OPTIMIZATION OF STOCHASTIC MODELS IN HEALTH CARE: APPOINTMENT SCHEDULING AND DISEASE TESTING

Nikolai Dabney Lipscomb

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2023

Approved by:

Nilay T. Argon

Vidyadhar G. Kulkarni

Yao Li

Xin Liu

Quoc Tran-Dinh

Serhan Ziya

ABSTRACT

Nikolai Dabney Lipscomb: Optimization of Stochastic Models in Health Care:
Appointment Scheduling and Disease Testing
(Under the direction of Vidyadhar G. Kulkarni and Serhan Ziya)

We consider two different problems: appointment scheduling and asymptomatic disease testing.

For the appointment scheduling problem, the goal is to assign appointment times to minimize a weighted sum of patient wait times, doctor idle time, and clinic overtime. We make the assumption that patients are unpunctual with respect to assigned appointment times and distributional information on unpunctuality is available. We first consider a model with heterogeneous patient distributions in both service time and unpunctuality. This is a complex system that requires heuristic approaches. We are able to show the benefits of capturing patient heterogeneity in addition to the superior performance of our heuristics. Our best methods do not scale well to large patient systems; thus, we consider a second model that allows a large number of patients. For large systems, we assume patient homogeneity; however, patient unpunctuality is permitted to be time-heterogeneous. With this model, we examine the fluid limits of the queue processes to develop a fluid control problem that seeks an asymptotically optimal appointment schedule in the form of an RCLL function. This problem is difficult to solve analytically, so we propose a numerical scheme that converts the control problem into a quadratic program. We examine asymptotically optimal appointment schedules under various unpunctuality distributions, then the superior performance of these schedules in discrete-event simulations.

For the asymptomatic disease testing problem, we consider the individual decision-maker problem of choosing when to use disease test kits from a limited supply. We assume an underlying SIR Markov model with split states for asymptomatic and symptomatic states. As only symptoms are directly observable, the decision process is modeled as a partially-observable Markov decision process for deciding when to use tests. The goal is to produce simple instructions for the average consumer to follow. We derive policies that do not require probability computations by the user.

Under certain assumptions, we are able to prove that these policies are optimal. Last, we examine a community simulation where infection probabilities are dependent on community infected. Our methods are shown to outperform existing baselines.

To William Lowndes Lipscomb and Bernadette Sanchez Lipscomb.

ACKNOWLEDGEMENTS

I would like to thank Dr. Vidyadhar G. Kulkarni and Dr. Serhan Ziya, my advisers at the University of North Carolina at Chapel Hill, for their guidance and patience.

I would like to further thank Dr. Xin Liu of Amazon.com, Inc. and Dr. Alex F. Mills of the Zicklin School of Business at Baruch College for their guidance and contributions to this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Introduction to Appointment Scheduling in Healthcare

Outpatient appointment scheduling is a topic examined since the 1950's beginning with the pioneering work of (Bailey, 1952). As healthcare costs increase and professionals in the industry advocate for more preventative care to help avoid expensive treatments and overnight hospital stays, a greater focus has been placed on outpatient healthcare services, a much cheaper alternative. In order to boost the efficiency of these outpatient clinics, careful attention has been paid to the scheduling approaches that are designed to avoid overloading clinic waiting rooms while similarly seeking to maximize the utilization of the clinics' healthcare providers (Cayirli and Veral, 2003; Deceuninck and Vuyst, 2018). Many healthcare systems often see unacceptable wait times whether due to oversights or inefficient design choices. The importance of reducing wait times have been typically viewed by examining the opportunity and psychological costs associated with waiting (Osuna, 1985; Kocas, 2015); however, in light of the COVID-19 pandemic, the health-related costs of a busy waiting room have become even more pronounced (Thunstrom and Shogren, 2020).

Today's clinics are in an excellent position to make decisions informed by historical data, as online check-ins and electronically-assisted arrival questionnaires provide the means to gather not just questionnaire responses themselves, but the time of patient check-in relative to scheduled appointment time. Such information could be leveraged to examine the unpunctuality behavior (lateness or earliness) of patients for use in scheduling decisions. With a goal of minimizing waiting-room wait times for patients who arrive before they can enter service, unpunctuality has recently become a topic of interest in outpatient appointment scheduling (Ahmadi-Javid and Klassen, 2017). For patients that regularly visit a clinic, more data will naturally be available, making it much more feasible to estimate both individual unpunctuality and service time distributions.

For the first chapter, we first examine a UNC clinic dataset and gather some insights into patient unpunctuality behavior and patient service times. We also provide justification for the idea of individualized patient service and unpunctuality; that is, patients should not be treated as homogeneous with respect to their service times nor their unpunctuality distributions. We then consider the following appointment setup. We have a fixed set of patients that we need to schedule for a given day. These patients are regularly returning patients and the clinic has accrued a slew of historical data used to estimate their unpunctuality distributions; we treat patients are heterogeneous with respect to their individual unpunctuality distributions. Patients may have different types of service scheduled, meaning they are also heterogeneous with respect to their service time distributions. The objective seeks an appointment to minimize a weighted sum of total patient wait time, server idle time, and server overtime. The introduction of individualized distributions and service disciplines different from first-come first-serve leads to a difficult nonconvex problem. We propose various heuristics for solving this problem. More effective heuristics, however, are computationally expensive in the number of patients to be scheduled, so this methodology is reserved for low-capacity (few patient) systems.

The subsequent chapter has us examine the scheduling problem for a high-capacity system. As discussed in the previous chapter, the heuristics perform poorly as the number of patients grows large. We dispose of all patient heterogeneity assumptions–a reasonable assumption in high-capacity settings such as vaccine clinics or dialysis centers. In this situation, however, we allow unpunctuality to be heterogeneous with respect to time-of-day that the appointment is scheduled. We look at the fluid limit of the queueing processes and formulate a fluid control problem that admits solutions for the appointment schedule distribution, also referred to as an *appointment profile*. We are then able to numerically solve the fluid control problem by using a time-discretization of the optimal control problem; the discretized form is a quadratic program. Leveraging Lindley's equations, this quadratic program has a linear number of variables and constraints in the time-discretization. We then proceed to evaluate schedules derived via our quadratic program in a discrete-event simulation setting using our UNC clinic data.

1.2 Literature Review for Appointment Scheduling Problems

(Bailey, 1952) is widely regarded as having published the first paper in the study of outpatient appointment systems in the mid-1900s along with (Jansson, 1966). Several comprehensive literature review papers exist. One by (Ahmadi-Javid and Klassen, 2017) provides a review of the optimization approaches across outpatient appointment systems in healthcare. Another by (Gupta and Denton, 2008) examines appointment systems in healthcare and discusses the difficulties of outpatient appointment scheduling, including future research areas. (Cayirli and Veral, 2003) provide a review of research in appointment scheduling in outpatient services.

We summarize outpatient scheduling papers that focus on multiple patient classes, patient unpunctuality, data-driven & simulation-based appointment scheduling, and fluid modeling of relevant queues. Note that, in most of the literature, when we refer to multiple patient classes, the classes are referring to unique service time distributions; these are often aligned with the idea of different surgeries, treatments, and consultation types.

When examining multiple patient classes, in addition to deciding the appointment times themselves, it is important to consider the sequence of patients within that schedule. (Heaney and Porter, 1991) examine 85 general practitioners and determine a large contribution to consultation time is the number of patients awaiting service within a particular service class in addition to other operational findings. (Ho and Lau, 1992) evaluate various scheduling rules and determine the biggest factors influencing patient and provider-centered costs are the no-show probability of patients, the coefficient of variation of service times, and the number of patients per scheduling window, in decreasing importance. (Klassen and Rohleder, 1996) address the sequencing of scheduled patients with respect to low-variance and high-variance service times. (Gerchak and Henig, 1996) examine the issue of scheduling surgeries in the presence of both elective surgeries and emergency surgeries using a stochastic dynamic programming formulation. (Dexter and Lubarsky, 1999) examine operating room scheduling strategies in the presence of variable surgery types; they perform computer simulation of patient schedules and examine patient preferences when it comes to surgery waiting time. (Denton and Gupta, 2003) examine a job scheduling system in the presence of different job duration distribution classes by deriving a two-stage stochastic linear program and producing sequential upper bounds on the problem via sequentially finer partitions on the job duration support

space. (Cayirli and Rosen, 2006) examine appointment schedules, particularly with attention to appointment sequencing when it comes to different service types by patient class. While testing several sequencing rules and appointment schedules, they conclude that the sequence patients are seen in can be just as important as the appointment schedule itself. (Kolisch and Sickinger, 2008) examine admission rules for 3 classes of customers seeking radiology healthcare services at a hospital while incorporating no-shows. (Mancilla and Storer, 2012) develop stochastic integer program solution algorithms for patient scheduling and sequencing rules with heterogeneous patients, with the main result being a heuristic based on Bender's decomposition. (Oh and Ptaszkiewicz, 2013) examine different patient types in addition to multiple server types (nurses and doctors) in a primary care scheduling context; they develop a stochastic integer program model for scheduling and sequencing of the heterogeneous patients in addition to a two-stage model that leverages the two different types of servers. (Erdogan and Denton, 2015) examine the analytical solution to sequencing two patients in an outpatient system relative to service completion time and waiting time. (Mak and Zhang, 2014) and (Kemper and Mandjes, 2014) derive results that support an ordered-by-variance sequencing rule as preferable over other index rules, particularly so when the service time distributions come from a single family of distributions with different scale parameters. (Berg and Huschka, 2014) and (Erdogan and Denton, 2015) also recommended the ordered-by-variance rule. (Tsai and Teng, 2014) examine a stochastic appointment scheduling procedure for physical therapy clinics where multiple resources can be utilized by a single patient depending on therapy type. (Kong and Zheng, 2016) determine the ordered-by-variance policy will perform well under certain assumptions, but is not generally optimal and loses its advantages depending on service-time distribution shape and size of the system. (Riise and Burke, 2016) develop a model for general surgery scheduling problems in addition to a search algorithm for efficiently solving it. (Mansourifard and Krishnamachari, 2018) derive a sequencing heuristic using a newsvendor formulation that addresses the issue of asymmetric patient waiting and doctor idle costs not captured by the ordered-by-variance rule. (Gocgun, 2018) examines dynamic scheduling within chemotherapy clinics, with patients coming from multiple classes with respect to preferences and appointment-time windows; an approximate dynamic programming technique is leveraged and performs better than myopic baseline heuristics. (Li and Fung, 2018) examine a dynamic scheduling model with heterogeneous patient preferences formulated as a Markov decision process; to address the curse of dimensional-

ity, two approximate dynamic programming algorithms are proposed. (Mandelbaum and Bunnell, 2020) derive a heuristic approach for scheduling and sequencing problems in the context of multiple servers, using an infinite-server approximation to derive a sequencing heuristic. (Jafarnia-Jahromi and Jain, 2020) proved that there does not exist an index-based optimal sequencing policy for the appointment scheduling problem. (Zhou and Yue, 2021) examine the customer wait and server idle-time minimization problem over multiple service stages via a sample-average approximation mathematical programming approach.

The effects of patient unpunctuality on appointment systems has been examined as early as 1964 by (White and Pike, 1964). However, the focus of incorporating unpunctuality in mathematical optimization models for appointment scheduling is a much more recent trend, likely due to the increasing availability of data provided by electronic check-ins and pre-visit surveys. Further, the complexity and analytical intractability induced by incorporating unpunctuality into appointment scheduling models has led to most studies keeping unpunctuality homogeneous (Klassen and Yoogalingam, 2014) and not dependent on either individual patient characteristics or scheduled appointment time. (Dexter, 1999) provide managerial insights regarding the source and improvement of large patient wait times in the presence of several factors, including patient unpunctuality and no-shows, an extreme form of unpunctuality. (Ho and Lau, 1999) examine nine scheduling rules in addition to the impact of several environmental factors on performance, such as no-show probability. (Cayirli and Veral, 2003) draw attention to patient unpunctuality in the context of appointment scheduling models. (Kim and Giachetti, 2006) develop a stochastic mathematical overbooking model for determining the number of patients to schedule in the presence of no-shows and walk-ins. (LaGanga and Lawrence, 2007) examine the overbooking problem with respect to improving patient access and provider productivity; they examine several overbooking levels in simulation studies to show the improvements from overbooking policies. (Alexopoulos and Wilson, 2008) examine the arrival and tardiness (unpunctuality) behaviors in community clinics and propose a nonhomogeneous Poisson process for modeling random patient arrivals. (Muthuraman and Lawley, 2008) examine the sequential call-in problem with overbooking to account to patient no-shows; practical observations and conditions on the unimodal evolution of the objective function are derived. (Lin and Lawley, 2011) develop a Markov decision process model of dynamic overbooking in the presence of no-shows with can be solved to optimality in small cases and solved approximately for larger problems; their

methods outperform myopic methods by exploiting information on the call-in process. (Tai and Williams, 2012) examine the optimization of patient appointment schedules while incorporating a new distribution for characterizing unpunctuality, known as the "F3" distribution. (Cheong and Fontanesi, 2013) examine real infusion center data and determine a mixture-exponential distribution to be a better fit to patient unpunctuality than prior Normal-based fitting attempts. (Luo and Guo, 2016) examine outpatient appointment scheduling with patient unpunctuality and develop a simulation model for determining the best appointment scheduling rules. (Samorani and Ganguly, 2016) analytically solve the 2-patient problem of deciding whether to wait for an earlier scheduled patient when patients arrive out of order (wait-preempt dilemma). (Deceuninck and Vuyst, 2018) also examine the wait-preempt dilemma while also incorporating patient unpunctuality into a sequencing problem with 2 patient classes: low-variance and high-variance unpunctuality; using local searches they find one of the best performing sequences involves putting low-variance patients towards the beginning and end of the sequence, with the high-variance patients in the middle. (Zhu and Liu, 2018) examine the multi-class patient scheduling problem with the addition of patient unpunctuality; they develop an analytical two-patient model that motivates an easy-to-implement heuristic policy. (Asadi and Yaghoobi, 2019) develop a stochastic mixed integer programming model for appointment scheduling that incorporates homogeneous unpunctuality assumptions of patient arrivals. (Deceuninck and Fiems, 2019) develop a variance-reduction technique to improve evaluation times of a schedule for use in a simulation optimization approach; they examine the change in the optimal schedule between punctual and unpunctual derivations. (Cayirli and Yang, 2019) examine the clinic performance metrics as a result of editing clinical environments such as patient-doctor cost ratio, service time variability, and the no-show probability. (Jiang and Yan, 2019) develop a mathematical programming formulation that seeks to minimize a sample average approximation of the system costs using Bender's Decomposition while incorporating homogeneous patient unpunctuality. Of particular interest is their in-depth analysis on how system performance is dependent on unpunctuality parameters. (Jiang and Xu, 2019) examine a distributionally robust optimization approach for appointment systems that incorporate no-shows. (Pan and Xie, 2019) develop a two-stage stochastic programming model approximately solved via Bender's decomposition for the appointment scheduling problem with patient unpunctuality and multiple servers. (Mandelbaum and Bunnell, 2020) develop a data-driven appointment scheduling and sequencing

heuristic that is robust to systems with large servers. This method uses an infinite-server relaxation to find a locally optimal appointment sequence that is able to handle individual patient characteristics with respect to unpunctuality, service time, and no show probability. They further develop a data-driven robust optimization approach via mathematical programming and Bender's Decomposition; however, this does not incorporate unpunctuality in its formulation. (Pan and Xie, 2021b) develop a two-stage mixed-integer program model for managing patient wait and clinic overtime in the presence of unpunctual patients, multiple servers, and no-shows and solve efficiently via a stochastic approximation technique. (Pan and Xie, 2021a) then further examine real-time patient sequencing in the presence of unpunctuality; that is, the prioritization of patients when multiple are to be selected from within a queue. However, there is little work that examines heterogeneity between patients with respect to unpunctuality and no work when the individual patients fall within their own distributional classes. (Moradi and Zolfagharinia, 2022) examine the difficulties introduced by patient unpunctuality to radiotherapy centers, developing a mixed integer program for determining the optimal sequence of patients for treatment. (Chen and Latina, 2023) simulate a hospital ultrasound department and examine the effects of varying patient unpunctuality and service discipline on system efficiency.

Data-driven and simulation-based optimization approaches are one of the most recent foci in the outpatient appointment scheduling literature. Most datasets are used to estimate parametric distributions for use in optimization models, but recently researchers have sought to use data directly in the optimization itself through both simulation-based and non-simulation-based approaches. (Brahimi and Worthington, 1991) design a scheduling approach to the Royal Lancaster Infirmary in the United Kingdom using queuing models. (Liu and Liu, 1998) examine a dynamic multi-server job-assignment model which is extended to a static outpatient appointment scheduling problem; Poisson approximation and simulation-based methods are used to compute the capacity distribution, with the latter method allowing for the incorporation of general service time distributions. (Harper and Gamlin, 2003) simulate an ENT clinic that allows comparison of various schedules for picking the best one based on patient and clinic factors. (Castaing and Weizer, 2016) develop a two-stage stochastic mixed-integer programming result for chemotherapy center scheduling with a single nurse serving several patients using real-world data for deriving managerial insights and schedule performance. (Alvarado and Ntaimo, 2018) develop a mean-risk stochastic programming

model for chemotherapy appointment scheduling under different appointment durations, acuity levels, and nurse availabilities that outperform deterministic-based approaches significantly. (Kim and Cha, 2018) provide a data-driven study on an appointment system using data provided by an outpatient clinic. Their main contributions are not in developing an optimal schedule, but rather to model and understand the arrival process and also provide some insight into what assumptions are valid to make in mathematical models based on their data. (Jiang and Yan, 2019) examine a single-server system with a fixed patient sequence that uses historical data to develop uncertainty sets for stochastic optimization. Using a specific construction methodology, they are able to produce tractable models that provide good approximations to the more computationally intensive stochastic programs. (Hribar and Chiang, 2019) collect electronic health record timestamp data and design a new scheduling template based on 3 appointment lengths (short/medium/long). This new template is tested on simulation models and compared to observed implementations. Overall, the schedules are shown to improve clinical efficiency measures such as mean clinic volume, patient wait times, and examination times. (Dogru and Melouk, 2019) develop a data-driven simulation-optimization approach for outpatient appointment scheduling in a patient-centered medical home setting that incorporates patient preferences, no shows, unpunctuality, and future appointment requests. (Mandelbaum and Bunnell, 2020) provide a data-driven infinite-server heuristic and data-driven robust optimization approach to problems involving potentially large numbers of identical servers motivated by a chemotherapy infusion centre setting.

Fluid or diffusion models of appointment systems are often used when the underlying model meets some kind of limiting condition on its parameters. These are often used as approximations of actual processes that may be intractable if viewed through a discrete-event lens. An understandable and concise introduction to fluid and diffusion limits of queues is provided in Chapter 8 of (Gautam, 2012). (Whitt, 2006) examines deterministic fluid models of multiserver queues with customer abandonment. (Lee and Zenios, 2009) examine regularly-visiting patients and address the idleness induced by frequent hospitalization via the consideration of overbooking; local diffusion approximations are used to derive closed-form expressions of the optimal capacity and overbooking level. (Honnappa, 2015) provide the analytical framework for patients arriving to a queue according to independent sampling from a common arrival distribution. Due to the intractability of the discrete-event model, the fluid limit and diffusion limit of the queue are examined to provide

tractable results. (Chen and Thomas, 2016) examine the fluid process of patient arrivals in complex, network-like health care settings. (Lu and Ying, 2017) examine a fluid-approximation queueing network on a large hospital system for the purpose of optimally allocating cashiers and pharmacists. (Zacharias and Armony, 2017) examine the joint problem of determining panel size and number of appointments per day with a diffusion approximation for evaluating performance. (Armony and Honnappa, 2019) examine the asymptotically optimal appointment scheduling problem in the context of fluid and diffusion limits and determine analytical results for optimal appointment schedules. They do not, however, incorporate generalized patient unpunctuality in their work, incorporating no-shows. (Lee and Zhang, 2021) examine a production queueing system with nonstationary demand by developing a fluid-control problem that serves a lower bound for the performance of the queueing formulation. (Mehrizi and Faradonbeh, 2022) examine the diffusion limits of multi-class advance patient scheduling and formulate the optimal scheduling problem as a Brownian control problem.

1.3 Introduction to Asymptomatic Disease Testing

Ever since the outbreak of the COVID-19 (SARS-CoV-2) pandemic, disease testing, quarantining, and vaccination policies have drawn greater attention. While waiting for vaccinations to become publicly available, testing and quarantining became of paramount concern for public health officials. Individuals demonstrating mild to moderate symptoms would be encouraged to quarantine themselves for a period of around 2 weeks whereas those with severe symptoms would check into hospitals and remain separated from the general public until discharge. For symptomatic patients, disease tests were generally used to confirm the disease itself as other potential diseases with similar symptoms could be the culprit. In the absence of symptoms, however, there remains the risk that an individual is in fact infected, infectious, and completely unaware of their ability to spread the disease to others. Despite the lack of symptoms that can contribute to the spread of a disease (coughing, sneezing, etc.), asymptomatic diseases can still be very transmissible, exacerbated by the fact that a host is unaware and unlikely to take precautionary actions; further, official miscommunications lead to the underestimation of the transmissibility of asymptomatic cases (Lippi and Plebani, 2020). Various case studies have shown the strong impact that asymptomatic cases have

had on the global community: (Li and Shaman, 2020) estimated that almost 80% of SARS-CoV-2 infections were caused by asymptomatic individuals. A different study (Oran and Topol, 2020) estimated that around 40-45% of SARS-CoV-2 infections resulted from asymptomatic individuals and that the infectious period lasts for as many as 14 days. Further, (Chau and Tan, 2020) found that the viral load in asymptomatic infections of SARS-CoV-2 is ultimately comparable to that in symptomatic patients. In light of these facts, testing also provide an important function of helping to detect an asymptomatic infection.

During the pandemic in the United States, free COVID-19 at-home tests were delivered to households in limited quantities, providing a safe, convenient alternative to visiting crowded and sometimes distant testing centers. However, these tests were in limited supply in the early stages of the pandemic and guidelines were lacking on how to properly use them. The Center for Disease Control (CDC) did provide the sage advice that tests should be used before visiting vulnerable people (elderly, immunocompromised, etc.) and a few days after attending larger gatherings; however, for people going about their daily lives, how should the tests best be used? As these tests tend to expire in a matter of months, holding on to them in case of a special event may not be an option. More advanced tests with medical-grade accuracy, such as the BINAXNOW COVID-19 Antigen Self-Test, arrived; while supply would not be limited by manufacturing capabilities, the large price tag on these tests equate to a very limited supply for the average person.

Our goal in Chapter 4 of this work is to develop guidelines for the average user to follow when an asymptomatic infection is a real possibility. The ultimate goal is to minimize the number of undetected days the individual experiences. We develop several analytical-backed policies for individual test use, then proceed to examine the performance of these policies in a community-based simulation.

1.4 Literature Review for Asymptomatic Disease Testing

We will examine relevant disease testing literature which generally falls into one of two classes: community-based disease testing and economic analyses of disease testing procedures.

