}$$

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 (3.16). We have

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

where (a) follows from (C.80), (b) follows from the optimality of $(\bar{N}_1, \bar{N}_2(\bar{N}_1, Y))$ to problem (3.18), and (c) follows from (C.80) again.

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

$$\begin{aligned}
\mathscr{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) &:= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h+a\gamma)\,\mathbb{E}\left[Q(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
&= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h/\gamma + a)\,\mathbb{P}\left(AB, N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), Y\right)\right],
\end{aligned} \tag{C.81}$$

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

$$\begin{aligned}
&\bar{\mathscr{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&:= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma\right].
\end{aligned} \tag{C.82}$$

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

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 (C.81) and (C.82) for Sequence II are

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

and

$$\begin{aligned}
&\bar{\mathscr{C}}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&:= c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[\left(\Lambda^\lambda - \mu^{II}\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma^{II}\right] \\
&= c_1 N_1^\lambda + \mathbb{E}\left[c_2 N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \mu\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma\right] \\
&= \bar{\mathscr{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)).
\end{aligned} \tag{C.83}$$

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

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

which implies that

$$\mathscr{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \leq \mathscr{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)). \tag{C.84}$$

Applying (C.80) to Sequence II, we get that

$$\begin{aligned}
&\mathscr{C}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&= c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[Q^{II}(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
&\leq c_1^{II} N_1^\lambda + \mathbb{E}\left[c_2^{II} N_2^\lambda(N_1^\lambda, Y) + (h^{II} + a^{II}\gamma^{II})\mathbb{E}\left[\left(\Lambda^\lambda/\mu^{II} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] \\
&\quad + O(\sqrt{\lambda}) \\
&= \bar{\mathscr{C}}^{e,\lambda,II}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) + O(\sqrt{\lambda})
\end{aligned} \tag{C.85}$$

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 (C.81) and (C.82) for Sequence III are

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

and

$$\begin{aligned}
&\bar{\mathscr{C}}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&:= \mathbb{E}\left[c_2^{III} N_2^\lambda(N_1^\lambda, Y) + (h^{III} + a^{III}\gamma^{III})\mathbb{E}\left[\left(\Lambda^\lambda - \mu^{III}\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma^{III}\right] \\
&\quad + c_1^{III} N_1^\lambda \\
&= c_1\gamma/\mu N_1^\lambda + \mathbb{E}\left[c_2\gamma/\mu N_2^\lambda(N_1^\lambda, Y) + (h + a\gamma)\mathbb{E}\left[\left(\Lambda^\lambda - \gamma\left(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y)\right)\right)^+ |Y\right]/\gamma\right] \\
&= \bar{\mathscr{C}}^{e,\lambda}(\gamma/\mu N_1^\lambda, \gamma/\mu N_2^\lambda(N_1^\lambda, Y)).
\end{aligned} \tag{C.86}$$

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

$$\mathscr{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \geq \mathscr{C}^{e,\lambda,III}(\mu/\gamma N_1^\lambda, \mu/\gamma N_2^\lambda(N_1^\lambda, Y)). \tag{C.87}$$

Applying (C.80) to Sequence III, we get that

$$\begin{aligned}
&\mathscr{C}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&= c_1^{III} N_1^\lambda + \mathbb{E}\left[c_2^{III} N_2^\lambda(N_1^\lambda, Y) + (h^{III} + a^{III}\gamma^{III})\mathbb{E}\left[Q^{III}(N_1^\lambda + N_2^\lambda(N_1^\lambda, Y), \Lambda^\lambda)|Y\right]\right] \\
&\geq c_1^{III} N_1^\lambda + \mathbb{E}\left[c_2^{III} N_2^\lambda(N_1^\lambda, Y) + (h^{III} + a^{III}\gamma^{III})\mathbb{E}\left[\left(\Lambda^\lambda/\mu^{III} - N_1^\lambda - N_2^\lambda(N_1^\lambda, Y)\right)^+ \middle| Y\right]\right] \\
&= \bar{\mathscr{C}}^{e,\lambda,III}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)),
\end{aligned} \tag{C.88}$$

which implies that

$$\begin{aligned}
\mathscr{C}^{e,\lambda}(N_1^{\lambda,*}, N_2^{\lambda,*}(N_1^{\lambda,*}, Y)) &= \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \mathscr{C}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&\overset{(d)}{\geq} \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \mathscr{C}^{e,\lambda,III}(\mu/\gamma N_1^\lambda, \mu/\gamma N_2^\lambda(N_1^\lambda, Y)) \\
&\overset{(e)}{\geq} \min_{N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)} \bar{\mathscr{C}}^{e,\lambda}(N_1^\lambda, N_2^\lambda(N_1^\lambda, Y)) \\
&= \bar{\mathscr{C}}^{e,\lambda}(\bar{N}_1^\lambda, \bar{N}_2^\lambda(\bar{N}_1^\lambda, Y)),
\end{aligned} \tag{C.89}$$

where (d) follows from (C.87), and (e) follows from (C.86) and (C.88).