(Tsay and Baldea, 2020) develop a partially-observable optimal control problem for sequencing social distancing and testing events to minimize the number of infections. They find that quar-

antine procedures are most effective when implemented early and "on-off" policies for quarantine procedures can effectively flatten the infectious curve while maintaining a low socioeconomic cost. (Sasmita and Chongsuvivatwong, 2020) develop mathematical control models with 5 strategies for control: lockdowns, contact tracing, mass testing, case detection and treatment, and face mask wearing. They determine a combination of lockdowns, contact tracing, case detection and treatment, along with facemasks produces the best result for limiting the spread. (Madubueze, 2020) examine time-dependent intervention strategies: quarantine, isolation, and public health education, in the context of an optimal control problem. Time-dependent interventions were shown to outperform time-independent interventions, with implementation within the first 100 days of an outbreak producing the best results. (Ghosh and Ghosh, 2021) examine complex, heterogeneous community networks where test kits can be implement at specific nodes to impact the spread of an epidemic; their numerical results confirm the reduction in spread by targeting well-connected nodes within the network. Their work ends with a numerical study based on data from two real networks: the global airport network and the transportation network of Kolkata, India. (Ely and Steiner, 2021) propose an optimization model that works with a set of tests with varied sensitivities and specificities and seeks to optimally assign them to a heterogeneous set of individuals; they derive a simple necessary condition for the optimality of test allocation among other structural insights. Last, they characterize the marginal benefit of a test and provide an algorithm for solving the test-allocation problem. (Choi and Shim, 2021) examine optimal strategies for social distancing and testing in order to control the spread of COVID-19. They determine that testing alone will not affect the overall spread, but will delay the peak of the pandemic. Combining social distancing with testing will produce the best results, with the most testing occurring in the early phases of the pandemic and after the peak and social distancing measures growing in intensity as the disease grows more prevalent. (Vatcheva and Villalobos, 2021) examine data from the Rio Grande Valley of Texas in 2019 to determine the optimal levels of social distancing and testing to slow virus spread at the onset of the COVID-19 pandemic. A simulated 4-month period develops social distancing and testing policies that consider heterogeneous classes of community members such as adults, children, and seniors. (Lyng and Berke, 2021) examine various population settings for COVID-19 testing along with testing goals. Using simulated data, they examine the performance of various testing policies within a susceptible-infectious-recovered (SIR) model. Schools and businesses are chosen

as the key population settings with more frequent testing correlating with reduced disease burden. (Buhat and Mamplata, 2021) develop a nonlinear programming allocation model for COVID-19 test kits across 11 testing centers in the Philippines. Their model incorporates heterogeneity in testing accessibility, population density of municipalities, and testing facility capacity. (Wells and Galvani, 2021) develop a mathematical model for combining testing with quarantines in a COVID-19 setting; they find discover what testing policies allow for shorter quarantine periods to maintain the same effectiveness as longer periods, with testing-on-exit policies leading to 50% reductions on a typical quarantine period. (Olivares and Staffetti, 2021) develop an optimal control model for vaccination and disease testing with limited resources. Their controls are the daily vaccination and testing rates; using a direct transcription technique, they are able to produce solutions that perform well in various numerical experiments. (Piguillem and Shi, 2022) examine the economic improvements provided by lockdowns (large-scale quarantines) and testing; they determine that testing is a cost-efficient substitute for lockdowns that render the lockdowns almost unnecessary, allowing for increased economic output. They further examine an economic-driven analysis of lockdown policy types that show intensity and duration of the lockdowns are invariant to several parameters: the weighting between lives and economic output and the aversion to GDP variability; instead, the virus dynamics best dictate the optimal policy. (Aldila and Samiadji, 2022) develop an optimal control model for rapid-test deployment and find, from their numerical experiments, that a time-dependent rapid-test intervention can successfully suppress the spread of COVID-19 with a low intervention cost. (Tatsuki and Igarashi, 2023) develop a SEIRD model that incorporates testing characteristics and limited testing resources. Simulations studies saw that cumulative deaths could range up to thousands less depending on the testing policies implemented.

In Section 4.4, we use model parameters influenced by the following papers. (Ma and Liu, 2021) examine asymptomatic infections of COVID-19 using data from PubMed, EMBASE, and ScienceDirect and conducted several analysis of different strata within the population; their most general infected-wide estimate of the asymptomatic infection rate was 40.50%. (Wells and Galvani, 2021) state that the recommended COVID-19 quarantine is 14 days, noting the mental health strain of such a lengthy period of near-isolation.

CHAPTER 2

Data-driven Scheduling and Sequencing of Unpunctual Patients in Health-care Clinics

2.1 Motivation

In today's environment, regularly visiting patients to a clinic can accrue large amounts of data regarding their unpunctuality behavior. With sufficient data for multiple patients, their individual distributions can be examined and leveraged for optimal scheduling of future appointments. We first examine real UNC clinic data that provides insight into differing individual patient characteristics. Considering the prospect of fully individualized distributions for patients in both unpunctuality and service time, we develop several heuristic methods for solving the general appointment scheduling problem. The most intensive heuristics are designed to improve upon the state-of-the-art by incorporating the individualized unpunctuality as a facet of the optimization process. Using our data again, we perform several simulation studies that show the improvements introduced by incorporating individualized unpunctuality as opposed to the typical homogeneous assumption. Further, we examine the performance of our heuristics against common-sense methods and the state-of-the-art.

2.2 UNC Clinic Data Analysis

In this section, we examine the unpunctuality behavior of patients using anonymized data obtained from UNC health care clinics. The dataset consists of approximately 150,000 records of patient appointments over a period of 2 years (730 days). Each record consists of data to determine patient unpunctuality and their service times in addition to unique patient, doctor, and clinic identifiers. Further details on the variables and their characteristics are presented in the supplement.

Using this data, we examine two particular values of interest: patient unpunctuality and patient service times. For each record, we use the appointment time and actual arrival time to determine

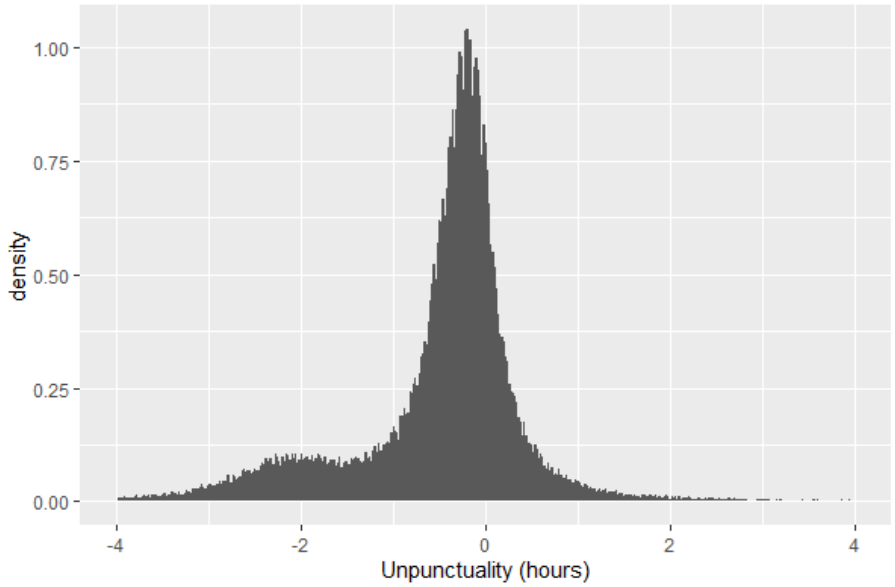

Figure 2.1: Unpunctuality distribution across the full dataset.

unpunctuality and use the time the doctor entered the room and the patient's check-out time to determine the service time.

We define unpunctuality of a patient as the difference between their scheduled appointment time and their actual arrival time. Thus, the patient unpunctuality is 0 if the patient is perfectly on time for their appointment, positive if the patient is late for their appointment, and negative if the patient is early for their appointment. We define the service time for a patient as the difference between the actual completion time of the appointment and the time the doctor entered the room.

We perform a simple filter of the data: any appointment with either an unpunctuality of magnitude greater than 4 hours or a service time less than 0.25 minutes or greater than 8 hours is removed from the analysis. Such extreme values often occur due to data entry error.

Figure 2.1 shows the histogram of unpunctuality across all patient visits. From this figure we see that the distribution appears to have two local modes, with the dominant mode being slightly earlier than 0 and the second mode being around 2 hours early.

One important question is whether and how much unpunctuality behavior changes from patient to patient. While our dataset does not provide the service or doctor types associated with certain appointments and patients, we are able to use the anonymized doctor IDs, anonymized patient IDs,

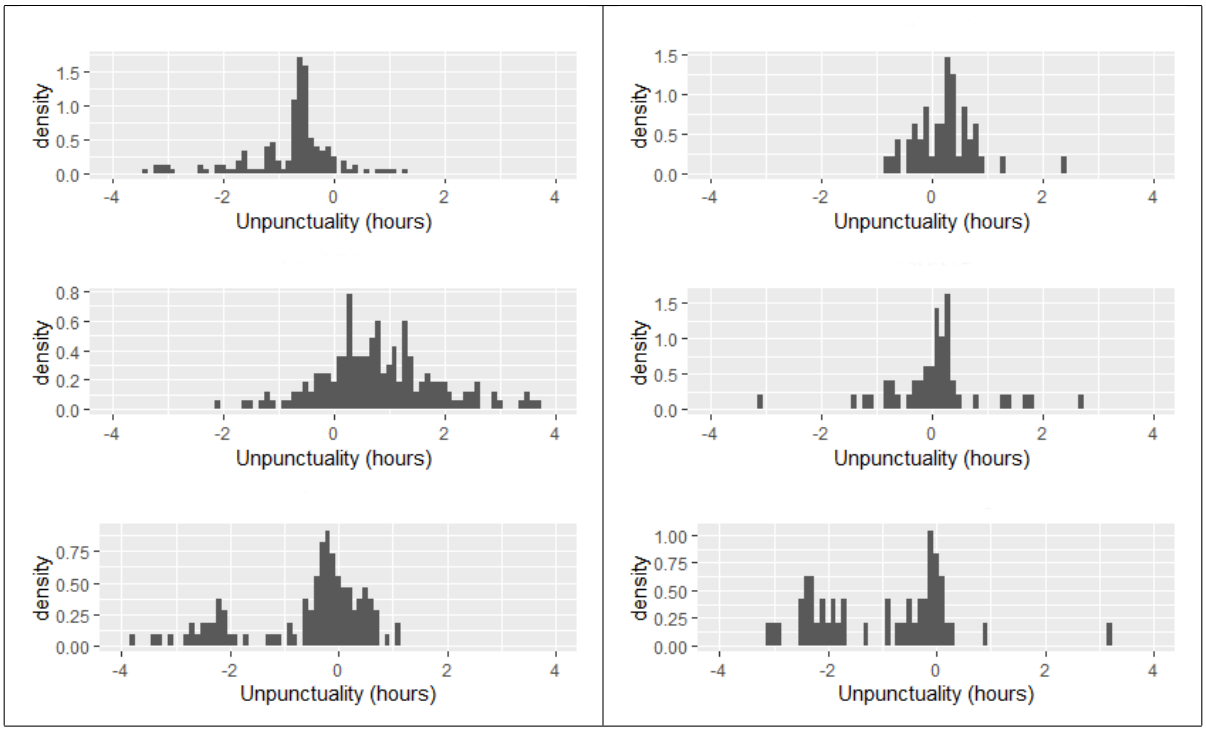

Figure 2.2: Unpunctuality distributions of 3 high-frequency patients on the left and 3 regular-frequency patients on the right

and anonymized clinic IDs to track the unpunctuality of a given patient over several appointments or cluster appointments and patients. We first divided the patients into two groups: those with three or more appointments a week on average (high-frequency) and those with less than three visits per week on average (regular-frequency). We picked three patients with the most visits from each of the two groups, thus giving us six patients with non-sparse distributions for their unpunctuality. We examined individual unpunctuality of these six patients. Figure 2.2 displays the full distribution of unpunctuality of the high-frequency patients (the left side) and the three regular-frequency patients (the right side).

As we can see from this figure, the idea of heterogeneous patients with respect to unpunctuality distributions makes sense as we can see different types of unpunctuality behavior; for example, the first high-frequency patient appears to be almost always early whereas the second high-frequency patient tends to be late.

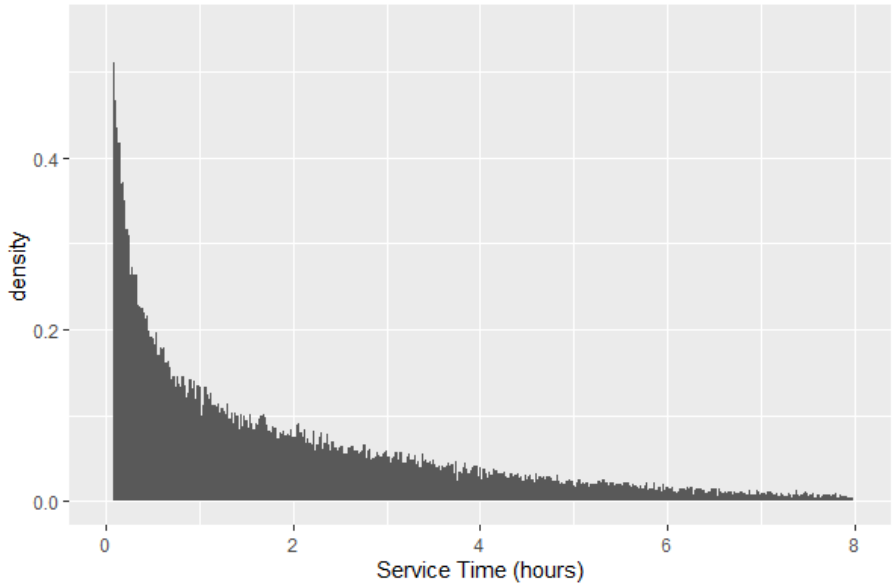

Figure 2.3: Service time distribution across the full dataset.

We present the overall service time distribution across all patients in Figure 2.3, which appears exponential in its general shape.

We also examined the individual patient distributions with respect to service times. Figure 2.4 examines the same six high-frequency and regular-frequency patients with respect to their individual service time distributions.

2.3 Heterogeneous Patient Model

In this section we formulate the problem of optimally scheduling P patients indexed by $i = 1, 2, \cdots, P$ over $[0, T]$. We split the problem into two pieces: one involving the determination of appointment times:

$$\boldsymbol{a} = [a_1, a_2, \cdots, a_P]$$

such that

$$0 \leq a_{\leq} a_2 \leq \cdots \leq a_P \leq T,$$

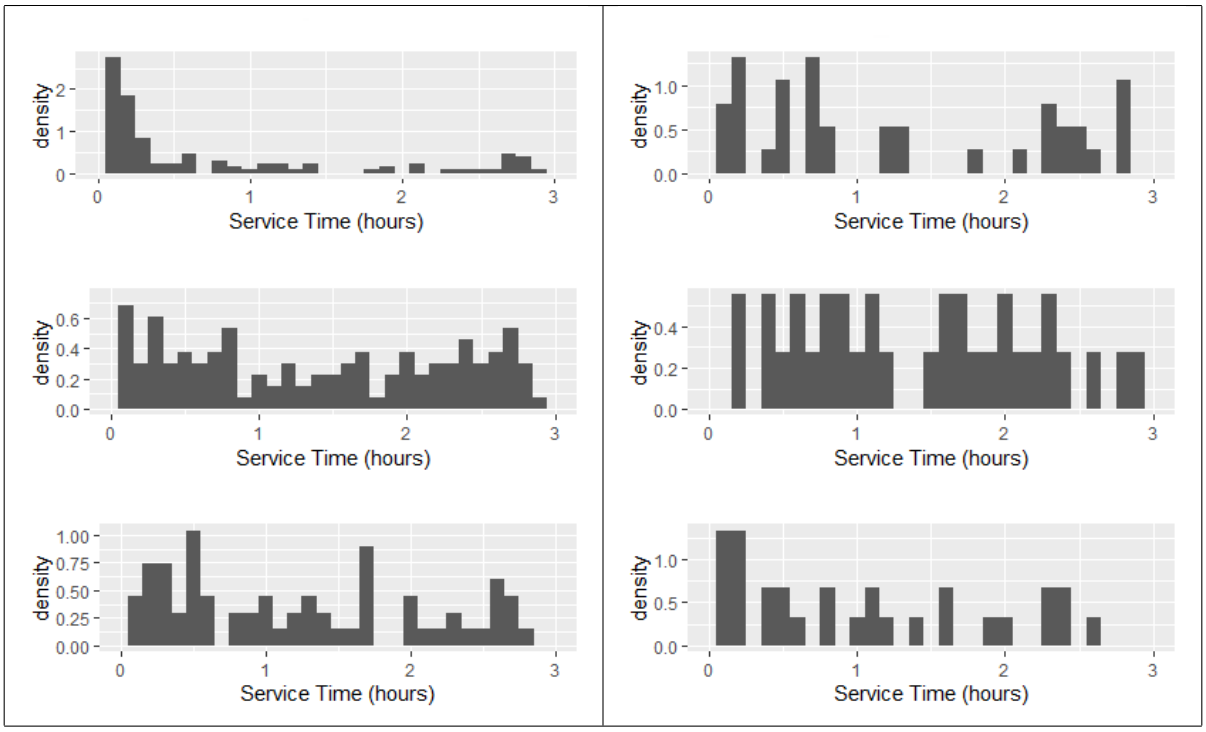

Figure 2.4: Service time distributions of the usual 3 high-frequency patients on the left and 3 regular-frequency patients on the right.

and the other determining a patient sequence vector:

$$\boldsymbol{q} = [q_1, q_2, \cdots, q_P]$$

that is a permutation of $\{1, 2, \cdots, P\}$. Under a given schedule $\mathbf{a}$ and $\mathbf{q}$, the patient with index q_j is scheduled for time a_j for $j = 1, \ldots, P$.

Let U_i be the unpunctuality and S_i the service time of the ith patient. Then the patient scheduled to arrive at time a_k actually arrives at time

$$A_k = a_k + U_{q_k},$$

and their service time is given by S_{q_k}. Note that the order in which patients arrive can be different than the order of their scheduled appointment times.

We consider a cost structure that incorporates waiting costs for the patients, and idle-time and overtime costs for the servers. We distinguish between the waiting time incurred by the patient before their appointment times, and after their appointment times. Specifically, c_b denotes the cost

per patient per unit time for a patient waiting in the queue before their scheduled appointment time, and c_a denotes the cost per patient per unit time for a patient waiting in the queue after their scheduled appointment time. We choose to separate these two types of waits as it is reasonable to assess them differently. Patients expect to be seen in a timely manner and can be expected to accrue a higher wait cost when waiting past their scheduled appointment time than when waiting prior to the scheduled appointment time. Finally, c_I denotes the cost per unit time per server that is idle during the day (i.e., during $[0, T]$), and c_o denotes the cost per unit time per server that is working beyond time T; i.e., when they are accruing overtime.

The presence of unpunctuality gives rise to the phenomenon of patients arriving out of order. This creates a question: in what order should patients be served? We consider the following two non-idling non-preemptive service disciplines:

1. Appointment-Based Priority (ABP). Under ABP, when a server completes service, he starts serving the patient with the earliest appointment time.

2. Early/Late Hybrid (ELH). ELH maintains two queues: an early queue of patients with unpunctuality less than or equal to 0, and a late queue of patients with unpunctuality greater than 0. When a server completes service, and the early queue is nonempty, a patient in that queue is selected according to ABP. If the early queue is empty, then the server takes a patient from the late queue according to a first-come first-serve (FCFS) service discipline.

Our objective is to find a schedule that minimizes the total expected cost of our system, assuming one of the above service disciplines is followed. Note that we are considering an "offline" setting where the schedule is to be set before the start of the day and cannot be changed during the day.

To make this more precise, we introduce the following notation. Let T_i be the time when the service of patient q_i starts. Then the queueing time of customer q_i prior to time $a_i$is given by

$$W_b(i, \boldsymbol{a}, \boldsymbol{q}) = \max(\min(a_i - A_i, T_i - A_i), 0),$$

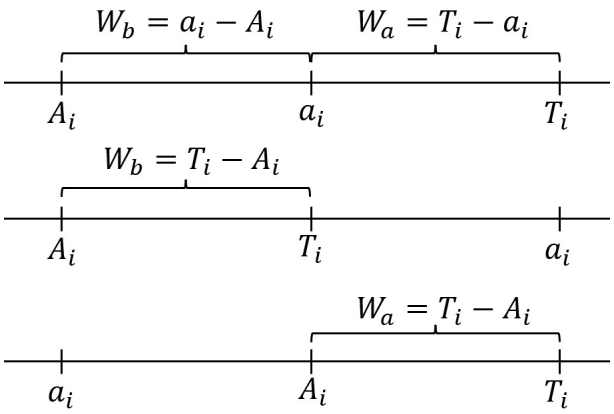

Figure 2.5: W_b and W_a values for 3 different scenarios.

and the queueing time of customer q_i after time a_i is given by

$$W_a(i, \boldsymbol{a}, \boldsymbol{q}) = \max(\min(T_i - A_i, T_i - a_i), 0).$$

Let $I(\boldsymbol{a}, \boldsymbol{q})$ be the total idle time across all servers from time $t = 0$ until all customers have entered service, and $O(\boldsymbol{a}, \boldsymbol{q})$ be the total overtime across all servers (summation of service times beyond time $t = T$). We also denote the set of all permutations of $\{1, 2, \cdots, P\}$ by Q.

Our general optimization problem is

$$\begin{aligned} \min_{\boldsymbol{a},\boldsymbol{q}} \quad & E\left[c_b \sum_{i=1}^{P} W_b(i, \boldsymbol{a}, \boldsymbol{q}) + c_a \sum_{i=1}^{P} W_a(i, \boldsymbol{a}, \boldsymbol{q}) + c_I I(\boldsymbol{a}, \boldsymbol{q}) + c_o O((\boldsymbol{a}, \boldsymbol{q}))\right] \\ \text{s.t.} \quad & 0 \le a_i \le T \quad \forall i = 1, 2, \ldots, P \\ & a_i \le a_j \ \forall i < j \\ & \boldsymbol{q} \in Q. \end{aligned} \tag{2.1}$$

In full generality, the objective function is hard to evaluate. In simple cases, such as when there is a single server and the service discipline is first-come first-serve, one can write recursive formulas (Lindley's Equations) which can be used to calculate successive wait times, overtime, and idle times given a sample path of unpunctuality, service times, and an appointment schedule.

However, for our system in its full generality, one cannot express these analytically and we must rely on a simulation-type method for calculating these values. Therefore, we use a sample-average approximation (SAA) in lieu of our original objective function.

Specifically, to approximate the objective function for a given $\mathbf{a}$, $\mathbf{q}$, and service discipline π, we generate M sample paths. For sample path $m \in \{1, \ldots, M\}$, let

$$\left(W_a^{(m)}(i, \boldsymbol{a}, \boldsymbol{q}), W_b^{(m)}(i, \boldsymbol{a}, \boldsymbol{q}), I^{(m)}(i, \boldsymbol{a}, \boldsymbol{q}), O^{(m)}(i, \boldsymbol{a}, \boldsymbol{q})\right)$$

be the computed values of

$$(W_a(i, \boldsymbol{a}, \boldsymbol{q}), W_b(i, \boldsymbol{a}, \boldsymbol{q}), I(i, \boldsymbol{a}, \boldsymbol{q}), O(i, \boldsymbol{a}, \boldsymbol{q}))$$

respectively for the m-th sample path. We use these quantities to estimate the objective function in the above optimization problem by defining

$$\Phi^{(M)}(\mathbf{a}, \mathbf{q}) = \frac{1}{M} \sum_{m=1}^{M} \left[\sum_{i=1}^{P} \left[c_a W_a^{(m)}(i, \boldsymbol{a}, \boldsymbol{q}) + c_b W_b^{(m)}(i, \boldsymbol{a}, \boldsymbol{q}) \right] + c_I I^{(m)}(\boldsymbol{a}, \boldsymbol{q}) + c_o O^{(m)}((\boldsymbol{a}, \boldsymbol{q})) \right]$$

as the sample average approximation of the objective in Problem 2.1 across M sample paths. This yields the following SAA minimization problem:

$$\begin{aligned} \min_{\boldsymbol{a}, \boldsymbol{q}} \quad & \Phi^{(M)}(\mathbf{a}, \mathbf{q}) \\ \text{s.t.} \quad & 0 \leq a_i \leq T \quad \forall i = 1, 2, \ldots, P \\ & a_i \leq a_j \ \forall i < j \\ & \boldsymbol{q} \in Q. \end{aligned} \tag{2.2}$$

This is a nonconvex optimization problem that is difficult to solve to global optimality; however, we were able to produce high-performance appointment schedules using our heuristics.

2.4 Analytical Sequencing Results

Classic results from job scheduling literature in manufacturing, such as by (Pinedo, 2012) have established optimal sequencing rules for punctual jobs with heterogeneous stochastic completion times. However, there are no analytical results for how heterogeneous unpunctual patients should be sequenced. Due to the complexity of Problem 2.1, provable analytical results can only be established under various restrictions. We will present some analytical results regarding the sequencing of patients relative to their unpunctuality distributions.

Definition 2.1. A patient sequence $\boldsymbol{q} = (1, 2, \ldots, P)$ is said be a stochastically increasing unpunctuality (SIU) sequence if $U_1 \leq_{st} U_2 \leq_{st} \cdots \leq_{st} U_P$.

We can prove the following proposition.

Proposition 2.4.1. Consider an equally spaced appointment schedule $\boldsymbol{a} = (a_1, \cdots, a_P)$ with $a_k = (k-1)a$. Suppose all service times $S_k, k = 1, \ldots, P$ are i.i.d. such that $P(S_k \geq a) = 1$. The patients are served in the order of appointments. Then the total patient wait time of the system is stochastically minimized if $\boldsymbol{q}$ is an SIU sequence.

The proof of Proposition 2.4.1 is located in the appendix. This theorem follows the intuition that, if we want to minimize overall wait-time, we should sequence in such a way that patients arrive as spread out as possible. The idea that $P(S_k \geq a) = 1$ corresponds to an *overloaded* clinic, as patients cannot be served in their allotted times.

Definition 2.2. A patient sequence $\boldsymbol{q} = (1, 2, \ldots, P)$ is said be a stochastically decreasing unpunctuality (SDU) sequence if $U_1 \geq_{st} U_2 \geq_{st} \cdots \geq_{st} U_P$.

Proposition 2.4.2. Consider an equally spaced appointment schedule $\boldsymbol{a} = (a_1, \cdots, a_P)$ with $a_k = (k-1)a$. Suppose all service times $S_k, k = 1, \ldots, P$ are i.i.d. such that $P(S_k \leq a) = 1$. The patients are served in the order of appointments. Then the total idle time of the system is stochastically minimized if $\boldsymbol{q}$ is an SDU sequence.

The proof of this proposition follows an identical structure to Proposition 2.4.1 by using the fact that the inter-patient idle time is the negative of the inter-patient wait time. This corollary follows the intuition that, if we want to minimize the overall idle-time, we should sequence in such

a way that patients arrive with the smallest gaps between them. The idea that $P(S_k \leq a) = 1$ corresponds to an *underloaded* clinic, as the patients will finish faster than their allotted times.

2.5 Alternating Iterative Perturbation and Resequencing (AIPR) Heuristic

The objective function of Problem (2.2) is nonconvex and therefore standard gradient-descent algorithms do not necessarily perform well. Further, the computation of a gradient is computationally expensive for general systems that require simulation to compute the objective function. Hence, in this section, we propose an alternative heuristic method. Specifically, we propose a two-phase method that alternates between optimising with respect to $\boldsymbol{a}$ while keeping $\boldsymbol{q}$ fixed and vice-versa. Phase-1 of the heuristic is based upon the local search method detailed by (Deceuninck and Vuyst, 2018). The Phase-1 algorithm seeks to improve only the appointment vector $\boldsymbol{a}$ while keeping the sequence $\boldsymbol{q}$ fixed. The Phase-2 algorithm seeks to find a better sequence vector $\boldsymbol{q}$, while keeping the appointment vector $\boldsymbol{a}$ fixed. We will first define a $\tau-$neighbourhood of appointment vector $\mathbf{a}$.

Definition 2.3. Let $N(\tau, \mathbf{a}) = \{\mathbf{a} + \tau\mathbf{e}_i \mid i \in \{1, 2, \ldots, P\}\}$ where $\mathbf{e}_i$ is the ith column of the $P \times P$ identity matrix. Then $N(\tau, \mathbf{a})$ is a $\tau-$neighbourhood of $\mathbf{a}$.

Using this definition, we will now describe our Phase-1 Heuristic.

Algorithm 2.5.1. Iterative Perturbation

Initial Input: $\mathbf{a}, \mathbf{q}, \tau_1 > \tau_2 > \cdots > \tau_N > 0, M \in \mathbb{Z}^+$

for τ in $(\tau_1, \tau_2, \ldots, \tau_N)$ **do**

Set $\hat{\Phi}(\mathbf{a}) = \min_{\mathbf{a}' \in N(\tau, \mathbf{a})} \Phi^{(M)}(\mathbf{a}', q)$

if $\hat{\Phi}(\mathbf{a}) < \Phi^{(M)}(\mathbf{a}, \mathbf{q})$ **then**

$\mathbf{a} \leftarrow \mathbf{a}'$

Repeat for loop for current τ

else

if $\tau = \tau_N$ **then**

Terminate algorithm

else

Proceed to next τ value in loop

end if

end if

end for

In this algorithm, we repeatedly pick the best single-appointment perturbation of size τ. Such a rule can be replaced with a different approach; for example, considering multiple appointment perturbations at once. It is important to note that picking large τ means the algorithm will likely converge quickly; however to increase accuracy and speed, it is recommended to repeat the algorithm for a decreasing sequence of positive τ values defined by $\tau_1 > \tau_2 > \cdots > \tau_N > 0$. Iterative Perturbation converges as the termination criteria is non-improvement and the step-sizes are a fixed sequence.

Next we describe Phase-2, which keeps a fixed appointment vector $\boldsymbol{a}$ but adjusts the sequence vector $\boldsymbol{q}$. For this, we will define a swap-neighbourhood of sequence vector $\mathbf{q}$.

Definition 2.4. Let $\mathbf{q}$ be a sequence vector and let

$N_s(\mathbf{q}) = \left\{\mathbf{q}' \mid \mathbf{q}'_i = \mathbf{q}_j, \mathbf{q}'_j = \mathbf{q}_i \text{ for some } i < j, \mathbf{q}'_k = \mathbf{q}_k \ \forall k \notin \{i, j\}\right\}$.

Then we call $N_s(\mathbf{q})$ a swap-neighbourhood of sequence vector $\mathbf{q}$.

Note that $N_s(\mathbf{q}) \subset Q$ created by interchanging a pair of elements in the original sequence $\mathbf{q}$.

Algorithm 2.5.2. Iterative Resequencing

Initial Input: $\mathbf{a}, \mathbf{q}, M \in \mathbb{Z}^+$

Set $\hat{\Phi}(\mathbf{q}) = \min_{\mathbf{q}' \in N_s(\mathbf{q})} \Phi^{(M)}(\mathbf{a}, q)$

if $\hat{\Phi}(\mathbf{q}) < \Phi^{(M)}(\mathbf{a}, \mathbf{q}')$ **then**

$\mathbf{q} \leftarrow \mathbf{q}'$

Repeat algorithm

else

Terminate algorithm

end if

Remark: One does not need to limit themselves to greedy pair-wise interchanges, other resequencing approaches are equally as plausible.

Iterative Resequencing converges since the number of possible sequences is finite and thus we will eventually find a sequence such that any pair-wise interchange will not lead to an improvement in the objective function.

Finally, we describe our 2-phase heuristic, alternating the iterative perturbation and iterative resequencing processes.

Algorithm 2.5.3. Alternating Iterative Perturbation and Resequencing (AIPR)

Initial Input: $\mathbf{a}, \mathbf{q}, \tau_1 > \tau_2 > \cdots > \tau_N > 0, M \in \mathbb{Z}^+$

Perform Iterative Perturbation with Initial Input: $\mathbf{a}, \mathbf{q}, \tau_1 > \tau_2 > \cdots > \tau_N, M \in \mathbb{Z}^+$

$\mathbf{a}'$ = final output from Iterative Perturbation.

Perform Iterative Resequencing with Initial Input: $\mathbf{a}', \mathbf{q}, M \in \mathbb{Z}^+$

$\mathbf{q}'$ = final output from Iterative Resequencing.

if $\Phi^{(M)}(\mathbf{a}', \mathbf{q}') < \Phi^{(M)}(\mathbf{a}, \mathbf{q})$ **then**

$\mathbf{a} \leftarrow \mathbf{a}'$

$\mathbf{q} \leftarrow \mathbf{q}'$

Repeat algorithm

else

Terminate algorithm

end if

Note that computation of $\Phi^{(M)}(\mathbf{a}, \mathbf{q})$ can be expensive depending on service discipline, number of patients P, and number of sample paths M.

2.6 Myopic Scheduling Heuristic

In this section we propose a computationally simpler alternative to AIPR. This method requires a decided-upon sequence beforehand. The well-regarded rule of scheduling with shortest variance service times first performs well in practice. We first analyze a two-patient system and then heuristically generalize to more than two patients. Let c_w be per unit cost of queueing (before or after the appointment). Let W_i be the queueing time for patient $i = 1, 2$ and I_i be the idle time of the server waiting on patient $i = 1, 2$ to arrive. We will assume that patients are seen in the order of appointment, making $W_1 > 0$ if and only if the patient arrives before the system start time of 0. Let $g = a_2 - a_1$, be the inter-appointment time. Given the system assumptions, we can express

the wait times and idle times as follows:

$$W_1 = \max\{0, -a_1 - U_1\}, \quad W_2 = \max\{0, U_1 + S_1 - U_2 - g\},$$

$$I_1 = \max\{0, a_1 + U_1\}, \quad I_2 = \max\{0, g + U_2 - U_1 - S_1\}.$$

We choose a_1 and g by solving the following optimization problem:

$$\begin{aligned} \min_{a_1, g} \quad & E\left[c_w(W_1 + W_2) + c_I(I_1 + I_2)\right] \\ \text{s.t.} \quad & a_1 \geq 0, \ g \geq 0. \end{aligned} \tag{2.3}$$

Proposition 2.6.1. Let

$$a_1^\star = -F_{U_1}^{-1}\left(\frac{c_w}{c_w + c_I}\right), \tag{2.4}$$

and

$$g^\star = F_{R_{1,2}}^{-1}\left(\frac{c_w}{c_w + c_I}\right), \tag{2.5}$$

where $R_{1,2} = U_1 + S_1 - U_2$, and F_{U_1} is the CDF for U_1 and $F_{R_{1,2}}$ is the CDF of $R_{1,2}$. Then, $a_1^\star$ and $g^\star$ are optimal solutions to Problem 2.3.

The proof of this proposition is located in the appendix.

We now use this result to construct a heuristic to decide the appointment times a_i ($i = 1, 2, \cdots, P$) in a system with $P \geq 2$ patients with a given sequence $\boldsymbol{q} = (1, 2, \cdots, P)$. We determine a_1 a given by Equation 2.4 and determine a_i for $i = 2, \ldots, P$ recursively as follows:

$$a_{i+1} = a_i + \max\left\{0, F_{R_{i,i+1}}^{-1}\left(\frac{k_i c_w}{k_i c_w + c_I}\right)\right\}$$

for some integer $k_i \geq 1$ with

$$R_{i,i+1} = U_i + S_i - U_{i+1}.$$

In this recursion, the unit waiting cost is adjusted by k_i to account for the fact that the waiting cost to be incurred will be more than what would be incurred in a 2-patient system as assumed in the simplified model.

Specifically, we set k_i as follows. Suppose appointments $a_1, \ldots, a_i$ have already been scheduled and we seek to schedule appointment a_{i+1}. Let $\bar{u}_j$ be the expected unpunctuality of the patient scheduled in the $j-$th appointment slot for $j \leq i$. Similarly, let $\bar{s}_j$ be the expected service time of the patient scheduled in the $j-$th appointment slot for $j \leq i$. Using these deterministic values, one can calculate the arrival times $\tilde{A}_j$ and departure times $\tilde{D}_j$ of patients $j = 1, \ldots, i$ using Lindley's equations:

$$\tilde{D}_j = \tilde{A}_j + \tilde{W}_j + \bar{s}_j,$$

where $\tilde{A}_j = a_j + \bar{u}_j$, $\tilde{W}_1 = -\min\{0, \tilde{A}_1\}$, $\tilde{W}_j = \max\{0, \tilde{D}_{j-1} - \tilde{A}_j\}$ for $j = 2, \ldots, i$.

If we define

$$N_i = \sum_{j=1}^{i} \mathbb{I}\{\tilde{A}_j \leq a_i\}\mathbb{I}\{\tilde{D}_j > a_i\},$$

then N_i is the number of people in the system at time a_i under the deterministic unpunctuality and service times. We then set $k_i = N_i + 1$ as the weighting factor for scheduling appointment a_{i+1}.

Using our above methodology, we propose the following Myopic Scheduling Heuristic that uses a fixed patient sequence $\boldsymbol{q} = (q_1, q_2, \ldots, q_P)$:

Algorithm 2.6.1. Myopic Scheduling Heuristic

Initial Input: $\mathbf{q} = (q_1, q_2, \ldots, q_P)$

Set $a_1 = \max\left\{0, -F^{-1}_{U_{q_1}}\left(\frac{c_w}{c_w + c_I}\right)\right\}$

for i=2, ..., P **do**

Calculate $k_i = N_i + 1$

Set $a_i = a_{i-1} + \max\left\{0, F^{-1}_{R_{q_{i-1},q_i}}\left(\frac{k_i c_w}{k_i c_w + c_I}\right)\right\}$

end for

2.7 Data-driven Numerical Studies

We use the clinic dataset to develop several scenarios for testing. In particular, we use individual patient distributions for generating unpunctuality and service times for generating schedules and testing our heuristics. Patients are selected as the 12 patients with the most amount of available data post-filtering. The filtering process involves removing any observations with unpunctuality with magnitude greater than 2 hours or service times beyond 3 hours as well as observations missing

Patient No.	U Mean	U SD	S Mean	S SD
1	-22.8274	29.5802	59.7859	45.3275
2	-44.4541	42.4770	52.0795	46.3084
3	-14.7445	37.0057	64.2493	46.6515
4	-37.4505	31.0536	55.3563	48.1651
5	-2.4112	19.5408	68.5667	49.6875
6	-22.4285	36.9639	55.2124	50.9274
7	-22.6192	31.3403	46.3822	53.6960
8	-35.0116	31.6447	41.0520	54.1557
9	-41.9465	25.9345	59.6438	54.6848
10	-20.0375	29.5642	61.5680	56.6795
11	-15.6299	31.6945	68.7111	57.2360
12	34.8698	46.2589	81.8488	58.5234

Table 2.1: Summary statistics of 12 patients used in the numerical studies.

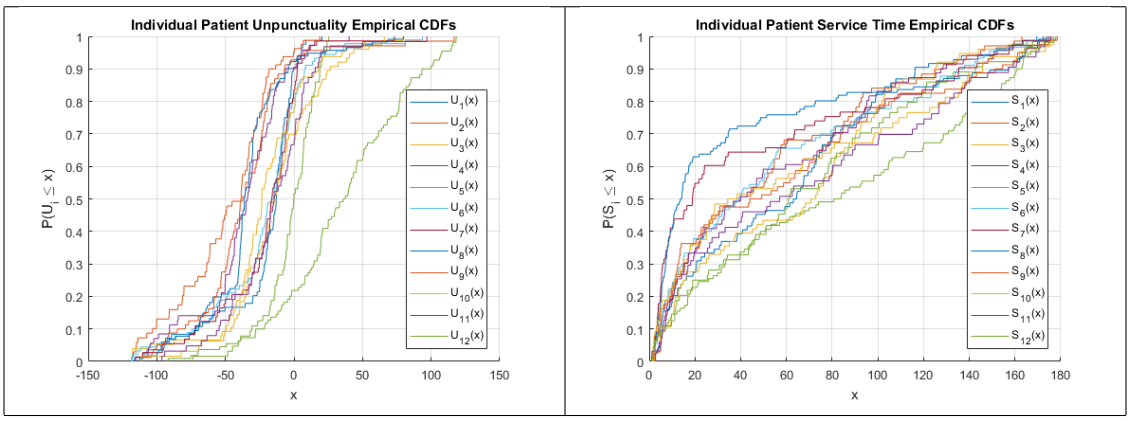

Figure 2.6: Empirical CDFs of the unpunctuality and service time distributions of the 12 patients used in the numerical studies.

either an unpunctuality or a service time. We present the means and standard deviations for the unpunctuality (U Mean and U SD, respectively) and service times (S Mean and S SD, respectively) of these 12 patients in Table 2.1, listed in the order of increasing variance of the service time. Figure 2.6 presents the empirical distributions of the unpunctuality and service times of the patients in the table. All time representations are in minutes.

The purpose of these scenarios is to picture how well we can perform when scheduling and sequencing patients by leveraging information on each of the patients' unpunctuality and service time distributions. In terms of scheduling methodologies, we propose several baseline heuristics in addi-

tion to our Myopic Scheduling heuristic and AIPR. The performance of the different appointment schedules is evaluated via simulation and presented with confidence intervals.

These scenarios cover different cost parameter values and service disciplines using a shared time horizon of $T = 720$ minutes. The different heuristics are listed below:

- Expected Service (ES): an appointment schedule derived from the heuristic that spaces out appointments according to expected service time: $a_1 = 0$ and $a_{i+1} = a_i + E(S_{q_i})$ for $i = 1, \ldots, P-1$.

- Expected Service and Expected Unpunctuality (ESEU): an appointment scheduled derived from the heuristic that spaces out appointments according to their expected service time, then additionally perturbs them by the expected unpunctuality: $a_1 = -E(U_{q_1})$ and $a_{i+1} = a_i + E(S_{q_i}) - E(U_{q_{i+1}})$ for $i = 1, \ldots, P-1$. Note: any negative appointment times are rounded up to 0 and any appointments beyond T are rounded down to T to keep all appointments within the scheduling window.

- Myopic Scheduling (MS): an appointment schedule derived via the Myopic Scheduling heuristic.

- Iterative Perturbation (IP): an appointment schedule derived via Iterative Perturbation.

- Alternating Iterative Perturbation and Resequencing (AIPR): an appointment schedule derived using the AIPR heuristic.

Each heuristic is associated with a single schedule per scenario. The different scenarios lead to different schedule outcomes for MS, IP, and AIPR as, even if the same heuristic was used, the heuristics depend on the cost parameters and/or service discipline unique to each scenario. Using IP (Iterative Perturbation) is presented as a separate scheduling method as it is simply a type of local search heuristic commonly used to solve these types of problems (Deceuninck and Vuyst, 2018). All simulation-optimization methods (IP and AIPR methods) have schedules derived using 2000 sample paths independently drawn from the patient distributions.

The only scheduling rules that did not incorporate the cost parameters into the derivation are ES and ESEU. However, these methods were incorporated as they are simple, easily justifiable

rules. ES acted as our baseline for examining percent improvements, as it was the most simple, "obvious" policy we could use.

2.7.1 Performance of the Heuristics

We first compared heuristics in the fully individualized patient characteristics setting. In this situation, we examined 4 cases for comparing heuristics:

Case 1: $c_b = c_a = 1$, $c_I \in \{1, 2, 5, 10\}$, $1.5c_I = c_o$, ABP service discipline.

Case 2: $c_b = 0, c_a = 1$, $c_I \in \{1, 2, 5, 10\}$, $1.5c_I = c_o$, ABP service discipline.

Case 3: $c_b = c_a = 1$, $c_I \in \{1, 2, 5, 10\}$, $1.5c_I = c_o$, ELH service discipline.

Case 4: $c_b = 0, c_a = 1$, $c_I \in \{1, 2, 5, 10\}$, $1.5c_I = c_o$, ELH service discipline.

Case 1 is the most basic scenario, which follows the "fair" ABP service discipline and does not distinguish between before-appointment wait time and after-appointment wait time. Case 2 keeps the ABP service discipline, but assigns zero cost to the before-appointment wait times. Case 3 and 4 mirror Case 1 and 2 respectively in terms of the cost structure, but switch to the ELH service discipline.

We evaluated the appointment schedules derived via various heuristics across 100,000 simulations per case. Mean total cost and their confidence intervals are presented across several tables. The acronym ESPI stands for the percent improvement over ES which will be presented with respect to mean values.

Tables 2.2 and 2.3 summarize Cases 1 and 2 respectively. We saw that the AIPR heuristic performed the best, but only marginally better than IP. That improvement gap appeared to increase as the idle cost of the system increased. This was, however, the simplest scenario and we expected patient sequences to play a much bigger role as we added further complication to the system. The MS heuristic provided modest improvements on ES; however, it should be noted that while the derivation of the MS heuristic does not incorporate overtime costs, the actual scenarios themselves do include overtime calculations.

Tables 2.4 and 2.5 present results from Cases 3 and 4 respectively. It is interesting to note that, in spite of the fact that MS's derivation is a serious departure from the service discipline used

c_I	Method	Mean TC (ESPI)	95% CI
1	ES	1178.2 (0%)	(1173.9, 1182.6)
	ESEU	1154.2 (2.04%)	(1149.8, 1158.6)
	MS	1114.5 (5.41%)	(1111.7, 1117.3)
	IP	1024.8 (13.02%)	(1021.7, 1027.9)
	AIPR	1019.7 (13.45%)	(1016.7, 1022.7)
2	ES	1467.4 (0%)	(1462.5, 1472.2)
	ESEU	1464.8 (0.18%)	(1459.7, 1469.9)
	MS	1440.9 (1.81%)	(1435.8, 1445.9)
	IP	1402.6 (4.42%)	(1398.0, 1407.3)
	AIPR	1396.8 (4.81%)	(1392.2, 1401.4)
5	ES	2331.4 (0%)	(2324.4, 2338.4)
	ESEU	2372.9 (-1.78%)	(2365.4, 2380.4)
	MS	2328.1 (0.14%)	(2318.5, 2337.6)
	IP	2098.0 (10.01%)	(2090.2, 2105.9)
	AIPR	2087.7 (10.45%)	(2079.8, 2095.7)
10	ES	3779.1 (0%)	(3767.7, 3790.6)
	ESEU	3901.8 (-3.25%)	(3889.5, 3914.1)
	MS	3514.7 (7.00%)	(3499.5, 3530.0)
	IP	2959.0 (21.70%)	(2946.0, 2971.9)
	AIPR	2915.4 (22.85%)	(2902.0, 2928.7)

Table 2.2: Simulation Results for Heuristic Comparison for Case 1: even wait costs and ABP service discipline.

c_I	Method	Mean TC (ESPI)	95% CI
1	ES	942.3 (0%)	(938.3, 946.4)
	ESEU	956.2 (-1.48%)	(952.0, 960.3)
	MS	974.5 (-3.42%)	(971.9, 977.1)
	IP	833.2 (11.58%)	(830.3, 836.1)
	AIPR	829.0 (12.02%)	(826.2, 831.8)
2	ES	1235.2 (0%)	(1230.5, 1239.9)
	ESEU	1262.4 (-2.20%)	(1257.6, 1267.2)
	MS	1229.6 (0.45%)	(1224.8, 1234.5)
	IP	1178.6 (4.58%)	(1174.1, 1183.1)
	AIPR	1178.6 (4.58%)	(1174.1, 1183.1)
5	ES	2108.9 (0%)	(2101.3, 2115.3)
	ESEU	2171.8 (-2.98%)	(2164.4, 2179.2)
	MS	2047.4 (2.92%)	(2037.9, 2056.8)
	IP	1841.5 (12.68%)	(1833.6, 1849.5)
	AIPR	1832.8 (13.09%)	(1824.9, 1840.8)
10	ES	3557.6 (0%)	(3546.1, 3569.1)
	ESEU	3712.8 (-4.36%)	(3700.4, 3725.1)
	MS	3206.8 (9.86%)	(3191.6, 3222.0)
	IP	2703.1 (24.02%)	(2690.3, 2716.0)
	AIPR	2646.3 (25.62%)	(2632.9, 2659.7)

Table 2.3: Simulation Results for Heuristic Comparison for Case 2: uneven wait costs and ABP service discipline.

c_I	Method	Mean TC (ESPI)	95% CI
1	ES	881.9 (0%)	(879.0, 884.9)
	ESEU	840.1 (4.74%)	(837.2, 843.1)
	MS	929.7 (-5.42%)	(927.6, 931.8)
	IP	805.2 (8.70%)	(802.5, 807.9)
	AIPR	805.2 (8.70%)	(802.5, 807.9)
2	ES	1172.4 (0%)	(1168.9, 1175.9)
	ESEU	1145.8 (2.27%)	(1142.2, 1149.4)
	MS	1132.8 (3.38%)	(1129.0, 1136.5)
	IP	1053.1 (10.18%)	(1049.3, 1056.8)
	AIPR	1046.7 (10.72%)	(1042.9, 1050.6)
5	ES	2035.4 (0%)	(2029.6, 2041.3)
	ESEU	2058.4 (-1.13%)	(2052.2, 2064.7)
	MS	1649.4 (18.96%)	(1641.9, 1656.9)
	IP	1603.0 (21.24%)	(1596.5, 1609.5)
	AIPR	1485.7 (27.01%)	(1479.0, 1492.4)
10	ES	3487.8 (0%)	(3477.3, 3498.3)
	ESEU	3589.9 (-2.93%)	(3578.6, 3601.1)
	MS	2534.7 (27.33%)	(2521.7, 2547.8)
	IP	2338.8 (32.94%)	(2327.6, 2350.0)
	AIPR	2125.9 (39.05%)	(2114.2, 2137.5)

Table 2.4: Simulation Results for Heuristic Comparison for Case 3: even wait costs and ELH service discipline.

in Cases 3 and 4 (ELH), the MS heuristic performs extremely well as the idle cost grew larger. However, in low idle cost cases, such as $c_I = 1$, the MS heuristic performs quite poorly, in one case performing approximately 20% worse than the ES heuristic (Case 4, $c_I = 1$). This was an unexpected development and it warrants caution for testing to determine what types of service disciplines might work well with the MS heuristic.

As expected, however, the IP and AIPR heuristic performed the best. It is also interesting to note that in this service discipline, as idle cost increases, the performance improvement gap between IP and AIPR became significant, which captures the importance of the resequencing aspect of the optimization process in more complicated systems, such as when an unpunctuality-dependent prioritization scheme is implemented.