Lastly, we can write

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

where (f) follows from (C.84), (g) follows from (C.83), and (h) follows from (C.85) and (C.89).

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

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

Applying (C.80) to Sequence II, we get that

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

which implies that

$$\begin{aligned}\mathscr{C}^{e,\lambda}(N_1^{\lambda,*},N_2^{\lambda,*}(N_1^{\lambda,*},Y)) &= \min_{N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)} \mathscr{C}^{e,\lambda}(N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)) \\ &\geq \min_{N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)} \mathscr{C}^{e,\lambda,II}(N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)) \\ &\geq \min_{N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)} \bar{\mathscr{C}}^{e,\lambda}(N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)) \\ &= \bar{\mathscr{C}}^{e,\lambda}(\bar{N}_1^{\lambda},\bar{N}_2^{\lambda}(\bar{N}_1^{\lambda},Y)).\end{aligned} \tag{C.90}$$

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

$$\mathscr{C}^{e,\lambda}(N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)) \leq \mathscr{C}^{e,\lambda,III}(\mu/\gamma N_1^{\lambda},\mu/\gamma N_2^{\lambda}(N_1^{\lambda},Y)). \tag{C.91}$$

Applying (C.80) to Sequence III, we get that

$$\begin{aligned}&\mathscr{C}^{e,\lambda,III}(N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y)) \\ &= c_1^{III}N_1^{\lambda}+\mathbb{E}\left[c_2^{III}N_2^{\lambda}(N_1^{\lambda},Y)+(h^{III}+a^{III}\gamma^{III})\mathbb{E}\left[Q^{III}(N_1^{\lambda}+N_2^{\lambda}(N_1^{\lambda},Y),\Lambda^{\lambda})|Y\right]\right] \\ &\leq c_1^{III}N_1^{\lambda}+\mathbb{E}\left[c_2^{III}N_2^{\lambda}(N_1^{\lambda},Y)+(h^{III}+a^{III}\gamma^{III})\mathbb{E}\left[\left(\Lambda^{\lambda}/\mu^{III}-N_1^{\lambda}-N_2^{\lambda}(N_1^{\lambda},Y)\right)^{+}\middle|Y\right]\right] \\ &\quad +O(\sqrt{\lambda}) \\ &= \bar{\mathscr{C}}^{e,\lambda,III}(N_1^{\lambda},N_2^{\lambda}(N_1^{\lambda},Y))+O(\sqrt{\lambda})\end{aligned} \tag{C.92}$$

Lastly, we can write

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

where (i) follows from (C.91), (j) follows from (C.86), and (k) follows from (C.90) and (C.92).

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

$$
\min_{N_1^\lambda} \left\{ c_1 N_1^\lambda + \mathbb{E}\left[\min_{N_2^\lambda(N_1^\lambda, \Lambda^\lambda)} \left\{ c_2 N_2^\lambda(N_1^\lambda, \Lambda^\lambda) + (h\mu/\gamma + a\mu)\mathbb{E}\left[\left(\Lambda^\lambda/\mu - N_1^\lambda - N_2^\lambda(N_1^\lambda, \Lambda^\lambda) \right)^+ | \Lambda^\lambda \right] \right\} \right] \right\}. \tag{C.93}
$$

whose optimal solution is given by

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

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

$$
\mathscr{C}_{\hat{u}}^{o,\lambda} - \mathscr{C}_{2,*}^{o,\lambda} = O(\sqrt{\lambda}). \tag{C.94}
$$