2.7.2 Significance of Unpunctuality

As we discussed in Section 1.1, prior work on appointment scheduling has largely ignored unpunctuality and an investigation of how it should be taken into account is one of the main

c_I	Method	Mean TC (ESPI)	95% CI
1	ES	656.7 (0%)	(654.1, 659.4)
	ESEU	650.3 (0.97%)	(647.6, 653.0)
	MS	790.7 (-20.41%)	(788.8, 792.6)
	IP	592.5 (9.78%)	(590.0, 594.9)
	AIPR	592.5 (9.78%)	(590.0, 594.9)
2	ES	947.7 (0%)	(944.4, 951.0)
	ESEU	957.1 (-0.99%)	(953.7, 960.5)
	MS	927.1 (2.17%)	(923.6, 930.6)
	IP	786.5 (17.01%)	(782.9, 790.2)
	AIPR	786.5 (17.01%)	(782.9, 790.2)
5	ES	1813.7 (0%)	(1807.9, 1819.5)
	ESEU	1873.1 (-3.28%)	(1866.9, 1879.2)
	MS	1372.3 (24.34%)	(1365.0, 1379.6)
	IP	1319.5 (27.25%)	(1313.2, 1325.9)
	AIPR	1230.1 (32.18%)	(1223.4, 1236.8)
10	ES	3258.1 (0%)	(3247.7, 3268.6)
	ESEU	3401.7 (-4.41%)	(3390.5, 3413.0)
	MS	2216.6 (31.97%)	(2203.7, 2229.6)
	IP	2077.9 (36.22%)	(2066.8, 2088.9)
	AIPR	1863.1 (42.82%)	(1851.5, 1874.6)

Table 2.5: Simulation Results for Heuristic Comparison for Case 4: uneven wait costs and ELH service discipline.

contributions of this paper. In Section 2.7.1, we found the the methods we developed significantly improve upon the simple alternatives. However, this does answer the question of whether capturing unpunctuality is worthwhile in the first place because one could have also used the IP and AIPR heuristics without taking into account patients' unpunctuality behavior. To investigate the worth of incorporating unpunctuality, in this section, we run the IP and AIPR methods under three different settings: one that assumes all patients are punctual, one that assumes patients are unpunctual, but the behavior is homogeneous across all patients, and one that assumes that each patient has their own unique unpunctual behavior that can be estimated based on their own historical behavior. Specifically, in the following we present the performances under six different heuristics. IP-ZU, IP-HU, and IP-IU all use the IP heuristic, but respectively under the assumptions of "zero unpunctuality", "homogeneous unpunctuality", and "individual unpunctuality". AIPR-ZU, AIPR-HU, and AIPR-IU make the same respective assumptions on unpunctuality but use the AIPR heuristic instead. Note that going from ZU to HU and then to IU, unpunctuality is captured at increasingly more granular levels. What is mainly of interest here is how much can we improve upon ZU policies using HU or IU policies, which would reveal the potential benefits of capturing unpunctuality, and how much one can improve upon HU policies by using IU policies, which would reveal the potential benefits of using unpunctuality data at the individual level.

Table 2.6 summarizes the mean and the 95% confidence intervals (CI) of the total costs (TC) associated with the 3 different treatments of unpunctuality for Case 1. ZUPI stands for "zero unpunctuality percent improvement", which is the percent cost improvement relative to the ZU schedule derived under the matching heuristic.

It is important to note that, under AIPR-ZU and AIPR-HU, resequencing is still performed relative to the individualized service distributions; however, unpunctuality is treated as homogeneous in both cases. This leads to AIPR methods under-performing relative to IP in the true heterogeneous setting. This is important because most prior research assumes homogeneous unpunctuality and, if resequencing is done strictly with respect to service times without accounting for heterogeneous unpunctuality, a worse solution may be reached as an incorrect distributional assumption is driving the optimization.

Case 2 results are provided in Table 2.7. As with Case 1, we saw improvements of around 2-3 percent when accounting for heterogeneous unpunctuality.

c_I	Method	Mean TC (ZUPI)	95% CI
1	IP-ZU	1054.8 (0%)	(1051.6, 1058.1)
	IP-HU	1047.7 (0.67%)	(1044.5, 1050.9)
	IP-IU	1024.8 (2.84%)	(1021.7, 1027.9)
2	IP-ZU	1455.1 (0%)	(1450.5, 1459.8)
	IP-HU	1425.6 (2.03%)	(1421.2, 1430.0)
	IP-IU	1402.6 (3.61%)	(1398.0, 1407.3)
5	IP-ZU	2164.2 (0%)	(2156.2, 2172.2)
	IP-HU	2122.0 (1.95%)	(2114.0, 2129.9)
	IP-IU	2098.0 (3.06%)	(2090.2, 2105.9)
10	IP-ZU	3084.1 (0%)	(3071.7, 3096.4)
	IP-HU	2996.3 (2.85%)	(2983.7, 3008.9)
	IP-IU	2959.0 (4.06%)	(2946.0, 2971.9)
1	AIPR-ZU	1057.4 (0%)	(1054.1, 1060.6)
	AIPR-HU	1061.6 (-0.40%)	(1058.5, 1064.6)
	AIPR-IU	1019.7 (3.57%)	(1016.7, 1022.7)
2	AIPR-ZU	1457.0 (0%)	(1452.3, 1461.6)
	AIPR-HU	1424.8 (2.21%)	(1420.3, 1429.2)
	AIPR-IU	1396.8 (4.13%)	(1392.2, 1401.4)
5	AIPR-ZU	2168.8 (0%)	(2160.7, 2176.9)
	AIPR-HU	2111.0 (2.67%)	(2103.1, 2119.0)
	AIPR-IU	2087.7 (3.74%)	(2079.8, 2095.7)
10	AIPR-ZU	3002.5 (0%)	(2989.0, 3015.9)
	AIPR-HU	2982.3 (0.67%)	(2969.0, 2995.6)
	AIPR-IU	2915.4 (3.35%)	(2902.0, 2928.7)

Table 2.6: Simulation Results for Unpunctuality Comparison for Case 1: even wait costs and ABP service discipline.

c_I	Method	Mean TC (ZUPI)	95% CI
1	IP-ZU	846.9 (0%)	(844.0, 849.9)
	IP-HU	854.0 (-0.84%)	(851.0, 856.9)
	IP-IU	833.2 (1.62%)	(830.3, 836.1)
2	IP-ZU	1215.4 (0%)	(1211.0, 1219.8)
	IP-HU	1213.6 (0.15%)	(1209.2, 1218.0)
	IP-IU	1178.6 (3.03%)	(1174.1, 1183.1)
5	IP-ZU	1889.4 (0%)	(1881.6, 1897.3)
	IP-HU	1859.1 (1.60%)	(1851.2, 1866.9)
	IP-IU	1841.5 (2.54%)	(1833.6, 1849.5)
10	IP-ZU	2797.9 (0%)	(2785.7, 2810.1)
	IP-HU	2732.2 (2.35%)	(2719.7, 2744.6)
	IP-IU	2703.1 (3.39%)	(2690.3, 2716.0)
1	AIPR-ZU	854.0 (0%)	(851.0, 857.0)
	AIPR-HU	850.4 (0.42%)	(847.5, 853.4)
	AIPR-IU	829.0 (2.93%)	(826.2, 831.8)
2	AIPR-ZU	1212.4 (0%)	(1208.0, 1216.8)
	AIPR-HU	1208.4 (0.33%)	(1208.0, 1216.8)
	AIPR-IU	1178.6 (2.79%)	(1174.1, 1183.1)
5	AIPR-ZU	1899.8 (0%)	(1891.8, 1907.8)
	AIPR-HU	1852.2 (2.51%)	(1844.0, 1860.5)
	AIPR-IU	1832.8 (3.53%)	(1824.9, 1840.8)
10	AIPR-ZU	2709.0 (0%)	(2695.6, 2722.3)
	AIPR-HU	2662.8 (1.71%)	(2649.6, 2675.9)
	AIPR-IU	2646.3 (2.31%)	(2632.9, 2659.7)

Table 2.7: Simulation Results for Unpunctuality Comparison for Case 2: uneven wait costs and ABP service discipline.

c_I	Method	Mean TC (ZUPI)	95% CI
1	IP-ZU	875.4 (0%)	(873.1, 877.6)
	IP-HU	862.7 (1.45%)	(860.0, 865.4)
	IP-IU	805.2 (8.02%)	(802.5, 807.9)
2	IP-ZU	1183.9 (0%)	(1180.6, 1187.2)
	IP-HU	1108.0 (6.41%)	(1104.1, 1111.8)
	IP-IU	1053.1 (11.05%)	(1049.3, 1056.8)
5	IP-ZU	1730.1 (0%)	(1723.7, 1736.4)
	IP-HU	1635.0 (5.50%)	(1628.5, 1641.5)
	IP-IU	1603.0 (7.35%)	(1596.5, 1609.5)
10	IP-ZU	2619.4 (0%)	(2608.6, 2630.2)
	IP-HU	2405.9 (8.15%)	(2394.9, 2417.0)
	IP-IU	2338.8 (10.71%)	(2327.6, 2350.0)
1	AIPR-ZU	875.4 (0%)	(873.1, 877.7)
	AIPR-HU	851.2 (2.76%)	(848.5, 853.8)
	AIPR-IU	805.2 (8.02%)	(802.5, 807.9)
2	AIPR-ZU	1179.0 (0%)	(1175.7, 1182.4)
	AIPR-HU	1123.9 (4.67%)	(1119.8, 1127.9)
	AIPR-IU	1046.7 (11.22%)	(1042.9, 1050.6)
5	AIPR-ZU	1725.8 (0%)	(1719.4, 1732.2)
	AIPR-HU	1599.5 (7.32%)	(1592.5, 1606.5)
	AIPR-IU	1485.7 (14.91%)	(1479.0, 1492.4)
10	AIPR-ZU	2375.3 (0%)	(2363.9, 2386.6)
	AIPR-HU	2221.1 (6.49%)	(2209.3, 2232.9)
	AIPR-IU	2125.9 (10.50%)	(2114.2, 2137.5)

Table 2.8: Simulation Results for Unpunctuality Comparison for Case 3: even wait costs and ELH service discipline.

Case 3 and Case 4 scenarios used the ELH service discipline. Their unpunctuality comparison results are presented in Table 2.8 and Table 2.9. Since this discipline changed the cost structure depending on whether a patient is early or late, we found that using individualized unpunctuality leads to higher improvements against homogeneous unpunctuality than in Cases 1 and 2. That said, optimization with respect to homogeneous unpunctuality still lead to significant improvements.

In summary, we concluded that incorporating homogeneous unpunctuality is better than not including it and including individualized unpunctuality provides the best results, if that data is available.

A plot of the inter-appointment gaps from three appointment schedules derived via different unpunctuality types for Case 1, $c_I = 5$, can be seen in Figure 2.7. For all heuristics, the intra-appointment gaps seem to alternate between large and small values. In particular, for fully-

c_I	Method	Mean TC (ZUPI)	95% CI
1	IP-ZU	673.2 (0%)	(671.2, 675.2)
	IP-HU	654.0 (2.85%)	(651.5, 656.4)
	IP-IU	592.5 (11.99%)	(590.0, 594.9)
2	IP-ZU	950.4 (0%)	(947.3, 953.5)
	IP-HU	816.7 (14.07%)	(813.1, 820.4)
	IP-IU	786.5 (17.25%)	(782.9, 790.2)
5	IP-ZU	1471.3 (0%)	(1465.1, 1477.5)
	IP-HU	1406.4 (4.41%)	(1400.0, 1412.8)
	IP-IU	1319.5 (10.32%)	(1313.2, 1325.9)
10	IP-ZU	2358.0 (0%)	(2347.2, 2368.7)
	IP-HU	2183.2 (7.41%)	(2172.2, 2194.3)
	IP-IU	2077.9 (11.88%)	(2066.8, 2088.9)
1	AIPR-ZU	677.5 (0%)	(675.5, 679.5)
	AIPR-HU	646.6 (4.56%)	(644.1, 649.0)
	AIPR-IU	592.5 (12.54%)	(590.0, 594.9)
2	AIPR-ZU	949.7 (0%)	(946.5, 952.8)
	IP-HU	816.7 (14.07%)	(813.1, 820.4)
	AIPR-IU	786.5 (17.18%)	(782.9, 790.2)
5	AIPR-ZU	1463.1 (0%)	(1456.8, 1469.3)
	AIPR-HU	1323.0 (9.58%)	(1316.1, 1329.8)
	AIPR-IU	1230.1 (15.93%)	(1223.4, 1236.8)
10	AIPR-ZU	2121.6 (0%)	(2110.3, 2133.0)
	AIPR-HU	1994.3 (6.00%)	(1982.3, 2006.2)
	AIPR-IU	1863.1 (12.18%)	(1851.5, 1874.6)

Table 2.9: Simulation Results for Unpunctuality Comparison for Case 4: uneven wait costs and ELH service discipline.

individualized unpunctuality, some of the gaps are at or near zero, implying small blocks of patients. This is different than the dome-shaped curves expected in the literature without unpunctuality.

A plot of the inter-appointment gaps for three appointment schedules derived from different unpunctuality types for Case 3, $c_I = 2$, can be seen in Figure 2.8. We noticed a similar trend to our earlier inter-appointment plot in that using individualized unpunctuality tended the schedule towards a more block-ish appearance, with higher and lower peaks than the other schedules.

2.8 Conclusion

In this chapter we developed several heuristic methods for cost-effective scheduling of patients under several cost models and service disciplines. In particular we considered two service disciplines and five different heuristics. We used real data to evaluate and compare these heuristics under various appointment scheduling scenarios. Since our highest-performing method is simulation-based, it can easily be extended to further patient heterogeneous characteristics such as no-show probabilities and late-cancellation probabilities by generalizing the simulation code. The most effective algorithm for reducing the total expected cost of the system was the AIPR heuristic that involves a local search (IP) heuristic paired with a separate resequencing heuristic that leverages the individual characteristics of patients. However, just using the IP heuristic alone still lead to high-performing solutions. Our MS heuristic also performed surprisingly well on complex systems, even if those systems do not resemble the analytical formulation used to derive MS. However, we caution practitioners to first examine how the MS heuristic performs via simulation on complex scheduling systems before implementation. We demonstrated that our simulation optimization heuristics can perform quite differently depending on the service discipline used for the system, with the resequencing heuristic likely playing a greater part in producing better solutions if priority classes are involved in the service discipline.

Our last conclusion is that incorporating individualized unpunctuality can lead to significant cost benefits. Of course, incorporating individualized unpunctuality requires availability and access

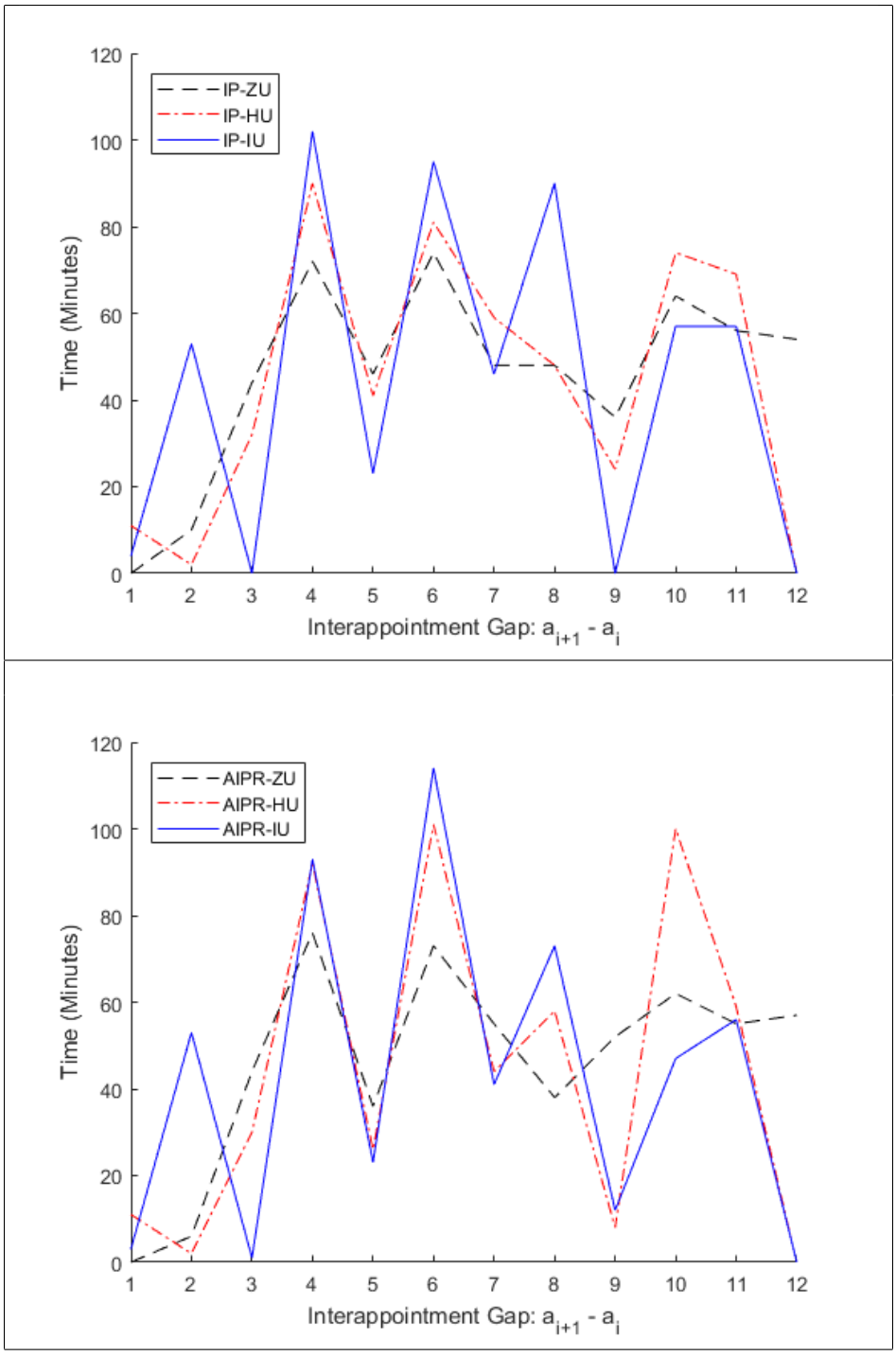

Figure 2.7: Inter-appointment plots for Case 1, $c_I = 5$ for appointments derived via IP and AIPR methods across 3 unpunctuality cases.

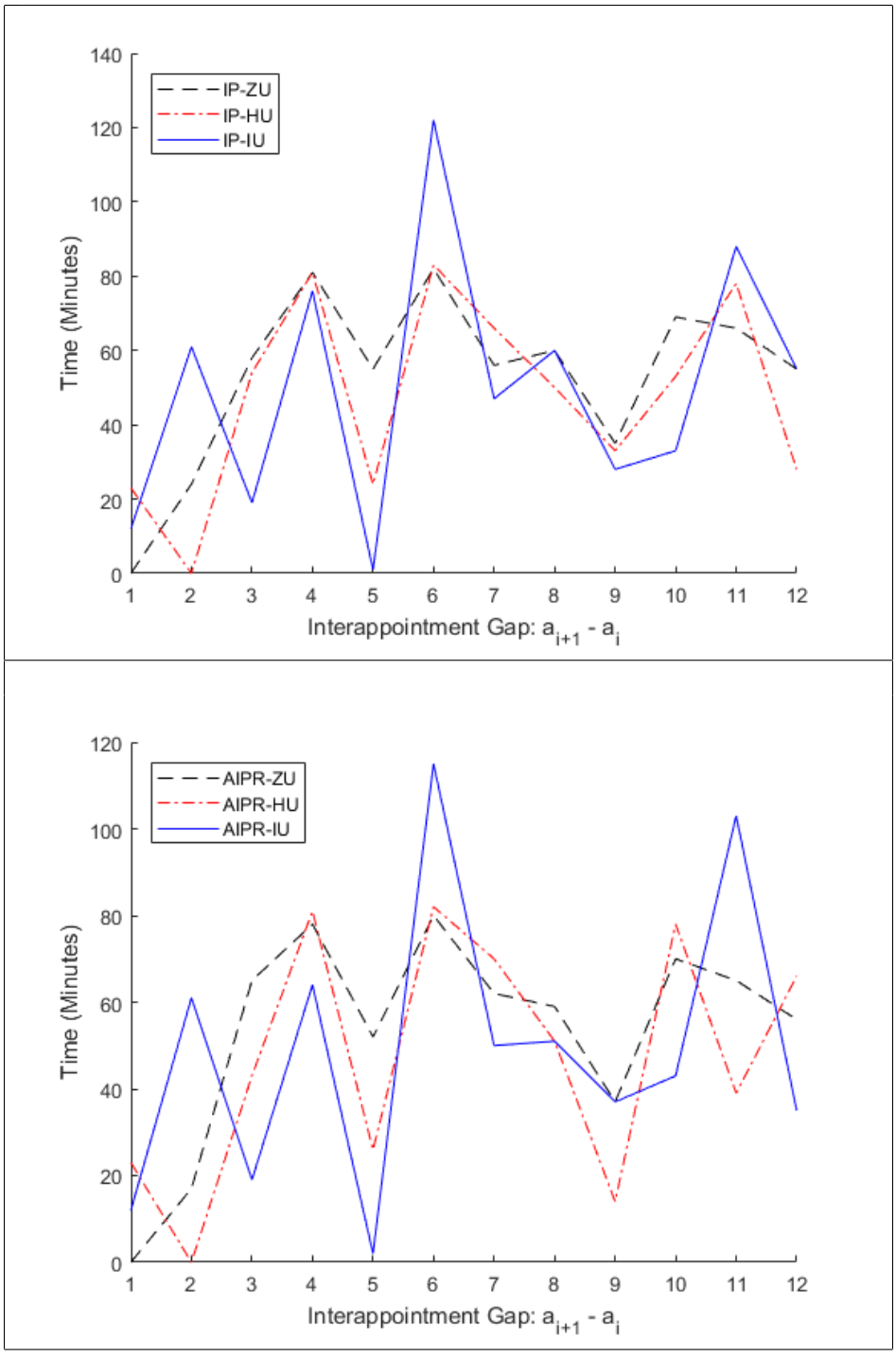

Figure 2.8: Inter-appointment plots for Case 3, $c_I = 2$ for appointments derived via IP and AIPR methods across 3 unpunctuality cases.

to a large amount of patient-specific data. If there is not enough data to estimate the individual unpunctuality distributions, one can pool the data and estimate a patient-independent homogeneous unpunctuality distribution. Regardless, incorporating homogeneous unpunctuality is better than ignoring the unpunctuality altogether.

CHAPTER 3

Asymptotically Optimal Appointment Schedules in the Presence of Unpunctuality Using Fluid Limits

3.1 Motivation

In the previous chapter, we developed various heuristics for solving a complex, nonconvex optimization problem. The best performing heuristics, however, do not scale well as the number of patients increase. In high-capacity clinic settings, such as vaccine clinics, the number of patients seen within a single day can be in the hundreds, making our simulation-based heuristics take far too long to run. However, in a high-capacity clinic, we can expect certain things: patients are more-or-less the same, allowing us to treat them as homogeneous in both unpunctuality and service time. Further, since service times will be shorter, a first-in first-out (FIFO) service discipline will generally be deemed acceptable by visiting patients. In light of these assumptions, we consider a method for deriving high-performance schedules when the number of patients is large. This method depends on the fluid limits of the arrival, departure, queueing, and idle-accumulation processes associated with an appointment system. Upon developing a method for finding asymptotically optimal appointment schedules in the fluid limit, we examine the performance of these schedules by using real clinic data.

3.2 The Model

We consider a clinic that operates over $[0, T]$ and schedules p patients to be seen over this time horizon. Let a_i be the scheduled appointment time of the i-th patient. Without loss of generality, assume

$$0 \leq a_1 \leq a_2 \leq \cdots \leq a_p \leq T.$$

Due to patient unpunctuality the ith scheduled patient actually arrives at time

$$T_i = a_i + U_i, \quad 1 \leq i \leq p, \tag{3.1}$$

where U_i is a random variable representing the unpunctuality of the ith patient. If $U_i < 0$, the patient arrives early for the scheduled appointment, if $U_i > 0$, the patient is late, and if $U_i = 0$, the patient is on time. The unpunctuality values $\{U_i\}_{i=1}^{p}$ are assumed to be independent random variables, and U_i has a distribution given by

$$F(t, a) = P(U_i \leq t | a_i = a), \quad t \in .$$

Thus, the unpunctuality of patient i can depend on its scheduled appointment time a_i. Let $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(p)}$ be the order statistics of the p actual arrival times. We observe that $T_{(i)}$ is the time of the ith arrival to the clinic.

We note that the arrival times could be negative, i.e., $T_i < 0$ with nonnegative probability. This happens when patients arrive before the clinic opens at time $t = 0$. Let $E(t)$ be the cumulative number of arrivals over $(-\infty, t]$, which can be formulated as

$$E(t) = \sum_{i=1}^{p} 1_{\{T_i \leq t\}}, \quad t \in .$$

All patients are served by a single doctor in the order of their arrivals (i.e., FIFO service discipline). Let ν_i be the service time of patient i. We assume that $\{\nu_i, i \in\}$ is an i.i.d. sequence with common mean $1/\mu$ and standard deviation σ, and a CDF G with support $[0, \infty)$. Let $\{S(t), t \geq 0\}$ be the renewal process generated by $\{\nu_i, i \in\}$, i.e.,

$$S(t) = \sum_{i=1}^{p} 1_{\{\nu_i \leq t\}}, \quad t \geq 0.$$

We note that $S(t)$ is the total number of departures over $[0, t]$ if the doctor is always busy. For $t \leq 0$, we define $S(t) = 0$. We will also denote μT as the *capacity* of the system; however, depending on a choice of p, a system can be booked over or under capacity.

Let $Q(t)$ be the number of customers in the system at time $t \in$. We assume that the patients who arrive after time T will not be allowed to enter the system. Then we have that for $t \in$,

$$Q(t) = \begin{cases} E(t \wedge T) - S(B(t)), & t \geq 0, \\ E(t), & t < 0, \end{cases} \tag{3.2}$$

where for $t \geq 0$,

$$B(t) = \int_0^t 1_{\{Q(s)>0\}} ds$$

representing the cumulative busy time of the server over $[0, t]$. The idle time of the server over $[0, t]$ can then be formulated as

$$I(t) = t - B(t), \;\; t \geq 0.$$

Finally, let $V(t)$ denote the queueing time of a customer (time spent in the system before starting service) who happens to arrive at time t. The process $\{V(t); t \geq 0\}$ is referred to as the virtual queueing time process. Under the FIFO service discipline, we have

$$V(t) = \begin{cases} \sum_{j=1}^{E(t)} \nu_j - B(t) & 0 \leq t \leq T, \\ \sum_{j=1}^{E(t)} \nu_j - t & t < 0. \end{cases} \tag{3.3}$$

Let O be a nonnegative random variable representing the amount of overtime it takes to empty the system, i.e.,

$$O(T) = \inf\{t \geq T : Q(t) = 0\} - T$$

Our goal is to determine the number of patients p and an optimal sequence of appointment times $\{a_i\}_{i=1}^p$ to maximize the profit of the clinic. Let r denote the reward of serving each patient, c_w the waiting time cost rate of each patient, c_i the idle time cost rate, and c_o the overtime cost

rate. Our optimization problem is:

$$\begin{aligned} \max J(p, \{a_i\}_{i=1}^{p}) = \quad & rE(T) - c_w \int_0^\infty Q(t)dt - c_i I(T) - c_o O(T) \\ \text{subject to} \quad & p \in \mathbb{Z}^+ \\ & 0 \le a_1 \le a_2 \le \cdots \le a_p \le T. \end{aligned} \tag{3.4}$$

This is a highly non-linear optimization problem, and solving it to optimality is intractable, as demonstrated in the previous chapter.

3.3 Heavy Traffic Fluid Limit

In this section, we study the asymptotic behavior of the system described in Section 3.2 under fluid scaling. To be precise, we consider a sequence of systems indexed by natural number n as follows. The service times in the nth system are iid random variables with mean $1/\mu^n$ and variance σ^n. All other parameters remain the same as in the previous section. We assume that the system capacity in the nth system approaches infinity as n approaches infinity. To be precise, we assume that

$$\lim_{n\to\infty} \frac{\mu^n}{n} = \mu, \quad \text{and} \quad \lim_{n\to\infty} \frac{\sigma^n}{n} = \sigma, \tag{3.5}$$

where μ and σ are positive constants. We consider the following processes defined in Section 3.3 under fluid scaling:

$$\bar{Q}^n(t) = Q(nt)/n, \ \bar{E}^n(t) = E(nt)/n, \ \bar{S}^n(t) = S(nt)/n, \ \bar{V}^n(t) = V(nt)/n, \tag{3.6}$$

$$B^n(t) = \frac{1}{n} \int_0^{nt} 1_{Q(s)>0} ds,$$

and

$$I^n(t) = \frac{nt - B^n(t)}{n}.$$

Next, we define the relative frequency process for the appointment times:

$$\bar{A}^n(t) = \begin{cases} 0, & t < 0, \\ \frac{1}{n}\sum_{i=1}^{n} 1_{\{a_i \le t\}}, & t \in [0, T], \\ 1, & t > T. \end{cases} \tag{3.7}$$

Assumptions: We make the following mathematical assumptions for our model:

1. We assume that 0 is in the interior of the convex hull for the support of the unpunctuality random variable for any patient. That is, $0 < F(0, a) < 1$ for any $a \in [0, T]$.

2. We assume that the function $F(t, a)$ is piecewise continuous in a and for each piece it is continuous uniformly for $t \in$. More precisely, there exists a finite (deterministic) partition $0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = T$ such that $\sup_{t\in} |F(t, a_1) - F(t, a_2)| \to 0$ as $|a_1 - a_2| \to 0$ when a_1 and a_2 are within any partition interval $[\tau_i, \tau_{i+1}), i = 0, \ldots, k-1$ or $[\tau_K, \tau_{K+1}]$.

3. We assume there exists a deterministic function $A(t), t \in [0, T]$ such that

$$\sup_{t\in[0,T]} \left|\bar{A}^n(t) - A(t)\right| \to 0.$$

   Let $A(t) = 0$ for $t < 0$ and $A(t) = 1$ for $t > T$. Then $\bar{A}^n(t) \to A(t)$ uniformly for $t \in$ and $\{A(t), t \in\}$ is a CDF (see the proof of Lemma 3.3.1).

Finally, we introduce the following Stieltjes-convolution of $F(\cdot, \cdot)$ and $A(\cdot)$:

$$H(t) = \int_{-\infty}^{\infty} F(t - s, s)dA(s) = \int_{0}^{T} F(t - s, s)dA(s), \quad t \in . \tag{3.8}$$

We have the following fluid limits.

Lemma 3.3.1. As $n \to \infty$,

$$\sup_{t\in} |\bar{E}^n(t) - H(t)| \to 0, \text{ almost surely.} \tag{3.9}$$

Proposition 3.3.1. As $n \to \infty$, for any $\tau > 0$,

$$\sup_{t \le \tau} |(\bar{Q}^n(t), I^n(t)) - (q(t), i(t))| \to 0, \text{ almost surely,}$$

where $\{(q(t), i(t)), t \ge 0\}$ is defined as follows.

(i) For $t \le 0$, $(q(t), i(t)) = (H(t), 0)$.

(ii) When $t \in [0, T]$, $\{(q(t), \mu i(t)), t \in [0, T]\}$ is the unique solution to a Skorokhod problem associated with $\{H(t) - \mu t, t \in [0, T]\}$. More precisely, $\{(q(t), \mu i(t)), t \in [0, T]\}$ satisfies, for $t \in [0, T]$,

$$q(t) = H(t) - \mu t + \mu i(t),$$

and $i(0) = 0$, $i(t)$ is nondecreasing, and it increases only when $q(t) = 0$. Equivalently, we have

$$i(t) = \mu^{-1} \sup_{0 \le s \le t} \max\{-H(s) + \mu s, 0\}, \; q(t) = H(t) - \mu t + \mu i(t), \quad t \in [0, T].$$

(iii) For $t \ge T$,

$$(q(t), i(t)) = \begin{cases} (q(T) - \mu(t - T), i(T)), & t \in [T, T + q(T)/\mu], \\ (0, i(T) + (t - T - q(T)/\mu)), & t > T + q(T)/\mu. \end{cases}$$

Remark 3.3.1. From Proposition 3.3.1, in the fluid limit, when $t < 0$, the server hasn't started to work, so the idle time process has value 0 and the queue length is increasing according to the CDF $H(\cdot)$. When $0 \le t \le T$, the fluid limit is a fluid $G_t/GI/1$ queue with initial value $H(0)$, time-varying cumulative arrivals $H(t) - H(0)$, and service rate μ. When $t > T$, there are no more external arrivals. The queue length is now decreasing linearly with rate μ till it becomes zero at time $T + q(T)/\mu$ and stays zero from then on. Further, we note that $q(t)$ and $\mu i(t)$ represent the Skorokhod map for RCLL function $H(t) - \mu t$.

3.4 Fluid Control Problem (FCP)

We construct the FCP that is the deterministic counterpart of the original stochastic control problem. Parallel to the stochastic control problem in (3.4) and using the fluid limits derived in Section 3.3, for a given service rate μ, the FCP is to choose an *appointment profile* $A = \{A(t), t \in [0, T]\}$ that optimizes the following optimal control problem:

$$\max J(A) = rH(T) - c_w \int_0^\infty q(t)dt - c_i i(T) - c_o \frac{q(T)}{\mu}, \tag{3.10}$$

subject to

$$H(t) = \int_0^T F(t-s, s)dA(s), \quad t \in, \tag{3.11}$$

$$(q(t), i(t)) = \begin{cases} (H(t), 0), & t < 0, \\ (\Phi(H - \mu\iota)(t), \Psi(H - \mu\iota)(t)), & t \in [0, T], \\ (q(T) - \mu(t - T), i(T)), & t \in (T, T + q(T)/\mu], \\ (0, i(T) + (t - T - q(T)/\mu)), & t > T + q(T)/\mu, \end{cases} \tag{3.12}$$

$$A(0) = 0 \text{ and } A(t) \text{ is RCLL nondecreasing over } [0, T], \tag{3.13}$$

where

$$\Phi(H - \mu\iota)(t) = H(t) - \mu t + \mu i(t) \tag{3.14}$$

and

$$\Psi(H - \mu\iota)(t) = \frac{1}{\mu} \sup_{0 \le s \le t} \max\{\mu s - H(s), 0\}. \tag{3.15}$$

The constraints (3.11) and (3.12) are derived from (3.8) and Proposition 3.3.1. We note that in the constraint (3.13) the control $A(t)$ is not required to be a CDF. This relaxation combines the two control variables n and $\{a_i^n\}_{i=1}^n$ of the stochastic control problem into a single control variable $\{A(t); t \in [0, T]\}$. The terminal value $A(T)$ would represent the number of appointments, and the normalized control $A(t)/A(T)$ provides the distribution of the appointment times. Equation (3.14) and Equation (3.15) represent the Skorokhod map associated with $H(t) - \mu t$.

From Proposition 3.3.1 (ii), for $t \in [0, T]$,

$$i(t) = \frac{1}{\mu}(q(t) + \mu t - H(t)). \tag{3.16}$$

Using (3.12) and (3.16), the objective function (3.10) becomes

$$\begin{aligned} J(A) &= rH(T) - c_w \int_0^T q(t)dt - c_w \int_T^\infty q(t)dt - \frac{c_i}{\mu}(q(T) + \mu T - H(T)) - c_o \frac{q(T)}{\mu} \\ &= \left(r + \frac{c_i}{\mu}\right) H(T) - c_w \int_0^T q(t)dt - \frac{c_w}{2\mu} q(T)^2 - \frac{c_i + c_o}{\mu} q(T) - c_i T. \end{aligned}$$

Consequently, the FCP is equivalent to select a function $A = \{A(t), t \in [0, T]\}$ to maximize

$$\tilde{J}(A) = \left(r + \frac{c_i}{\mu}\right) H(T) - c_w \int_0^T q(t)dt - \frac{c_i + c_o}{\mu} q(T) - \frac{c_w}{2\mu} q(T)^2 \tag{3.17}$$

subject to

$$H(t) = \int_0^T F(t - s, s)dA(s), \quad t \in, \tag{3.18}$$

$$q(t) = \Phi(H - \mu\iota)(t), \quad t \in [0, T], \tag{3.19}$$

$$A(0) = 0 \text{ and } A(t) \text{ is RCLL nondecreasing over } [0, T]. \tag{3.20}$$

3.4.1 Special case: No unpunctuality

We consider the special case when all patients arrive at their scheduled appointment times with zero unpunctuality. Here $H(t) = A(t)$ for all $t \in$ and the FCP is simplified to choose a RCLL nondecreasing function $H = \{H(t), t \in [0, T]\}$ satisfying $H(0) = 0$ to maximize

$$J(H) = rH(T) - c_w \int_0^\infty q(t)dt - c_i i(T) - c_o \frac{q(T)}{\mu},$$

subject to the constraint (3.19). Using the properties of the Skorokhod map, this control problem can be solved explicitly (Honnappa, 2015; Armony and Honnappa, 2019).

Proposition 3.4.1. The FCP with no unpunctuality is equivalent to the variational problem that selects a RCLL nondecreasing function $H = \{H(t); t \leq T\}$ satisfying $H(0) = 0$ and $H(t) \geq \mu t$ for

$t \in [0, T]$ to maximize

$$\check{J}(H) = (r - c_o/\mu + c_w T)H(T) - \frac{c_w}{2\mu}H(T)^2 - c_w \int_0^T H(t)dt, \tag{3.21}$$

and it admits the following optimal solution: For $t \in [0, T]$,

$$H^*(t) = \begin{cases} \mu t, & \text{if } r \leq c_o/\mu, \\ \mu t + \frac{\mu(r - c_o/\mu)}{c_w} 1_{\{t=T\}}, & \text{if } r > c_o/\mu. \end{cases} \tag{3.22}$$

The corresponding optimal state process and the FCP optimal value $J^* = \max_H J(H)$ are given as follows:

(i) When $r \leq c_o/\mu$, $q^*(t) = 0$ for all $t \geq 0$, and $J^* = r\mu T$.

(ii) When $r > c_o/\mu$,

$$q^*(t) = \begin{cases} 0, & t \in [0, T) \\ \frac{\mu(r - c_o/\mu)}{c_W} - \mu(t - T), & t \in [T, T + \frac{(r - c_o/\mu)}{c_w}], \\ 0, & t > T + \frac{(r - c_O/\mu)}{c_W}, \end{cases}$$

and $J^* = r\mu T - \frac{\mu[(r - c_0/\mu)^+]^2}{2c_W}$.

Remark 3.4.1. $H^\star(t)$ follows the intuition that, over $[0, T)$, if the rate of patient arrivals matches the rate of patient departures, the system is in perfect equilibrium, accruing no wait costs nor idle costs over this time horizon. The only time a wait cost and overtime cost would occur is if the reward is sufficiently high relative to the overtime and wait cost. As the wait cost of the system grows quadratically in $q(T)$, a finite amount of overbooking is guaranteed for finite r.

3.4.2 Special case: uniform unpunctuality

Recall that if $r \leq c_o/\mu$, then $H^*(t) = \mu t$ is an optimal solution of the FCP without considering unpunctuality. One possibility for finding the optimal control A analytically is to have the control satisfy $\int_0^T F(t - s)dA(s) = \mu t$. The following corollary is a direct result of Proposition 3.4.1.

Corollary 1. *Suppose $r \leq c_o/\mu$ and there exists RCLL A such that $\int_0^T F(t-s)dA(s) = \mu t$ for $t \in [0,T]$. Then A is an optimal control to the FCP.*

We consider the situation where the unpunctuality is uniformly distributed over the interval $[-a,b]$ for each patient, where $a,b>0$, independent of their arrival time. Our goal is to construct an RCLL nondecreasing function A such that $H^*(t) = \int_0^T F(t-s)dA(s)$, where $F(t) = (t+a)/(b+a),\ t \in [-a,b]$, is the CDF of the uniform unpunctuality time. This will be optimal since $H^\star(t) = \mu t$ will accrue a cost of 0.

Motivated from the numerical experiments, we consider piecewise constant A similar to the CDF of a discrete random variables. Note that if A puts some positive mass m at a time point s, in the convolution

$$F(t-s)dA(s) = \begin{cases} 0, & \text{if } t-s < -a, \\ 1, & \text{if } t-s > b, \\ \frac{(t-s+a)m}{b+a}, & \text{if } -a \leq t-s \leq b, \end{cases} \tag{3.23}$$

which says the influence of the mass m at time s is spread uniformly over the interval $[s-a, s+b]$. Following this observation, we consider an A function that jumps at points $a, 2a+b, 3a+2b, \ldots, Na+(N-1)b$, where $N = \max\{n \geq 1 : T-[na+(n-1)b] < a+b\}$, and the mass at each point is set to be $\mu(a+b)$. Then when $T = N(a+b)$, one can check that $H^*(t) = \int_0^T F(t-s)dA(s) = \mu t,\ \ 0 \leq t \leq T$.

For general $F(t)$, the FCP is difficult to solve to optimality with the exception of a few special cases of unpunctuality (zero unpunctuality, or certain types of uniform distributions). However, it can be solved numerically to arbitrary accuracy, if we discretize the time. We give the details in the next section.

3.5 Quadratic Programming Formulation

Suppose we discretize $[0,T]$ into K segments, with end points $t_0, t_1, \cdots, t_K$, where $t_k = \frac{kT}{K}$ for $k = 0,1,\ldots,K$. We call K the resolution of the discretization. Let $p_k = A(t_{k+1}) - A(t_k)$, $k = 0,1,\cdots,K-1$, and define

$$\mathbf{p} = [p_0, p_1, \cdots, p_{K-1}].$$

We replace the continuous convolution $H(t)$ by its the discrete convolution

$$\hat{H}(t) = \sum_{k=1}^{K} p_k F(t - t_{k-1}, t_{k-1}) = \boldsymbol{F}(t)^T \boldsymbol{p}.$$

We also discretize the Skorokhod map on $[0, T]$ as follows:

$$\hat{i}(k) = \frac{1}{\mu} \max_{i \in \{0,1,\ldots,k\}} \max\{\mu t_i - \hat{H}(t_i), 0\}$$

and

$$\hat{q}(k) = \hat{H}(t_k) - \mu t_k + \mu \hat{i}(t_k)$$

for $k = 0, 1, \ldots, K$.

Lemma 3.5.1. The discrete Lindley's recursion

$$\begin{cases} q_{k+1} = \max\left\{0, q_k + \hat{H}(t_k) - \hat{H}(t_{k-1}) - \frac{\mu T}{K}\right\} \text{ for } k = 0, \ldots, K-1 \\ I_k = \frac{1}{\mu} \max\left\{0, \frac{\mu T}{K} - (\hat{H}(t_k) - \hat{H}(t_{k-1})) - q_k\right\} \text{ for } k = 0, \ldots, K \end{cases} \tag{3.24}$$

with $q_0 = \hat{H}(0)$ satisfies the discretized Skorokhod map on $[0, T]$; i.e., $q_k = \hat{q}(k)$ and $\sum_{j=0}^{k} I_j = \hat{i}(k)$ for $k = 0, 1, \ldots, K$.

Note that q_k corresponds to the fluid queue length at time t_k and I_k corresponds to the idle time over the period $(t_{k-1}, t_k]$.

Using the discretized convolution and Lindley's recursion for the Skorokhod map, we are able to discretize the FCP as a quadratic program:

$$\begin{aligned}
\max_{\mathbf{p}} \hat{J}(\mathbf{p}) \quad & r\sum_{i=0}^{K-1} F(T-t_i,t_i)p_i - \frac{c_w}{2\mu}q_K^2 - \frac{c_o}{\mu}q_K - \sum_{i=0}^{K-1}\left(\frac{c_w T}{K}q_i + c_I I_i\right) \\
\text{subject to} \quad & I_0 = 0, \quad q_0 = \sum_{i=0}^{K-1} F(-t_k,t_k)p_k \\
& q_k = q_{k-1} - \tfrac{\mu T}{K} + \sum_{i=0}^{K-1} p_i\left(F(t_k-t_i,t_i) - F(t_{k-1}-t_i,t_i)\right),\ k=1,\dots,K \qquad (3.25)\\
& I_k = \tfrac{T}{K} - \tfrac{1}{\mu}q_{k-1} - \tfrac{1}{\mu}\sum_{i=0}^{K-1} p_i\left(F(t_k-t_i,t_i) - F(t_{k-1}-t_i,t_i)\right),\ k=1,\dots,K \\
& I_k \geq 0, q_k \geq,\ k=0,\dots,K.
\end{aligned}$$

This quadratic program has a linear number of variables and constraints in the resolution K. The quadratic component comes from the wait cost of q_K patients remaining in queue at time T as the queue length decreases linearly to 0 at rate μ. It is important to note that the quadratic program's optimal objective converges to the FCP's optimal objective as $K \to \infty$ which can be observed by examining the error bound between $J(A)$ and $\hat{J}(\mathbf{p})$ in Proposition 3.5.1.

Proposition 3.5.1. Let $J(A^\star)$ be the value of Equation (3.10) at optimality and let $\hat{J}(\mathbf{p}^\star)$ be the objective value of Problem (3.25) at optimality. Let $\bar{f}$ be the maximum value of the derivative of F on $[0,T]$, let $\bar{h}$ be the maximum derivative of $H^\star(t)$ over $[0,T]$, and let $\bar{A}(T) < \infty$ be the optimal control for the fluid control problem at time T. Then,

$$\begin{aligned}
|J(A^\star) - \hat{J}(\mathbf{p}^\star)| \quad \leq \quad & \tfrac{T}{K}\left(r\bar{f}\bar{A}(T) + c_w(\bar{h}T + \bar{f}T\bar{A}(T) + (\mu+1)T)\right) \\
& + \tfrac{T}{K}\left(\tfrac{c_w}{2\mu}\left(2\bar{A}(T) + 1 + \tfrac{(\mu+1)\bar{f}\bar{A}(T)}{\mu}\right)\left(1 + \tfrac{(2\mu+1)\bar{f}\bar{A}(T)}{\mu}\right)\right) \qquad (3.26)\\
& + \tfrac{T}{K}\left(c_I\left(1 + \tfrac{\bar{f}\bar{A}(T)}{\mu}\right) + \tfrac{c_o}{\mu}\left(1 + \tfrac{(\mu+1)\bar{f}\bar{A}(T)}{\mu}\right)\right).
\end{aligned}$$

Thus, $|J(A^\star) - \hat{J}(\mathbf{p}^\star)| \to 0$ as $K \to \infty$.

Remark 3.5.1. Proposition 3.5.1 assumes Lipschitz continuity on the unpunctuality CDF $F(t, a)$ and optimal arrival profile $H^\star(t)$. The left-Riemann sum approximation error of a Lipschitz continuous function $F(t)$ over $[0, T]$ with resolution K has error bound

$$\left| \sum_{i=0}^{K-1} t_i F(t_i) - \int_0^T F(t)dt \right| \leq M_1 \frac{T^2}{K}$$

where $|F'(t)| \leq M_1$ at all points F is continuous. However, for RCLL functions with a countable set of jump discontinuities, such as CDFs for discrete random unpunctuality, the error bound changes as follows:

$$\left| \sum_{i=0}^{K-1} t_i F(t_i) - \int_0^T F(t)dt \right| \leq M_1 \frac{T^2}{K} + \frac{dT}{K}$$

where $\mathcal{S} = \left\{ t \in \mathbb{R} \ : \ \lim_{x \uparrow t} F(x) \neq F(t) \right\}$ is the countable set of jump discontinuities point on the domain of F and $d = \sum_{t \in \mathcal{S}} \left(F(t) - \lim_{x \uparrow t} F(x) \right)$ is the total height across all jump discontinuities. If F is a CDF, we know that $d \leq 1$. We redefine the error bound as

$$\left| \sum_{i=0}^{K-1} t_i F(t_i) - \int_0^T F(t)dt \right| \leq M_2 \frac{T^2}{K}$$

where $M_2 = M_1 + \frac{1}{T}$ if F is a CDF and $M_2 = M_1 + \frac{d}{T}$ if $F(t) = H^\star(t)$ with $d \leq H^\star(T) < \infty$ by Equation (3.10). Such a substitution can be made in the error-bound formula to account for jump-discontinuities induced by discrete random variables.

3.6 Numerical Results

We present our numerical results in two sections. The first section will focus on initial insights with respect to two questions:

1. How do appointment profiles change relative to different types of unpunctuality distributions.
2. How do these appointment schedules change as we incorporate time-heterogeneous unpunctuality?

The second section of the numerical results will look at a real dataset from several clinics. In particular, we will look at a single doctor who sees a large number of patients in a single day (at least 60 patients), as we wish to motivate the asymptotically optimal schedule as high performing for systems in which a large number of patients need to be seen. The dataset provides us with actual appointment times throughout a day and each is associated with an actual unpunctuality for the patient provided via the check-in time of the patient and its difference from the scheduled appointment time. Reliable service times are not provided, so we consider 3 parametric cases for service time distribution: deterministic, exponential, and log-Normal distributions.

3.6.1 Preliminary Insights

We examine the optimal appointment schedules derived from various parametric unpunctuality distributions. We consider 3 classes of parametric distributions in this section: Uniform on (a, b), denoted $U(a, b)$, Normal with mean μ and standard deviation σ, denoted $N(\mu, \sigma)$, and Generalized Laplace, denoted $\mathbb{L}(\mu, \pi, \lambda_l, \lambda_r)$. The first two have well-known density curves. The generalized Laplace distribution has the following mixture density curve:

$$f(x; \mu, \pi, \lambda_l, \lambda_r) = \begin{cases} \pi\lambda_l e^{\lambda_l(x-\mu)}, & \text{if } x \leq \mu \\ (1-\pi)\lambda_r e^{-\lambda_r(x-\mu)}, & \text{if } x > \mu. \end{cases}$$

For all appointment schedules in this subsection, we use $c_w = 1$, $c_I = 5$, $c_o = 7.5$, $r = 0$, $\mu = 100$, and $T = 1$. For the numerical discretization, we use $K = 1000$.