Recall from the proof of Proposition 7 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

$$\begin{aligned} 0 =&c_1 - c_2 \mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^\alpha Y > \left(\bar{N}_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu} \right) \left(\frac{\lambda}{\mu}\right)^\nu \right) \right) \\ & - \left(\frac{h\mu}{\gamma} + a\mu \right) \mathbb{P}\left(\left(\frac{\lambda}{\mu}\right)^\alpha Y \le \left(\bar{N}_1^\lambda - \frac{\lambda}{\mu} - \bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu} \right) \left(\frac{\lambda}{\mu}\right)^\nu \right), \right. \\ & \left. \left(\frac{\lambda}{\mu}\right)^\alpha Y + \left(\frac{\lambda}{\mu}\right)^\nu Z > \bar{N}_1^\lambda - \frac{\lambda}{\mu} \right). \end{aligned} \tag{C.95}$$

Next, we compare the two inner-optimization problems in (3.18) and (C.93). It holds that

$$\begin{aligned} & \mathbb{E}\left[\min_{N_2^\lambda(\bar{N}_1^\lambda, Y)} \left\{ c_2 N_2^\lambda(\bar{N}_1^\lambda, Y) + \left(\frac{h\mu}{\gamma} + a\mu \right) \mathbb{E}\left[\left(\frac{\Lambda^\lambda}{\mu} - \bar{N}_1^\lambda - N_2^\lambda(\bar{N}_1^\lambda, Y) \right)^+ \middle| Y \right] \right\} \right] \\ & \quad - \mathbb{E}\left[\min_{N_2^\lambda(\bar{N}_1^\lambda, \Lambda^\lambda)} \left\{ c_2 N_2^\lambda(\bar{N}_1^\lambda, \Lambda^\lambda) + \left(\frac{h\mu}{\gamma} + a\mu \right) \left(\frac{\Lambda^\lambda}{\mu} - \bar{N}_1^\lambda - N_2^\lambda(\bar{N}_1^\lambda, \Lambda^\lambda) \right)^+ \right\} \right] \\ & = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[c_2 \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu} \right) \left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda \right)^+ \right. \\ & \quad + \left(\frac{h\mu}{\gamma} + a\mu \right) \left(\frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha + z \left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda \right. \\ & \quad - \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu} \right) \left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda \right)^+ \Bigg)^+ \\ & \quad \left. - c_2 \left(\frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha + z \left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda \right) \right] f_Y(y) f_Z(z) dy dz. \end{aligned} \tag{C.96}$$

Denote part of the integrand in (C.96) as

$$\begin{aligned} g^\lambda(y,z) := & \, c_2 \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu} \right) \left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda \right)^+ \\ & + \left(\frac{h\mu}{\gamma} + a\mu \right) \left(\frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha + z \left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda \right. \\ & - \left(\bar{F}_Z^{-1}\left(\frac{c_2}{h\mu/\gamma + a\mu} \right) \left(\frac{\lambda}{\mu}\right)^\nu + \frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha - \bar{N}_1^\lambda \right)^+ \Bigg)^+ \\ & - c_2 \left(\frac{\lambda}{\mu} + y \left(\frac{\lambda}{\mu}\right)^\alpha + z \left(\frac{\lambda}{\mu}\right)^\nu - \bar{N}_1^\lambda \right)^+ . \end{aligned}$$

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 (C.95) 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

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g^{\lambda}(y,z) f_Y(y) f_Z(z) dy dz = \Theta(\lambda^{\nu}). \tag{C.97}$$

It follows from (C.96), (C.97), and the construction of stochastic-fluid problems (3.18) and (C.93) that

$$\mathscr{C}_{2,ERR}^{e,\lambda} - \mathscr{C}_{\hat{u}}^{o,\lambda} = \Theta(\lambda^{\nu}). \tag{C.98}$$

The statement follows from (C.94), (C.98), and statement (II). □

C.7 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 (3.3). 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, $i \in \mathscr{I}$. Since $L_i | \Lambda_i \sim$Poisson(Λ_i), we have

$$\mathbb{E}[L_i] = \mathbb{E}[\mathbb{E}[L_i|\Lambda_i]] = \lambda_i$$
$$\mathrm{Var}(L_i) = \mathrm{Var}(\mathbb{E}[L_i|\Lambda_i]) + \mathbb{E}[\mathrm{Var}(L_i|\Lambda_i)] = \lambda_i^{2\alpha}\sigma^2 + \lambda_i, \quad i \in \mathscr{I}.$$

Thus,

$$\frac{\mathrm{Std}(L_i)}{\lambda_i^{\alpha}} = \left(\sigma^2 + \lambda_i^{1-2\alpha}\right)^{1/2}, \quad i \in \mathscr{I}.$$