We will use the following sets of mean-variance pairs across the three distributions: (μ, σ^2) of $(-0.1, 0.0025), (-0.05, 0.01), (0, 0.04)$. From left to right, the mean unpunctuality is increasing along with the variance. This means, for Normally distributed unpunctuality, we have the following distributions: $N(-0.1, 0.0025)$, $N(-0.05, 0.01)$, and $N(0, 0.04)$. Figure 3.1's left plot shows the appointment profile $A(t)$ over $[0, T]$ outputted by our quadratic program with the Normal unpunctuality distributions. It is interesting to note that Normal unpunctuality produces block-schedules. Upon close examination, we saw that $\sum_{i=0}^{K-1} p_i f(x - t_i; \mu, \sigma)$, where $f(x; \mu, \sigma)$ is the density curve for a $N(\mu, \sigma)$ distribution, would approximate a uniform-type density better and better for smaller

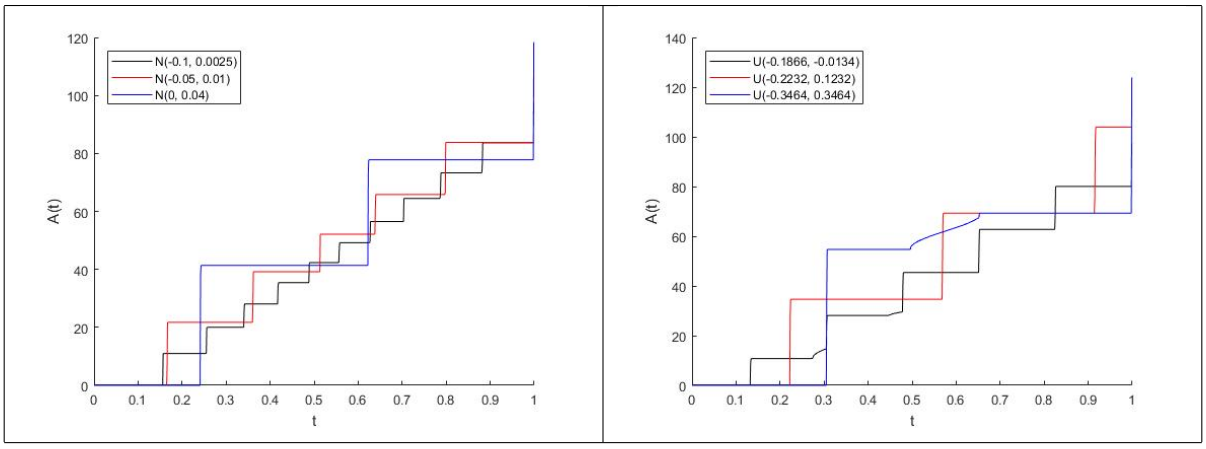

Figure 3.1: Left: appointment profiles derived for 3 different Normal unpunctuality distributions. Right: appointment profiles derived for 3 different Uniform unpunctuality distributions.

σ values. Thus, the solution to our quadratic program acted as a series of mixture densities for a Normal series approximation to a uniform distribution.

For the uniform unpunctuality case, we consider 3 distributions: $U(-0.1866, -0.0134)$, $U(-0.2232, 0.1232)$, and $U(-.3464, 0.3464)$. These parameters ensure that our mean and variances match the Normal cases. Figure 3.1's right plot shows the appointment profile $A(t)$ over $[0, T]$ outputted by our quadratic program with the uniform unpunctuality distributions.

For the generalized Laplace distribution, we consider three distributions. $\mathbb{L}(-0.1211, 0.35, 45, 22.5)$,
$\mathbb{L}(-0.05, 0.5, 14.15, 14.15)$, and $\mathbb{L}(0.085, 0.65, 5.59, 11.18)$. These parameters are once again chosen to match the same mean and variance as in the uniform and Normal distribution scenarios; however, the generalized Laplace distribution allows for skewness to be incorporated. Figure 3.2's left plot shows the PDFs associated with each of the parameter cases, whereas the right plot shows the associated appointment profiles $A(t)$ derived from our quadratic program.

It is interesting to note that the generalized Laplace distribution leads to appointment schedules that appear closer to uniform in distribution plus a shift relative to the expected unpunctuality. For example, the $\mathbb{L}(-0.1211, 0.35, 45, 22.5)$ distribution has a high probability of arriving early and leads

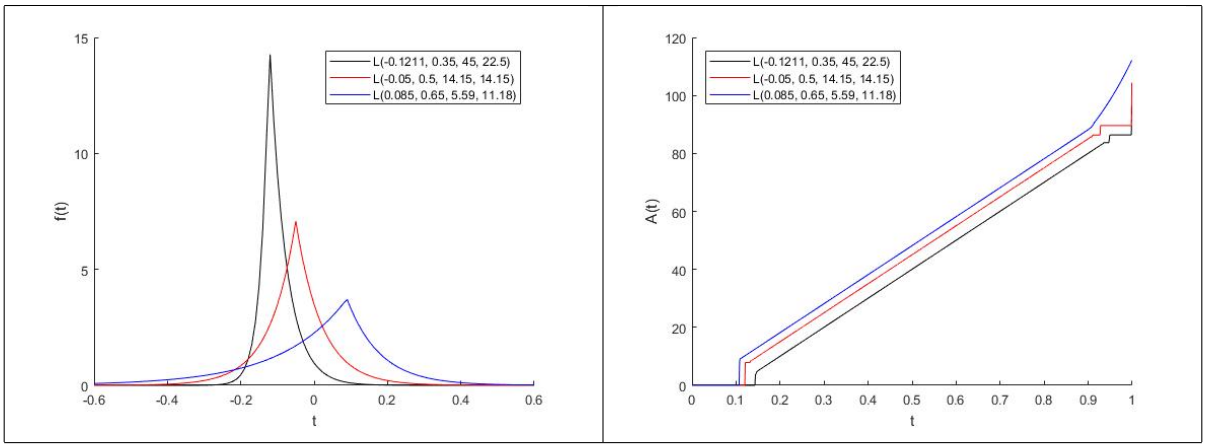

Figure 3.2: Left: PDFs for 3 different generalized Laplace unpunctuality distributions. Right: appointment profiles derived for the same 3 generalized Laplace unpunctuality distributions.

to appointments not being scheduled until around time 0.15 with a small block of patients to start; the schedule increases relatively uniformly until a block schedule at the end of the appointment profile.

We also considered the impact of time-heterogeneity when scheduling and how it impacted the appointment profiles. We consider two types of time-heterogeneity: a midday split scenario, which has a single distribution for the first half of the scheduling horizon and a different distribution for the second half of the horizon, and a parametric drift scenario where a single parametric family is picked for the distribution of unpunctuality, but its parameters change continuously over the time horizon $[0, 1]$.

For the midday split, we use the following unpunctuality distribution:

$$U_i|a_i = \begin{cases} U_{i,e} \sim N(0, 0.04), & \text{if } a_i \leq \frac{1}{2} \\ U_{i,l} \sim N(-0.1, 0.0025), & \text{otherwise.} \end{cases}$$

This causes patients scheduled in the earlier half of the horizon to arrive on time, on average, but with greater variance. If they are scheduled later in the day, they arrive earlier on average with less variance. These parameter choices match two of the earlier scenarios.

For the parametric drift, we have

$$U_i|a_i \sim N(\mu(a_i), \sigma^2(a_i)),$$

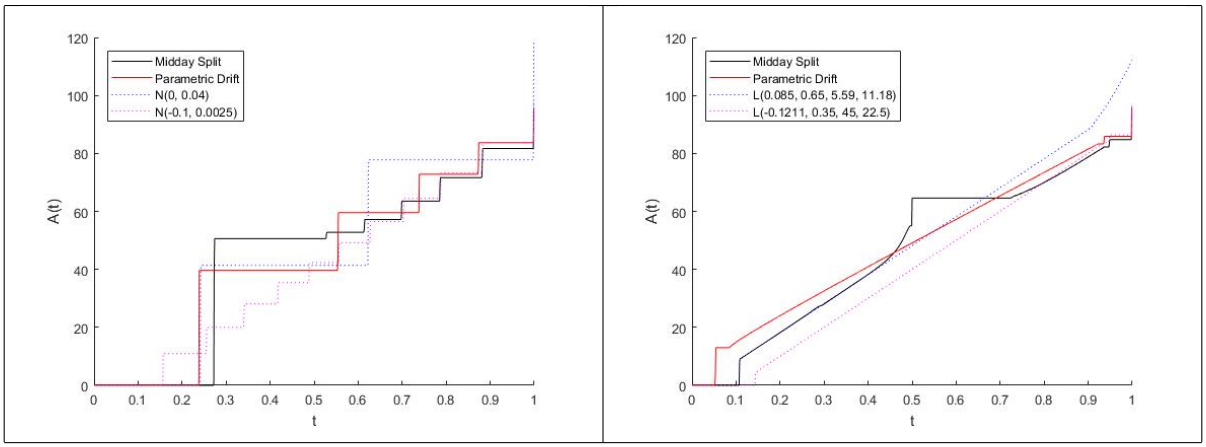

Figure 3.3: Left: appointment profiles derived with midday split and parametric drift time-heterogeneous unpunctuality based on Normal distributions with homogeneous-derived schedules for comparison (dotted lines). Right: appointment profiles derived with midday split and parametric drift time-heterogeneous unpunctuality based on generalized Laplace distributions with homogeneous-derived schedules for comparison (dotted lines).

where $\mu(t) = -0.1t$ and $\sigma(t) = 0.2 - 0.15t$ reflecting a constant linear drift in mean and standard deviation over the scheduling horizon.

The left plot of Figure 3.3 shows the schedules resulting from the midday split and parametric drift Normal distribution scenarios. To compare, the schedules derived from homogeneous $N(0, 0.04)$ and $N(-0.1, 0.0025)$ are presented as dotted lines for comparison.

We complete a similar analysis for the generalized Laplace distribution, with both a midday split and parametric drift scenario. The midday split is as follows:

$$U_i|a_i = \begin{cases} U_{i,e} \sim L(0.085, 0.65, 5.59, 11.18), & \text{if } a_i \leq \frac{1}{2} \\ U_{i,l} \sim L(-0.1211, 0.35, 45, 22.5), & \text{otherwise.} \end{cases}$$

The parametric drift case again is a continuously evolving distribution setup as follows:

$$U_i|a_i \sim L(\mu(a_i), \pi(a_i), \lambda_1(a_i), \lambda_2(a_i)),$$

with $\mu(t) = 0.085 - 0.2061t$, $\pi(t) = 0.65 - 0.3t$, $\lambda_1(t) = 5.59 + 39.41t$, and $\lambda_2(t) = 11.18 + 11.32t$.

The left plot of Figure 3.3 shows the schedules resulting from these two time-heterogeneous generalized Laplace scenarios with homogeneous-derived schedules for comparison.

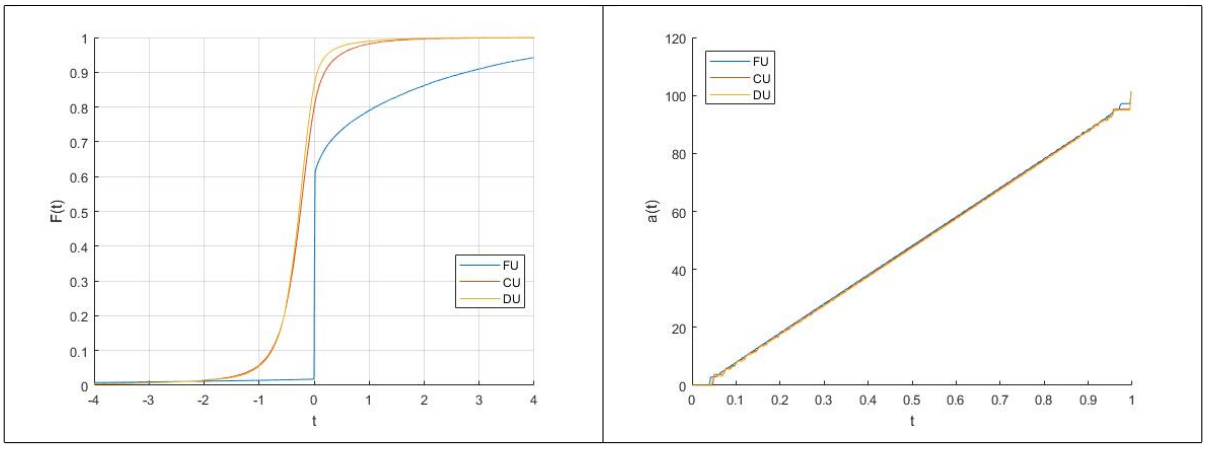

Figure 3.4: Left: Empirical CDFs for full data (FU), largest-booking clinic data (CU), and largest-booking doctor data (DU) Right: appointment profiles derived for the same 3 unpunctuality distributions. Unpunctuality is rescaled such that $T = 1$.

It can be seen that, on either side of a midday split, the schedules reflect the respective appointment schedules from the reference distributions with the dotted lines. In the parametric drift scenarios, the schedules vary their behavior throughout the day: in the Normal case, the blocks grow closer and closer together as the variance decreases in time; in the generalized Laplace case, the slope of the schedule appears to be shallower than the reference schedules; however, the schedule starts with a sizeable block near the beginning to make up for it.

3.6.2 Data-driven Experiments

We examined a new dataset that contains patient unpunctuality and appointment schedules. The data is across several clinics and each data point corresponds to a single appointment. An appointment includes anonymized clinic ID, anonymized doctor ID, in addition to scheduled appointment times and actual arrival times, from which we can derive the unpunctuality associated with the appointment. We examined the full data in addition to data for the single clinic and single doctor with the highest booking frequency. Figure 3.4 presents the empirical CDFs for unpunctuality across the 3 different subsets of the data.

We now examine how a schedule derived from our quadratic program performs relative to a baseline in a discrete event simulation setting. As we have access to the scheduled appointment

times in our data, we can set our baseline to be the actual schedules used by these clinics in the past. Each system will be a single-server queue following a FIFO service discipline. Due to this system setup, given an appointment schedule and a sample path of unpunctuality and service time values, we can easily calculate the components of the objective function using Lindley's equations. Sample paths for unpunctuality can be drawn from the data and service times will be generated as either deterministic, exponential, or log-Normal random variables. Looking at the largest-booking doctor, we can examine a particular day. However, the scheduling window of a day is not consistent across several days. To simplify things, we normalize a day's appointment schedule so it falls within the range $[0, 1]$–this is based off the desire to set $T = 1$. Suppose there are P patients scheduled for a single day and a_i is the time of the appointment for the ith patient for $i = 1, \ldots, P$. We normalize a_i as follows:

$$\tilde{a}_i = \frac{a_i - \min\limits_{i=1,\ldots,P}\{a_i\}}{\max\limits_{i=1,\ldots,P}\{a_i\} - \min\limits_{i=1,\ldots,P}\{a_i\}}$$

As appointments are being rescaled, we must similarly rescale unpunctuality. If u_i is the unpunctuality of the ith patients for $i = 1, \ldots, P$, we rescale u_i as follows:

$$\tilde{u}_i = \frac{u_i}{\max\limits_{i=1,\ldots,P}\{a_i\} - \min\limits_{i=1,\ldots,P}\{a_i\}}.$$

We have several days of data for the highest-booking doctor. However, not all days have the same number of patients, with some days having very few bookings. We wish to compare how our QP-based appointment scheduling compares to the existing schedules in the data. As our model is based off an asymptotically optimal formulation, we will consider only days where the number of bookings is on the large-side: the day must have at least 60 patients. This leaves us with 492 days to compare schedules to. Another issue to note is that, in the data, the number of patients seen on a particular day varies greatly; however, we have no indication of the cost structure used to derive these schedules. To address this, we keep $c_I = 10$ and $c_o = 15$ fixed, but vary c_w within the set $\{1, 2, 5, 10\}$. Also, we only need to solve a single quadratic program to produce an appointment schedule as we employ scalar multiplication of $\mathbf{p}$ in order to match the correct number of patients to be scheduled with data on a particular day.

As correct values for service times are not provided by the data, we instead simulate service times. Let P be the number of patients scheduled on a particular day. Service times are drawn from 3 different distributions:

1. $Det\left(\frac{1}{P}\right)$: service times are deterministic with fixed value $\frac{1}{P}$.
2. $Exp(P)$: service times are exponential with rate P.
3. $\log -N(\mu, 2)$: service times are log-Normally distributed with logarithmic standard deviation $\sigma = 2$ and logarithmic mean $\mu = -\log(P) - \frac{\sigma^2}{2}$.

The parameters of these distributions are chosen such that the expected value of the sum of the service times is equal to $T = 1$, meaning the clinic is neither overbooked nor underbooked relative to the service time durations.

Table 3.1 provides 95% bootstrap confidence intervals for the mean value of the objective function for each schedule across the 492 days with deterministic service times. We compare several schedules:

- Actual refers to using the actual schedule recorded in the dataset.
- ZU refers to an optimal schedule under zero unpunctuality according to Proposition 3.4.1.
- QP-HMG refers to a data-driven schedule derived from using the quadratic program with time-homogeneous unpunctuality.
- QP-Hour refers to a data-driven schedule derived from using the quadratic program with time-heterogeneous unpunctuality according to an hourly partition.

c_w	Actual	ZU	QP-HMG	QP-Hour
1	(34.18, 35.71)	(31.92, 33.39)	(29.28, 30.67)	(28.53, 29.89)
2	(67.88, 71.05)	(63.45, 66.35)	(57.99, 60.86)	(55.20, 57.85)
5	(169.19, 177.05)	(157.98, 165.37)	(144.33, 151.58)	(138.11, 145.00)
10	(338.29, 353.87)	(315.69, 330.38)	(288.13, 302.56)	(272.18, 285.77)

Table 3.1: Bootstrap mean confidence intervals of objective value and relative improvement across different cost structures with deterministic service times.

Table 3.2 provides the same results for the case that service times are exponentially distributed. We notice that switching from deterministic to exponential does not seem to have a strong impact on the mean values of the objective function for the different cost structures.

c_w	Actual	ZU	QP-HMG	QP-Hour
1	(34.93, 36.82)	(32.73, 34.50)	(30.05, 31.79)	(28.31, 30.02)
2	(68.83, 72.60)	(64.39, 67.96)	(58.99, 62.48)	(55.52, 59.13)
5	(170.53, 179.96)	(159.41, 168.33)	(145.93, 154.65)	(134.99, 143.61)
10	(340.03, 358.89)	(317.76, 335.55)	(290.70, 308.13)	(269.83, 288.11)

Table 3.2: Bootstrap mean confidence intervals of objective value and relative improvement across different cost structures with exponential service times.

Table 3.3 similarly provides the same simulation results, but with log-Normal simulated service times. In this case, we see that the log-Normal distribution leads to a more noticeable increase in objective function as opposed to switching from deterministic to exponential.

c_w	Actual	ZU	QP-HMG	QP-Hour
1	(37.33, 50.32)	(35.67, 48.59)	(33.54, 46.43)	(32.57, 45.22)
2	(70.38, 94.45)	(67.05, 90.85)	(62.66, 86.35)	(60.15, 83.17)
5	(169.52, 226.52)	(161.16, 217.39)	(150.08, 206.11)	(147.24, 198.29)
10	(334.95, 447.32)	(317.94, 428.42)	(295.73, 405.80)	(286.36, 390.93)

Table 3.3: Bootstrap mean confidence intervals of objective value and relative improvement across different cost structures with log-Normal service times.

We can see that, in general, the incorporation of unpunctuality will produce the best results, with the factoring-in of time-heterogeneous unpunctuality producing the best overall schedules.

3.7 Conclusion

In this paper, we developed a framework for computing asymptotically optimal appointment schedules with generalized patient unpunctuality that also considers the under/overbooking of patients in the form of a fluid control problem. As the general problem is not analytically solvable, we provided a time-discretized quadratic programming approximation that converges to the fluid control problem as the time-discretization grows finer and finer. We then finished with a numerical study that first examined the appointment profiles associated with several different types of unpunctuality. It was interesting to note that even in the Normal case, multi-block-type schedules were optimal under our model. However, other types of unpunctuality, such as generalized Laplace,

may lead to more fluid-arrival-type appointment profiles that appear shifted relative to expected unpunctuality. Further, we discovered that time-heterogeneous unpunctuality can lead to significant differences in the character of appointment schedules from the homogeneous case. In our final section of the numerical studies, we compared schedules derived from the FCP or QP against a real dataset with real schedules and unpunctuality values across several hundred days. We saw that the time-heterogeneous assumption performed the best on the real data and that, in general, it is better to use some form of unpunctuality, whether it be time-homogeneous or time-heterogenous, rather than assume zero unpunctuality.

CHAPTER 4

Optimal Asymptomatic Disease Testing Under Limited Test Supply

4.1 Motivation

In this chapter, we wish to develop easy-to-understand guidelines for the average person to follow when it comes to using a finite set of disease tests in a mathematically optimal manner. We develop an analytically tractable model that allows us to develop optimal policies and heuristics for a variety of individual-focused decision scenarios. While we make the modeling choice of using partially-observable Markov decision processes, which generally require state-probability knowledge for making decisions according to a policy, we choose a class of policies that allow us to instruct a decision-maker to behave optimally without knowledge of these probabilities, given an initial condition. Next, we examine how these policies perform in a community-based simulation. As the policies are derived based on an individual's decision-making process, our wish is for it to perform well as a heuristic for a community-wide optimal policy.

4.2 The Model

Our objective is to minimize the number of days an individual stays in an infected state undetected. If the individual is asymptomatic, they will not be able to detect the disease without the help of test kits. We developed a partially-observable Markov decision process (POMDP) for determining the optimal timing for using tests with the presence of both symptomatic and asymptomatic infections. For this model, we will need an underlying Markov model with state space $\mathcal{S}$ and transition probability matrix T that models the state of the individual; however, this model will not be fully-observable to the individual. We will also need action space $\mathcal{A}$ and a cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

For modeling an individual's state, we assume a discrete-time Markov chain (DTMC) where each state corresponds to the current health status of the individual. The DTMC is a variation

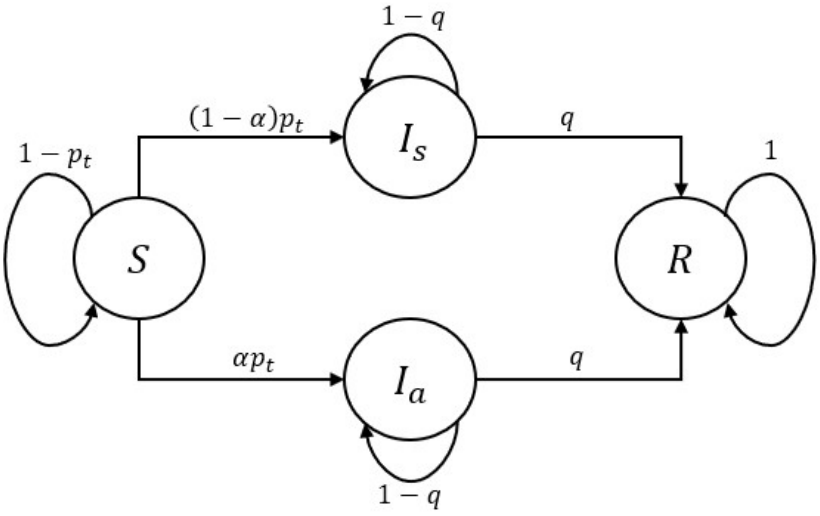

Figure 4.1: Underlying SIR model with symptomatic and asymptomatic infected states.

on a susceptible-infected-recovered (SIR) model; that is, the individual can fall into one of three categories: susceptible, infected, and recovered. However, the key difference is that due to the possibility of being asymptomatic, we split the infected state into two separate states: asymptomatic infected and symptomatic infected. Thus, our state space is defined as $\mathcal{S} = \{S, I_a, I_s, R\}$, where S stands for susceptible, I_a stands for infected-asymptomatic, I_s stands for infected-symptomatic, and R stands for recovered. Further, the DTMC is non-stationary in its infection probabilities. That is, at any time t, the probability of transitioning from S to either I_a or I_s is p_t which may vary depending on the value of t–this is designed to capture the state of the world at time t. Figure 4.1 is the transition probability diagram for the DTMC. From this, we also see the average infection time is $\frac{1}{q}$ and α is the probability of an infection being asymptomatic.

Letting $X_n \in \mathcal{S}$ represent the state the individual is in at time n, we must note that X_n cannot be directly observed by the individual as the model is not full-observable. Instead, we separately define the observation set $\mathcal{O} = \{0, 1\}$. Letting $\theta_n \in \mathcal{O}$ represent the observation the individual makes at time n, if $\theta_n = 0$ then symptoms are not observed at time n and if $\theta_n = 1$, the opposite is true. This information can be used to determine the probability of being in particular states. We define $\mathbf{b}_n = (b(S), b(I_a), b(I_s), b(R)) \in [0, 1]^4$ as the belief state at any time n. If we know $\mathbf{b}$ on day n and observe θ on the following day, we can determine $\mathbf{b}'$, the update belief on the following day,

as follows:

$$b_\prime(j|\theta, \mathbf{b}) = \frac{\Theta(\theta|j)}{P(\theta|\mathbf{b})} \sum_{i \in \mathcal{S}} T_{i,j} b(i) \tag{4.1}$$

where

$$\Theta(\theta|j) = \begin{cases} 1, & \text{if } \theta = 1, j = I_s \\ 1, & \text{if } \theta = 0, j \in \{S, I_a, R\} \\ 0, & \text{otherwise.} \end{cases}$$

is the probability of observation θ upon arriving in state j and

$$P(\theta|\mathbf{b}) = \sum_{j \in \mathcal{S}} \Theta(\theta|j) \sum_{k \in \mathcal{S}} T_{k,j} b(k).$$

Note that $\{\mathbf{b}_n : n \geq 0\}$ is a Markov process as $\mathbf{b}_n$ depends only on $\mathbf{b}_{n-1}$ and θ_n.

Our above description describes the evolution of the individual's knowledge in the absence of tests. Now we will consider the effects of having access to K tests. This expands the information state space of the POMDP to $\{(\mathbf{b}, k) : \mathbf{b} \in [0,1]^4, k \in \{0, 1, \ldots, K\}$ where $\mathbf{b}$ is the current belief state induced by θ and k is the number of remaining tests. Tests allow us to make a binary decision each day: to test or not to test. We represent the action space as $\mathcal{A} = \{u, c\}$ where u represents the action of *using* a test and c represents the action of not using a test and *continuing* to the next day. For a POMDP model, we also need costs associated with each state. If we choose to continue on a day when we are infected-asymptomatic, then we accrue a unit cost for the day. Note that we do not accrue a cost if infected-symptomatic as we automatically detect the disease and the problem is effectively over. This means that, if we use up all our tests too early, we will accrue the full cost associated with being infected-asymptomatic. In expectation this is simply $\frac{\alpha}{q}$.

We will define $V_\pi(\mathbf{b}, k)$ as the expected cost of following policy π starting from state $(\mathbf{b}, k)$ with $V_\pi(\mathbf{e}_S, 0) = \frac{\alpha}{q}$. We can express the state-action-dependent cost function $c : \mathcal{S} \times \mathbb{Z}_{\geq 0} \times \mathcal{A} \to [0, \infty)$ as follows:

$$c(i, k, a) = \begin{cases} 1, & \text{if } i = I_a, a = c \\ 0, & \text{otherwise,} \end{cases} \tag{4.2}$$

for $i \in \mathcal{S}, k \in \mathbb{Z}_{\geq 0}, a \in \mathcal{A}$.

Using a test is not guaranteed to produce a perfect result as they can produce both false negatives and false positives. We will define the probability of false negatives and false positives as follows:

$$P(-|X_n = I_a, A_n = u) = \zeta, \quad P(+|X_n \in \{S, R\}, A_n = u) = \eta \quad .$$

As our primary goal is the detection of a disease, the result of a false positive proves to be less of a risk than a false negative. We will make the following assumption about how the decision-maker handles a positive results:

1. A positive test results at time n causes the decision maker to immediately seek a professional healthcare provider that reveals whether the patient is infected-asymptomatic or not at time n.

2. If $X_n \in \{I_a, I_s\}$, the positive test was correct and the decision problem is over as the patient remains quarantined. If $X_n \in \{S, R\}$, then the positive test was incorrect and the patient restarts the problem with their beliefs updated on the conditional knowledge that they are either in state S or state R.

Last, we define $V^\star(\mathbf{b}, k)$ as the optimal value function under the optimal policy.

4.3 Analytical Results

Under various regularizations on the model, we are able to determine the optimal testing policy analytically. Since our main goal is detection, observing $\theta_t = 1$ at any time t results in the problem ending due to having detected the disease and does not require the use of tests. However, using a test when $\theta = 0$ will update our belief state based on the current belief state and the values for ζ and eta.

4.3.1 Perfect Tests

To start, let us assume that an individual's false positive and false negative rates are zero. In this case, we will consider two subcases: stationary infection probabilities and nonstationary infection probabilities.

4.3.1.1 Stationary Infection Probabilities

For this, we assume that neither false negative nor false positives may occur. Under these assumptions, if we use a test and test positive, then the problem is over as tests are perfect. If we instead test negative when in state $\mathbf{b}$, since tests are perfect, we update our belief state to $\mathbf{b}'$ as follows:

$$\mathbf{b}' = \left(\frac{b(S)}{b(S)+b(R)}, 0, 0, \frac{b(R)}{b(S)+b(R)} \right).$$

The negative test then leads us into a $k-1$ test, infinite horizon problem with starting state $\mathbf{b}'$. However, there is no cost associated with state R and R cannot return into earlier states, so we make the observation that

$$V^\star(\mathbf{b}', k-1) = b'(S)V^\star(\mathbf{e}_S, k-1) = \frac{b(S)}{b(S)+b(R)} V^\star(\mathbf{e}_S, k-1),$$

where $V^\star(\mathbf{e}_S, k)$ can be solved recursively, starting with $V^\star(\mathbf{e}_S, 0) = \frac{\alpha}{q}$.

Proposition 4.3.1. Assume an initial belief state of $\mathbf{b} \in \{\mathbf{x} : \mathbf{x} = \gamma \mathbf{e}_S + (1-\gamma)\mathbf{e}_R,\ \gamma \in [0,1]\}$ and k tests remain. If $p_t = p\ \forall t$ and $\zeta = \eta = 0$, then the optimal policy is to use a test if

$$b(S) \leq \frac{b(I_a)}{pV^\star(\mathbf{e}_S, k-1)} \tag{4.3}$$

for the first time, with $V^\star(\mathbf{e}_S, 0) = \frac{\alpha}{q}$. Moreover, if symptoms are never observed, all tests will be used in a finite amount of time.

This type of policy is what we call a one-step look-ahead (OSLA) policy. Such a policy allows us easily determine the optimal time to use tests given some initial starting start $\mathbf{b}_0$. Define τ_k as the time we would use the first of k tests under the optimal policy if we never observe symptoms and $X_0 = S$. Using the Equation System (5.24) from the proof of Proposition 4.3.1 allows us to derive a recursive method for calculating τ_k and $V^\star(\mathbf{e}_S, k)$ for $k = 0, 1, \ldots$.

Proposition 4.3.2. For $k \geq 1$,

$$\tau_k = \begin{cases} \inf\left\{ t \in \mathbb{Z}^+ : t \geq \frac{\log(\alpha - V^\star(\mathbf{e}_S, k-1)(q-p)) - \log(\alpha)}{\log(1-q) - \log(1-p)} \right\}, & \text{if } p \neq q \\ \inf\left\{ t \in \mathbb{Z}^+ : t \geq \frac{(1-p)V^\star(\mathbf{e}_S, k-1)}{\alpha} \right\}, & \text{if } p = q, \end{cases}$$

$$V^\star(\mathbf{e}_S, k) = V^\star(\mathbf{e}_S, k-1)(1-p)^{\tau_k} + \mathbb{I}_{\{\tau_k \geq 2\}} \sum_{j=1}^{\tau_k - 1} j\alpha p \left((1-p)^{\tau_k - j - 1}(1-q)^j + q(1-q)^{j-1} \sum_{i=0}^{\tau_k - j - 1} (1-p)^i \right),$$

and $V^\star(\mathbf{e}_S, 0) = \frac{\alpha}{q}$.

We also have the following structural properties of the optimal value function and the τ_k values:

Corollary 4.3.1. The optimal value functions from initial state $\mathbf{e}_S$ have the following ordering relative to k, the number of tests:

$$V^\star(\mathbf{e}_S, k) \leq V^\star(\mathbf{e}_S, k-1) \leq \cdots \leq V^\star(\mathbf{e}_S, 1) \leq \frac{\alpha}{q}.$$

It follows that

$$\tau_k \leq \tau_{k-1} \leq \cdots \leq \tau_1 < \infty.$$

The ordering of the τ_k values translates into the following insight: as we start to run low on tests, we use them more sparingly.

4.3.1.2 Nonstationary Infection Probabilities

The prior assumption that p is nonstationary in time does not capture the fluctuating nature of a pandemic at different points in time. Often, in population-level SIR models, $\frac{dI}{dt} = f(I(t), S(t))$; i.e., the infection rate is a function of the total number of infected and the remaining number of people who remain susceptible to infection. This usually corresponds to slow infection rates when the number of infected is low and high infection rates when the number of infected are high, tempered by the available pool of susceptibles. Often when a disease first surfaces, there is a chance of a pandemic which, in its early stages, will lead to both daily increase in cases and in turn a daily increase in the infection rate. Towards the end of a pandemic, we would expect the opposite to occur: the number of infections decreases day to day along with the infection rate. One of the main goals of this research is to produce easy-to-follow guidelines for the average person to follow and

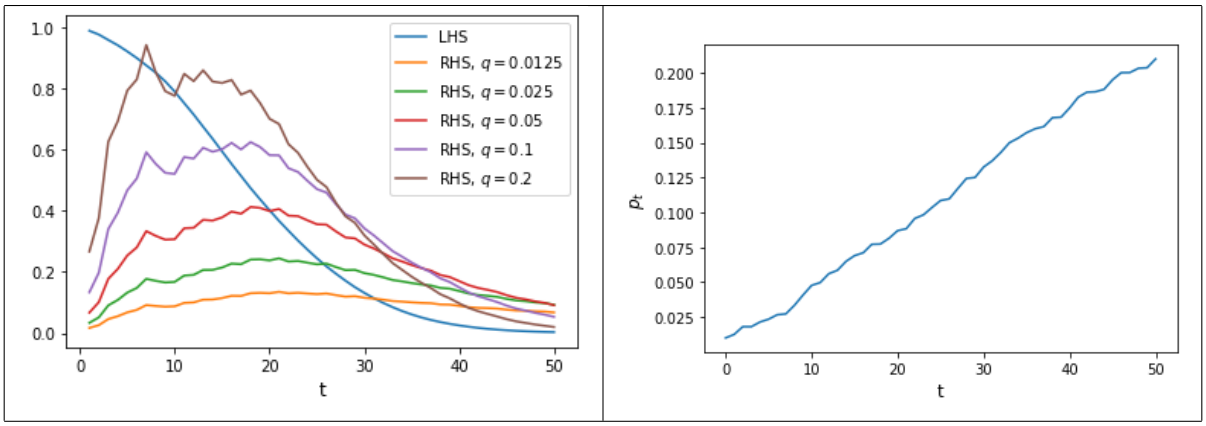

Figure 4.2: Left: plot of $LHS = \phi_s(t)$ against $RHS = \frac{q}{\alpha p_t}\phi_a(t)$ for different values of q. Right: Plot of monotone nondecreasing p_t for $t = 0, 1, \ldots, 50$.

we have leveraged optimal OSLA policies in order to easily compute what time to take the tests. However, in the nonstationary case, we make the following observation.

Remark 4.3.1. Suppose the SIR DTMC is nonstationary with infection probability sequence $\{p_t\}_{t=0}^{\infty}$. Let $\mathbf{b}_0 = \mathbf{e}_S$. The criterion for an OSLA policy being optimal does not unconditionally hold. Further, even if the sequence is monotone, the criterion for an OSLA policy being optimal does not unconditionally hold.

Let us define the following terms:

$$\phi_s(t) = \prod_{i=0}^{t-1}(1-p_t), \quad \phi_a(t) = \sum_{i=0}^{t-1} p_i(1-q)^{t-1-i}\prod_{j=0}^{i-1}(1-p_j). \tag{4.4}$$

In the nonstationary, perfect test case with $\mathbf{b}_0 = \mathbf{e}_S$, an OSLA policy is equivalent to stopping when $\phi_s(t) \leq \frac{q}{\alpha p_t} phi_a(t)$ for the first time. It can be shown via counterexample that there are cases where this may be true for some t_a, but not true for some t_b where $t_b > t_a$. This is demonstrated in Figure 4.2 by looking at the $q = 0.0125$ case which visibly reverses the inequality between $\phi_s(t)$ and $\frac{q}{\alpha p_t}\phi_a(t)$ three times instead of once. The infection probability sequence in this case is monotone nondecreasing.

However, under certain regularizing conditions on $\{p_t\}_{t=0}^{\infty}$, an OSLA policy will remain optimal. For now, we will assume perfect tests.

Proposition 4.3.3. Let $\zeta = \eta = 0$ and suppose $t = t_0$ is the first time $\phi_s(t) \leq \frac{q}{\alpha p_t}\phi_a(t)$ when starting from initial belief state $\mathbf{e}_s$. Then using the test at time $t = t_0$ is optimal if

$$\left(\frac{1-p_t}{p_t} - \frac{1-q}{p_{t+1}}\right)\phi_a(t) \leq \frac{p_t}{p_{t+1}}\phi_s(t) \tag{4.5}$$

for all $t \geq t_0$.

Note that this condition appears quite complicated; however, there is a simpler sufficient, yet tougher, condition that makes an OSLA policy optimal.

Corollary 2. *Let $\zeta = \eta = 0$ and suppose $t = t_0$ is the first time $\phi_s(t) \leq \frac{q}{\alpha p_t}\phi_a(t)$ when starting from initial belief state $\mathbf{e}_s$. Then using the test at time $t = t_0$ is optimal if*

$$\frac{p_{t+1}(1-p_t)}{p_t} \leq (1-q) \tag{4.6}$$

for all $t \geq t_0$.

Example functions that satisfy the Condition (4.6) include: the right half of a sigmoid function,

$$p_t = \frac{A}{1+e^{-t/L}}$$

where $A \leq 1-q$, $L >> 0$, and $q < 0.5$.

Note that this condition is not equivalent to concavity. A counterexample to demonstrate this is:

$$p_t = \begin{cases} 1-q^{t/L}, & \text{if } 0 \leq t \leq L \\ 1-q, & \text{if } t > L. \end{cases}$$

for $L >> 0$.

4.3.2 Imperfect Tests

We will now assume that $\zeta > 0$ and $\eta > 0$. Unlike before, we need to consider how to update the belief state as the result of a potential false negative or a false positive. In the case of a false negative when asymptomatic, we know there is a ζ probability of a negative result despite being in state I_a. Upon receiving a negative test result when in state $\mathbf{b}_n = (b_n(S), b_n(I_a), b_n(I_s) = 0, b_n(R))$

leads to the updated belief $\mathbf{b}_n^- = (b_n^-(S), b_n^-(I_a), b_n^-(I_s), b_n^-(R))$ where

$$\begin{aligned} b_n^-(S) &= \tfrac{b_n(S)}{b_n(S)+\zeta b_n(I_a)+b_n(R)}, \\ b_n^-(I_a) &= \tfrac{\zeta b_n(I_a)}{b_n(S)+\zeta b_n(I_a)+b_n(R)}, \\ b_n^-(I_s) &= 0, \\ b_n^-(R) &= \tfrac{b_n(R)}{b_n(S)+\zeta b_n(I_a)+b_n(R)}. \end{aligned}$$

For positive results, we make the following assumption: the patient, acting cautious, will consult with a medical professional upon receiving a positive test. If the result is in fact a true positive, then the patient successfully detected the disease and the problem is over. If it is a false positive, that information is revealed and the patient updates their belief state to $\mathbf{b}_n^+ = (b_n^+(S), b_n^+(I_a), b_n^+(I_s), b_n^+(R))$ where

$$\begin{aligned} b_n^+(S) &= \tfrac{b_n(S)}{b_n(S)+b_n(R)}, \\ b_n^+(I_a) &= 0, \\ b_n^+(I_s) &= 0, \\ b_n^+(R) &= \tfrac{b_n(R)}{b_n(S)+b_n(R)}. \end{aligned}$$

The long-run expected cost of using a test when in state $\mathbf{b}_n$ with $k \geq 1$ tests remaining and following policy π from then on is

$$c(\mathbf{b}_n, k, u) = b_n(S)\zeta V^\pi(b_n^+, k-1) + \left(b_n(S)(1-\zeta)b_n(I_a)\eta\right) V^\pi(b_n^-, k-1), \tag{4.7}$$

where $V^\pi(\mathbf{b}, 0) = \frac{\alpha}{q} b(S) + \frac{1}{q} b(I_a)$ regardless of policy π. Note that knowledge of the value space beyond state $\mathbf{e}_S$ is needed to calculate these costs, a major departure from the simplicity introduced by perfect tests.

It is important to note that in this case, we should consider a general starting belief state that allows $b_0(i) > 0$ for all $i \in \mathcal{S}$. In particular, we are interested in the case that $b_0(I_a) > 0$. For a single test, we have a more generalized result of the OSLA policy.

Proposition 4.3.4. Assume an initial belief state of $\mathbf{b}_0$ such that $b_0(S) > 0, b_0(I_a) \geq 0, b_0(R) \geq 0$, and $b_0(S) + b_0(I_a) + b_0(R) = 1$. Suppose we have a single test, $p_t = p \ \forall t$, $\zeta, \eta > 0$ and if $p < q$ then $b_0(I_a) \leq \frac{p}{q-p} b_0(S)$. Then the optimal policy is to use the test after t days have elapsed where t satisfies:

$$b_0(S)\left(\frac{1-p}{1-q}\right)^t \left(1 - \frac{q(1-\alpha\zeta}{(1-p)\alpha(1-\zeta)} \sum_{i=0}^{t-1} \left(\frac{1-q}{1-p}\right)^i\right) \leq b_0(I_a)\left(\frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}\right). \tag{4.8}$$

In general, it is difficult to show an OSLA policy is optimal for $k \geq 2$ tests in the presence of both false negative and false positives. However, the computation of a one-step look-ahead policy is still possible for any $k \geq 2$ which allows us to develop it as a heuristic. This requires us to be able to approximate $V^\pi(\mathbf{b}, k-1)$ for $\mathbf{b}$ in a nontrivial belief space. Fortunately, since costs only depend on knowing beliefs associated with states S and I_a, we can restrict ourselves to a 2-D interpolation grid for the value function with search space

$$\begin{aligned} b(S) + b(I_a) &\leq 1, \\ 0 \leq b(S) &\leq 1, \\ 0 \leq b(I_a) &\leq 1, \end{aligned}$$

as demonstrated in Figure 4.3.

4.4 Simulation Study

Our POMDP-based methods present optimization problems on an individual scale and treat infection probabilities as exogenous to the system. However, in a community, infection probabilities are also dependent on the number of infected and susceptibles remaining within the community itself, as seen in common population-wide SIR models. For this section, we examine the performance of our POMDP-based approaches in a community-based setting by simulating a community of a few thousand agents over a finite time horizon and examining how different policies affect the number

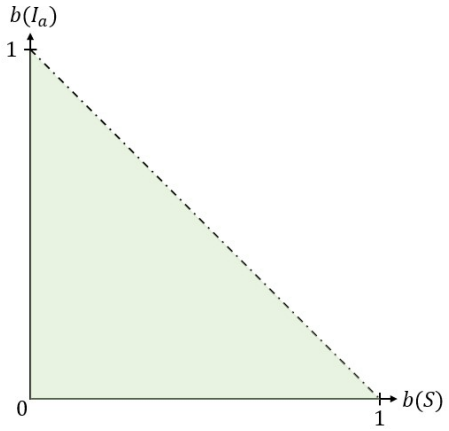

Figure 4.3: Search space for interpolation with imperfect tests.

of detected asymptomatic infections, the number of people who got infected over the time horizon, and the total number of undetected asymptomatic days across the community during the year.

For a single policy, we perform s i.i.d. simulations of a community with N identical people in it over a period of T days. Initial states are chosen for each member of the community for the beginning of the simulation out of susceptible, asymptomatic infected, symptomatic infected/quarantined, and recovered. Note that, when an asymptomatic infected successfully detects the disease, they are effectively the same as a symptomatic infected: their contributions to infections is treated the same as a symptomatic infected as both are now quarantining.

All N people follow the same policy; however, for the i-th person, there is a random testing delay D_i such that no tests are available until day D_i is reached for that individual. This means that people will not necessarily be testing on the same day across the population–the only patients who test on the exact same days as each other are patients with the same testing delay. Testing delay corresponds to the random roll-out of tests to the community caused by random shipping times and varied reaction speeds to the availability of tests by the public.

The most important part of this simulation is how infection probabilities are determined. For this, we use a discrete-time SIR model with the following day-to-day governing equations:

$$\begin{aligned} S_{n+1} &= S_n - \frac{\rho}{N} S_n (I_n + e) \\ \tilde{I}_{n+1} &= \tilde{I}_n + \frac{\rho}{N} S_n (I_n + e). \end{aligned} \tag{4.9}$$

S_n is the number of susceptibles on day n, I_n is the number of infected on day n, $\tilde{I}_n$ is the *cumulative* number of infections that have occurred by day n, N is the number of people in the community, and e is a constant corresponding to the exogenous contribution to infections. ρ is a parameter that determines how quickly the infection spreads and can be tuned to account for many different disease scenarios. We use the $\tilde{I}_n$ so that we do not need to incorporate a governing equation for the number of recovered. As this is a deterministic model, we use it to produce infection probabilities as follows: on day n, the infection probability is $p_n = \frac{I_{n+1}-I_n}{S_n}$. We must also address how $\tilde{I}_n$ handles both asymptomatic and symptomatic infected. If asymptomatic and symptomatic infected are both treated the same, then there is zero point to testing as detecting the disease carries no benefit. The whole premise of this work is that people that have detected their symptoms will behave in a manner that limits their exposure to the remaining susceptibles in the system by quarantining and behaving more cautiously in general. Let $I_{a,n}$ and $I_{s,n}$ be the number of asymptomatic and symptomatic infected at time n, respectively. For Equation (4.9), we set $I_n = \omega I_{s,n} + I_{a,n}$ where $0 \leq \omega < 1$ is a weighting on symptomatic individuals corresponding to how their quarantining behavior reduces risk.

The simulation proceeds on a day-by-day basis as follows:

1. On day n, determine which people are taking tests that day. There is a ζ probability of a test failing to detect an asymptomatic infection. Asymptomatic infected that successfully detect their infections have their state updated to quarantined.

2. Each susceptible individual has their state updated independently of each other according to infection probability p_n, asymptomatic probability α, and recovery probability q.

3. Update $n \to n+1$ and repeat until the time horizon is exhausted.

We are interested in three metrics of success:

1. End Susceptibles (ES): the number of susceptibles remaining in the system at the end of the time horizon. This tells us how many individuals became infected. A larger value is better as that means the spread of the infection was slowed.

2. Total Detected Asymptomatic Infections (TDAI): the total number of asymptomatic infections that were detected via using tests. A larger value is better as this means we successfully detected more infections.

3. Total Undetected Days (TUD): the total number of undetected days per person across the entire horizon. A smaller value is better as that means that the disease was detected earlier as opposed to later in individuals' infected periods.

The TUD metric has the closest relationship to the costs in our POMDP models. Note that the TDAI can be a misleading metric as a large number of asymptomatic cases can be detected; however, those detected cases may be towards the end of the infected periods, in which case a large TUD would still occur. On a related note, a low number for ES can boost the value for TDAI, as more infections lead to a larger number of potential detections.

For our first experiment we consider the following configuration: the simulation horizon is $T = 90$ days, the community has $N = 10,000$ individuals in it with $K = 6$ tests available per person and tests behave perfectly. The asymptomatic rate is $\alpha = 0.41$ and the average recovery time is $\frac{1}{q} = 14$ days. The delay distributions are *i.i.d.* discrete uniform from 3 to 7 days; i.e., $D_i \overset{i.i.d.}{\sim} U_d(3, 7)$. There are zero infected and recovered at time 0; however, the exogenous contribution to infections is $e = 100$, which is 1% of our community population. The symptomatic weighting factor is $\omega = 0.05$, meaning infected only contribute one-twentieth the weight of an asymptomatic individual due to quarantining procedures. The infection parameter is $\rho = 0.25$; Table 4.1 shows the expected ES, TDAI, and TUD values of the system after the 90 days have elapsed without the presence of tests.

In Experiment 1, we consider several policies and conduct 1000 independent simulations per policy. OSLA corresponds to our MDP-derived policy with a static infection probability estimated by averaging the infection probabilities from a 30-day warm-up period with no tests; this can be treated as the community observing the spread in a different community for 30 days and using that data to inform their decisions. Policies of form Dx for some integer x refer to policies that evenly space the test uses every x days. For example, the D15 policy has a user using a test every 15 days until depletion. For each policy, we provide Bonferroni corrected simultaneous 95% confidence intervals for our three metrics of interest in Table 4.1. We note that, while the Opt-S policy performs the best, it is statistically on-par with the D14 policy. However, the D14 policy is

rather arbitrary and coming up with that particular value is not particularly intuitive: one could perhaps take the mean delay, subtract it from 90, and then divide by the number of tests to get approximately 14 days for every test.