In addition, since $\alpha > 1/2$,

$$\lim_{\lambda \to \infty} (\log \mathrm{Std}(L_i) - \alpha \log \lambda_i) = \log \sigma, \quad i \in \mathscr{I}.$$

Hence, it holds for large λ_i that

$$\log \mathrm{Std}(L_i) \approx \alpha \log \lambda_i + \log \sigma, \quad i \in \mathscr{I}.$$

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

$$\log \Sigma_i \approx \hat{\alpha} \log \bar{L}_i + \log \hat{\sigma}, \quad i \in \mathscr{I},$$

which is equivalent to (3.20) in our parametric estimation setting.

C.8 Supplementary Numerical Experiments

C.8.1 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 (3.12). In what follows, we repeat the experiments in Tables 3.2 (with $c_2 = 2$) and 3.3 (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 5 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 (3.13) for small systems.

Table C.1: System performance under different specifications of the two-stage QED staffing rule with $\beta^* = 0.967, \eta^* = 0.120$
$(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 6)$

k λ	-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%

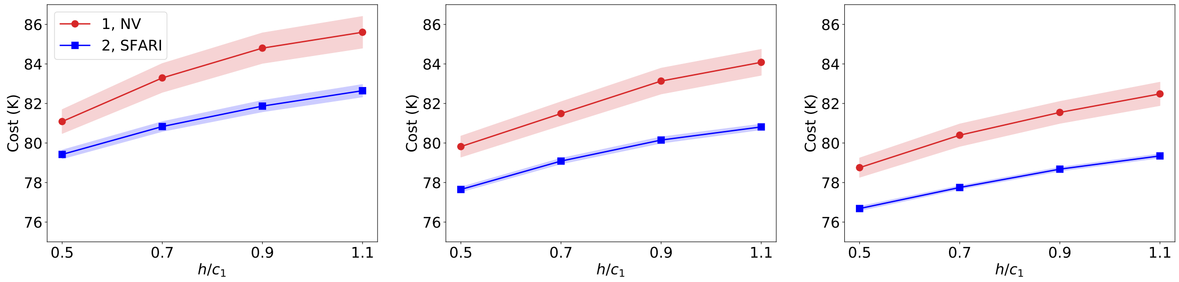
Table C.2: System performance under different specifications of the two-stage QED staffing rule with $\beta^* = 1.465, \eta^* = -0.380$
$(\mu = 1, \gamma = 0.1, \alpha = 0.75, h = 1.5, a = 3, c_1 = 1, c_2 = 14)$

k λ	-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%

C.8.2 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 C.3a 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 3.4), 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 C.3a, Figure C.3b assumes lognormal (as opposed to exponential) service times, and Figure C.3c considers both lognormal service times and hourly-varying arrival rates. We observe that the cost curves in Figures C.3a–C.3c 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 C.3: Impact of service time distribution and non-stationary arrival rate



(a) Theoretical setting

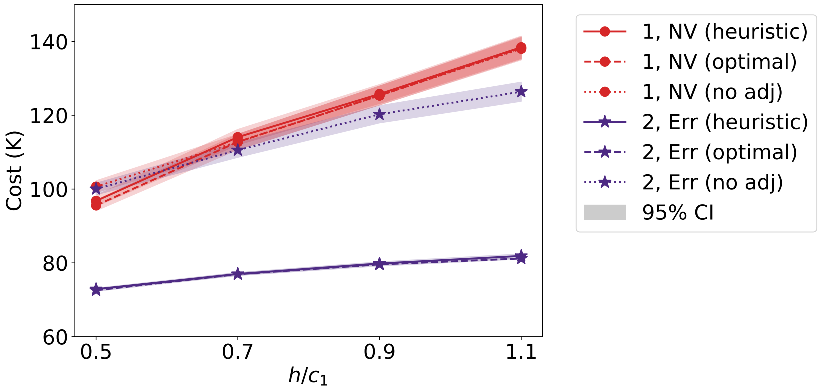
(b) Lognormal service time

(c) Lognormal service time and hourly-varying arrival rate

C.8.3 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 3.7.4.2 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 C.4 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 C.4: Comparison of tradeoff curves under the proposed and optimized adjustment parameters


Cost (K)
h/c_1
1, NV (heuristic)
1, NV (optimal)
1, NV (no adj)
2, Err (heuristic)
2, Err (optimal)
2, Err (no adj)
95% CI