Table 4.1: Experiment 1 Results: 95% Simultaneous Confidence Intervals by Method

Method	ES	TDAI	TUD
No Tests	3035.56 ± 11.38	0.00 ± 0.00	36114.27 ± 122.92
OSLA	6442.31 ± 7.07	832.073 ± 2.99	7734.99 ± 27.86
D15	6385.89 ± 7.09	894.50 ± 3.17	8014.08 ± 27.69
D14	6433.43 ± 7.41	848.96 ± 3.25	7754.85 ± 28.22
D13	6431.82 ± 6.73	804.52 ± 3.01	7805.46 ± 26.49
D12	6378.34 ± 7.00	754.45 ± 2.92	8152.74 ± 28.92
D7	5313.49 ± 11.69	466.41 ± 2.14	15264.05 ± 73.44

In Experiment 2, we once again conduct 1000 independent simulations per policy. We will use the exact same parameters are Experiment 1 with the single change that now, false negative results can occur with probability $\zeta = 0.1$. Table 4.2 displays the Bonferroni corrected simultaneous 95% confidence intervals for our three metrics of interest for Experiment 2. Note that, the *No Tests* case is identical to Experiment 1, so we simply repeat the same confidence intervals for that "policy".

Table 4.2: Experiment 2 Results: 95% Simultaneous Confidence Intervals by Method

Method	ES	TDAI	TUD
No Tests	3035.56 ± 11.38	0.00 ± 0.00	36114.27 ± 122.92
OSLA	6244.92 ± 7.88	815.24 ± 3.13	8923.37 ± 33.89
D15	6158.11 ± 8.10	880.68 ± 3.33	9403.04 ± 35.52
D14	6209.05 ± 7.78	840.84 ± 3.11	9087.38 ± 33.78
D13	6225.44 ± 7.83	796.47 ± 3.04	9006.63 ± 32.83
D12	6190.70 ± 7.75	748.33 ± 2.99	9274.92 ± 34.65
D7	5228.49 ± 12.09	463.65 ± 2.15	15852.13 ± 76.50

As opposed to Experiment 1, our OSLA policy outperforms the fixed-spacing policies significantly. It is interesting to note that D14 is still the best performing of the fixed-spacing policies in spite of the fact that tighter OSLA policies result from larger values of the false negative rate ζ.

4.5 Conclusion

Our main goal was to develop easy-to-follow guidelines for the average test user that still have a model-based analytical backing. This was achieved by deriving one-step look-ahead (OSLA) policies which, under certain model conditions, are provably optimal. An OSLA policy allows test

use times to be determined given an initial condition, allowing for the decision maker to follow a list of dates to use tests, as opposed to calculating their day-to-day state probabilities. In cases where OSLA optimality is not guaranteed, the rule can still be implemented to provide a heuristic policy. These rules are derived via an individual-scale model of an SIR model, where infection probabilities are exogenous factors. However, in a community-based setting where both exogenous factors (tourism, business, shipping, etc.) and endogenous factors (number of infected within the community itself) play a role, our individual-based rule systems still outperform common-sense based heuristics such as "every X days" rules.

CHAPTER 5
Future Work and Appendices

Potential future work for the scheduling problem could include multiple patient classes that demonstrate unique unpunctuality distributions and service time distributions in the fluid model. An approach to this would likely involve independent arrival processes for each patient class with a distinct schedule for each class. Further research can also be done on the highly heterogeneous model–currently the best performing methods are random search-type heuristics; more efficient heuristics or new analytical developments could be found in certain scenarios.

For the optimal disease testing problem, future work can consider transitions from the recovered state back to the susceptible state in the SIR model. This is more reflective of real life where individuals can contract diseases such as COVID-19 multiple times throughout their lives. This extension, however, will introduce much difficulty into determining the optimal policy as, due to the cyclic nature of the new Markov chain, the evolution of belief states is unlikely to allow a one-step look-ahead policy to be optimal, so the search for an optimal policy will be complicated. It may, in fact, not be possible to develop an n-step look-ahead policy, which may prevent us from developing the simple guidelines we have in the non-relapsing Markov chain.

APPENDIX 1: SUPPLEMENTS TO CHAPTER 1

5.1 UNC Clinic Dataset Description

We will provide further details on the clinic dataset in this appendix.

UNC Clinic Dataset Variables	
Name	Description
Appointment Status	Takes on the following categorical values: cancelled, completed, left without seen, no show.
Encounter ID	A numerical label for appointments, each observation has a unique label.
Patient ID	A serial number that identify the patient being seen during the appointment.
Doctor ID	A numerical label that identifies the doctor associated with the appointment.
Encounter Scheduled Date	A numerical variable that states which of the 729 days this appointment was scheduled for.
Encounter Scheduled Time	A numerical variable that states the time during the day the appointment was scheduled to start in hours.
Encounter Scheduled Duration	A numerical variable that states the scheduled appointment duration in minutes.
Actual Arrival Time	A numerical variable that states the actual arrival time of the patient in hours.
Doctor Entered Room	A numerical variable that states the actual time the doctor started seeing the patient in hours.
Actual Completion Time	A numerical variable that states what time the patient completed service with the doctor.

The original dataset contains an additional appointment status category called "arrived"; however, this data appears to be blank and only accounts for 6 observations, so it was excluded from the dataset. There are additional variables that state actual arrival date and what date the doctor

entered the room–if these differ from the actual scheduled dates, they are disregarded as data-entry errors–we do not expect any day-transitions to occur between arrival and being seen.

The dataset is fully anonymized; this is an important issue to deal with when examining medical records. The anonymization consists of the following:

- Patient IDs are randomly generated serial numbers and do not link up to any actual database systems, they are only useful for matching the same patient to multiple appointments.
- Encounter Scheduled Dates are anonymized by undergoing a simple transformation. However, this transformation still preserves the order and spacing of the days; that is, encounter scheduled dates that have a difference of n occured n days apart. The range of encounter scheduled dates is 42736 to 43465. We subtract 42735 to rescale all dates to the set $\{1, 2, \ldots, 730\}$.
- Doctor IDs are randomly generated numbers that also do not link up to any actual database systems, they are only useful for matching the same doctor to multiple appointments.

The dataset does not provide any patient or doctor characteristics that may act as identifiers. Further, we have no notion of the type-of-visit for each appointment, we can only infer from the data given; for example, very long service times may correspond to certain types of procedures.

APPENDIX 2: SUPPLEMENTS TO CHAPTER 2

5.2 Mathematical Details

Proof of Proposition 2.4.1. Suppose we have a patient sequence $\vec{s} = (s_1, s_2, \ldots, s_P)$ that is not an ELS. Then, there must exist two consecutive patients in positions k and $k+1$ such that $U_{s_k} \geq_{st} U_{s_{k+1}}$. Now consider the alternative schedule $\vec{\hat{s}} = (s_1, s_2, \ldots, s_{k-1}, s_{k+1}, s_k, s_{k+2}, \ldots, s_P)$ which is identical to $\vec{s}$ with the exception that the values of the k-th and $(k+1)$-th components are switched. We will consider the total patient wait between these two sequences. We determine the wait time of the i-th patient under an appointment-based priority discipline via Lindley's Equation to be

$$W_{i+1} = \max\{0, W_i + S_i + U_i - Ui + 1 - a\} \tag{5.1}$$

for $i \geq 0$, with $W_0 = 0$.

Let $\bar{W} = \sum_{i=1}^{P} \bar{W}_i$ be the total sum of wait times under sequence $\vec{s}$ and $\hat{W} = \sum_{i=1}^{P} \hat{W}_i$ be the total sum of wait times under sequence $\vec{\hat{s}}$. We will show that $\hat{W} \leq_{st} \bar{W}$. Assume, without loss of generality, that $U_{s_{k+1}} = U_i$ and $U_{s_k} = U_j$ for some $i < j$, so $U_i \leq_{st} U_j$.

Clearly $\bar{W}_i = \hat{W}_i$ for $i = 1, \ldots, k-1$ since the patient sequence is identical up until position k; so, for $i = 1, \ldots, k-1$ let $W_i := \bar{W}_i = \hat{W}_i$. We will now define $\hat{W}^* := \hat{W}_k + \hat{W}_{k+1}$ and $\bar{W}^* := \bar{W}_k + \bar{W}_{k+1}$. If $U_i \leq U_j$ implies $\hat{W}^* \leq \bar{W}^*$, then $\hat{W}^* \leq_{st} \bar{W}^*$. Let $Q = S_{k-1} - a + W_{k-1} + U_{s_{k-1}}$ and $R = S_k - a$, then we can write

$$\hat{W}^* = \max\{0, Q - U_i\} + \max\{0, R + U_i - U_j, R + Q - U_j\} \tag{5.2}$$

and

$$\bar{W}^* = \max\{0, Q - U_j\} + \max\{0, R + U_j - U_i, R + Q - U_i\} \tag{5.3}$$

We have six possibilities for $\hat{W}^*$ and six possibilities for $\bar{W}^*$. We will consider which combinations are feasible and proceed to show that $\hat{W}^* \leq \bar{W}^*$ in all feasible cases.

1. $\hat{W}^* = 0$; we do not need to evaluate this case; if it is true, there is nothing to show since $\hat{W}^* \geq 0$.

2. $\hat{W}^* = R + U_i - U_j$; this is feasible if $Q < U_i$ and $R + U_i - U_j > 0$. We must now consider the cases for $\bar{W}^*$

   - $\bar{W}^* = 0$ requires $Q < U_j$ and $R + U_j - U_i < 0$; however, $0 < R + U_i - U_j < R + U_j - U_i$ so this possibility is not feasible.
   - $\bar{W}^* = R + U_j - U_i \geq \hat{W}^* = R + U_i - U_j$ since $U_i \leq U_j$.
   - $\bar{W}^* = R + Q - U_i$ requires $Q < U_j$ and $R + U_j - U_i > R + Q - U_i$ which is a contradiction.
   - $\bar{W}^* = Q - U_j$ requires $Q > U_j$ but this contradicts the fact that $Q < U_i \leq U_j$.
   - $\bar{W}^* = Q - U_j + R + U_j - U_i = Q + R - U_i$ requires $Q > U_j$ but this also contradicts the fact that $Q < U_i \leq U_j$.
   - $\bar{W}^* = Q - U_j + R + Q - U_i$ requires $Q > U_j$ but is contradictory as with the above case.

3. $\hat{W}^* = R + Q - U_j$ requires $Q < U_i$ but also $R + Q - U_i > R + U_j - U_i$, which gives us a contradiction.

4. $\hat{W}^* = Q - U_i$ requires $Q > U_i$ and $R + Q - U_j < 0$.

   - $\bar{W}^* = 0$ requires $Q < U_j$ and $R + U_j - U_i < 0$. This means $R < U_i - U_j \leq 0$ which contradicts with the fact that $R \geq 0$ since $S_k \geq a$.
   - $\bar{W}^* = R + U_j - U_i$ requires $Q < U_j$ and $R + U_j - U_i > 0$. We note that $\bar{W}^* - \hat{W}^* = R + U_j - Q \geq 0$ since $R \geq Q - U_j$ since R must be nonnegative.
   - $\bar{W}^* = R + Q - U_i$ has the same contradiction as in Case 2.
   - $\bar{W}^* = Q - U_j$ needs $Q > U_j$ and $R + Q - U_i < 0$. However, we derive a contradiction from the fact that $R < U_i - Q \leq 0$, but R must be nonnegative.
   - $\bar{W}^* = Q - U_j + R + U_j - U_i = R + Q - U_j$ requires $Q > U_j$ and $R + U_j - U_i > R + Q - U_i$ which is a contradiction.
   - $\bar{W}^* = Q - U_j + R + Q - U_i = 2Q + R - U_j - U_i$ requires $Q > U_j$ and $R + Q - U_i > 0$. Examining $\bar{W}^* - \hat{W}^* = R + Q - U_j \geq$ if $R > U_j - Q$ which will be guaranteed by $R \geq 0$.

5. $\hat{W}^* = Q - U_i + R + U_i - U_j = R + Q - U_j$ is infeasible since $Q > U_i$ and $R + U_i - U_j > R + Q - U_j$ which is a contradiction.

6. $\hat{W}^* = Q - U_i + R + Q - U_j = 2Q + R - U_i - U_j$ requires $Q > U_i$ and $R + Q - U_j > 0$.

   - $\bar{W}^* = 0$ needs $Q < U_j$ and $R + U_j - U_i < 0$, but this means $R < U_i - U_j \leq 0$ so it is infeasible.
   - $\bar{W}^* = R + U_j - U_i$ needs $Q < U_j$ and $R + U_j - U_i > 0$. Then, $\bar{W}^* - \hat{W}^* = R + U_j - U_i - 2Q - R + U_i + U_j = 2U_j - 2Q > 0$ since $Q < U_j$.
   - $\bar{W}^* = R + Q - U_i$ has the same contradiction as in Case 2.
   - $\bar{W}^* = Q - U_j$ needs $Q > U_j and R + Q - U_i < 0$. So, $\hat{W}^* = (R + Q - U_i) + (Q - U_j) < Q - U_j = \bar{W}^*$.
   - $\bar{W}^* = Q - U_j + R + U_j - U_i = R + Q - U_j$ has the same contradiction as in Case 3.
   - $\bar{W}^* = Q - U_j + R + Q - U_i = 2Q + R - U_j - U_i = \hat{W}^*$, so there is nothing to show.

Given the above cases, we have $\hat{W}^* \leq \bar{W}^*$. This means

$$\sum_{i=1}^{k+1} \hat{W}_i \leq_{st} \sum_{i=1}^{k+1} \bar{W}_i.$$

It remains to show that

$$\sum_{i=k+2}^{P} \hat{W}_i \leq_{st} \sum_{i=k+2}^{P} \bar{W}_i.$$

To show this is the case, we consider the terms

$$\hat{W}_{k+2} = \max\{0, S_{k+1} + \hat{W}_{k+1} - a + U_j - U_{s_{k+2}}\} \tag{5.4}$$

and

$$\bar{W}_{k+2} = \max\{0, S_{k+1} + \hat{W}_{k+1} - a + U_i - U_{s_{k+2}}\}. \tag{5.5}$$

To show $\sum_{i=k+2}^{P} \hat{W}_i \leq_{st} \sum_{i=k+2}^{P} \bar{W}_i$, it is sufficient to show $\hat{W}_{k+2} \leq_{st} \bar{W}_{k+2}$ by Equation (5.1) and the fact that the two patient sequences are identical for components $k+2, \ldots, P$. We note that by Equation (5.4) and Equation (5.5), $\hat{W}_{k+2} \leq_{st} \bar{W}_{k+2}$ if and only if $\hat{W}_{k+1} + U_j \leq_{st} \bar{W}_{k+1} + U_i$. We note that

$$\hat{W}_{k+1} + U_j = \max\{0, R + U_i - U_j, R + Q - U_j\} + U_j$$

and

$$\bar{W}_{k+1} + U_i = \max\{0, R + U_j - U_i, R + Q - U_i\} + U_i.$$

We now consider nine possible cases. If for all feasible cases $\hat{W}_{k+1} + U_j \leq \bar{W}_{k+1} + U_i$ for $U_i \leq U_j$, we can say $\hat{W}_{k+1} + U_j \leq_{st} \bar{W}_{k+1} + U_i$. Let $A = \hat{W}_{k+1} + U_j$ and $B = \bar{W}_{k+1} + U_i = U_i$.

1. $A = U_j$ and $B = U_i$ requires $R < U_j - U_i$ for $Q \leq U_i$, $R < U_i - U_j$ for $Q \leq U_j$, $R < U_j - Q$ for $Q > U_i$, and $R < U_i - Q$ for $Q > U_j$. However, this situation is infeasible since $R \geq 0$.

2. $A = U_j$ and $B = R + U_j$, since $R \geq 0$, we have $A \leq B$.

3. $A = U_j$ and $B = R + Q$ requires $Q > U_j$ and since $R \geq 0$, $A < B$.

4. $A = R + U_i$ and $B = U_i$ requires $Q < U_i < U_j$ and $R < U_i - U_j$ from Equations (5.2–5.3), which contradicts $R \geq 0$, so this case is infeasible.

5. $A = R + U_i$ and $B = R + U_j$ does not require further inspection.

6. $A = R + U_i$ and $B = R + Q$ requires $Q < U_i$ and $Q > U_i$ which is contradictory.

7. $A = R + Q$ and $B = U_i$ requires $Q > U_i$ and $R > U_j - Q$ along with $R < U_i - Q$ which leads to a contradiction.

8. $A = R + Q$ and $B = R + U_j$ requires $Q > U_i$, $R > U_j - Q$, $Q < U_j$, and $R > U_i - U_j$. Since $Q < U_j$, $A < B$.

9. $A = R + Q = B$ leaves nothing to be shown.

So, we have $\hat{W}_{k+1} + U_j \leq_{st} \bar{W}_{k+1} + U_i$ and by our earlier argument, we have

$$\sum_{i=k+2}^{P} \hat{W}_i \leq_{st} \sum_{i=k+2}^{P} \bar{W}_i,$$

giving us the desired result that

$$\sum_{i=1}^{P} \hat{W}_i \leq_{st} \sum_{i=1}^{P} \bar{W}_i.$$

□

Lemma 5.2.1. Let X be a continuous random variable with a support set $S \subset \mathbb{R}$ and let $g \geq 0$ be a constant. Let $F(\cdot)$ be the cumulative distribution function of X. Define $Y = \min\{g, X\}$ and $Z = \max\{g, X\}$, then

$$E(Y) = \int_0^g (1 - F(x))dx - \int_{-\infty}^0 F(x)dx$$

and

$$E(Z) = g + \int_0^\infty (1 - F(x))dx - \int_0^g (1 - F(x))dx.$$

Proof. First we derive $E(Y)$,

$$\begin{aligned}
E(Y) &= \int_0^\infty Pr(Y \geq x)dx - \int_{-\infty}^0 Pr(Y \leq x)dx \\
&= \int_0^\infty Pr(X \geq x \cap g \geq x)dx - \int_{-\infty}^0 Pr(X \leq x \cup g \leq x)dx \\
&= \int_0^\infty Pr(X \geq x)\mathbb{I}\{g \geq x\}dx - \int_{-\infty}^0 Pr(X \leq x)dx \\
&= \int_0^g (1 - F(x))dx - \int_{-\infty}^0 F(x)dx.
\end{aligned}$$

Similarly, for $E(Z)$,

$$\begin{aligned}
E(Z) &= \int_0^\infty Pr(Z \geq x)dx - \int_{-\infty}^0 Pr(Z \leq x)dx \\
&= \int_0^\infty Pr(X \geq x \cup g \geq x)dx - \int_{-\infty}^0 Pr(X \geq x \cap g \leq x)dx \\
&= \int_0^\infty \left[Pr(X \geq x) + Pr(g \geq x) - Pr(X \geq x \cap g \geq x\right] dx \\
&= \int_0^\infty (1 - F(x))dx + g - \int_0^g (1 - F(x))dx.
\end{aligned}$$

□

Proof of Proposition 2.6.1. We notice that this can be separated as a minimization of two pieces with respect to a_1 and g separately. We use Lemma 5.2.1 when minimizing with respect to a_1 and g and rewrite

$$\begin{aligned} c_w E(W_1) + c_I E(I_1) &= c_I E(\max\{0, a_1 + U_1\}) - c_w E(\min\{0, a_1 + U_1\}) \\ &= c_I(a_1 - E(\min\{a_1, -U_1\})) - c_w(a_1 - E(\max\{a_1, -U_1\})) \\ &= a_1(c_I - c_w) - c_I\left[\int_0^{a_1}(1 - F_{-U_1}(x))dx - \int_{-\infty}^{0} F_{-U_1}(x)dx\right] \\ &\quad -c_w\left[\int_0^{a_1}(1 - F_{-U_1}(x))dx - a_1 - \int_0^{\infty}(1 - F_{-U_1}(x))dx\right] \end{aligned}$$

where $F_{-U_1}(x)$ is the cumulative distribution function for $-U_1$. Taking the derivative of the above expression, we get

$$\frac{d}{da_1}(c_w E(W_1) + c_I E(I_1)) = c_I - (c_I + c_w)(1 - F_{-U_1}(a_1)) = c_I - (c_I + c_w)F_{U_1}(-a_1).$$

Setting this equal to zero and solving for a_1 yields the optimal value for a_1,

$$a_1^\star = -F_{U_1}^{-1}\left(\frac{c_w}{c_w + c_I}\right).$$

Thus, the first patient's appointment time should only depend on the patient's unpunctuality distribution. However, when determine the optimal inter-appointment gap between two patients, we have to consider both their unpunctuality distributions and the service time distribution of the first patient. Let $R_{i,j} = U_i + S_i - U_j$ and $F_{R_{i,j}}(x)$ be the cumulative distribution function for $R_{i,j}$.

The remaining portion of our objective function can be rewritten as

$$\begin{aligned} c_w E(W_2) + c_I E(I_2) &= c_w E(\max\{0, U_1 + S_1 - U_2 - g\}) + c_I E(\max\{0, g + U_2 - U_1 - S_1\}) \\ &= c_w (E(\max\{g, R_{1,2}\}) - g) - c_I (E(\min\{g, R_{1,2}\}) - g) \\ &= c_I g + c_w \left[\int_0^\infty (1 - F_{R_{1,2}}(x)) dx - \int_0^g (1 - F_{R_{1,2}}(x)) dx \right] \\ &\quad -c_I \left[\int_0^g (1 - F_{R_{1,2}}(x)) dx - \int_{-\infty}^0 F_{R_{1,2}}(x) dx \right]. \end{aligned}$$

Taking the derivative of the above expression, setting it equal to zero, and solving for g, we get

$$g^\star = F_{R_{1,2}}^{-1} \left(\frac{c_w}{c_w + c_I} \right). \tag{5.6}$$

□

APPENDIX 3: SUPPLEMENTS TO CHAPTER 3

5.3 Mathematical Details

We introduce the one-dimensional Skorokhod problem based on the lecture notes of (Dai and Williams, 2017). First, let $\mathbb{D}_+$ be the set of all RCLL functions $x : [0, \infty) \to \mathbb{R}$ with the restriction that $x(0) \geq 0$.

Definition 5.1 One-dimensional Skorokhod Problem. Let $x \in \mathbb{D}_+$. A pair of functions $(z, y) \in \mathbb{D}_+ \times \mathbb{D}_+$ is a solution of the one-dimensional Skorokhod problem for x if the following conditions holds:

1. $z(t) = x(t) + y(t), \quad t \geq 0,$
2. $z(t) \geq 0, \quad t \geq 0,$
3. y satisfies the following:
   (a) $y(0) = 0,$
   (b) y is nondecreasing,
   (c) $\int_0^\infty z(t)dy(t) = 0.$

Note: (z, y) is called a regulation of x where y behaves as the regulator and z is the regulated path. The regulated path is made nonnegative and is comparable to the queuing process from our fluid model. The regulator ensures the boundary reflection property of the regulated path and is comparable to the $\mu i(t)$ process from our fluid model. The $x(t)$ is comparable to $H(t) - \mu t$ in our model.

(Dai and Williams, 2017) next prove that for each $x \in \mathbb{D}_+$, there exists a unique solution to the Skorokhod problem, defining a Skorokhod map $(\phi, \psi) : \mathbb{D}_+ \to \mathbb{D}_+ \times \mathbb{D}_+$ such that $(\phi(x), \psi(x)) = (z, y)$, where (z, y) is the unique solution of the Skorokhod problem for x. The theorem is formally stated as follows:

Theorem 5.3.1. Let $x \in \mathbb{D}_+$. Then there exists a unique solution $(z, y) \in \mathbb{D}_+ \times \mathbb{D}_+$ of the Skorokhod problem for x given by

$$\begin{aligned} y(t) = &\sup_{0\le s\le t} \max\{-x(s), 0\}, \quad t \ge 0, \\ z(t) = &\quad x(t) + y(t), \qquad\qquad t \ge 0. \end{aligned} \tag{5.7}$$

Further, if x is continuous, then both y and z are continuous.

For one of our numerical proofs, we require Theorem 5 from (Dragomir, 2000) as stated below:

Theorem 5.3.2. Suppose $u : [a, b] \to \mathbb{R}$ be a mapping of bounded variation on $[a, b]$ and $f : [a, b] \to \mathbb{R}$ be a p-H-Hölder mapping; i.e., $|f(x) - f(y)| \le H|x - y|^p \ \forall x, y \in [a, b]$. Let $I_n : a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ be a partition of the interval $[a, b]$, $h_i = x_{i+1} - x_i$ $(i = 0, \ldots, n - 1)$, $\xi_i \in [x_i, x_{i+1}]$ $(i = 0, \ldots, n - 1)$, and $\nu(h) = \max\{h_i | i = 0, \ldots, n - 1\}$. Then

$$\int_a^b f(t)du(t) = \sum_{i=0}^{n-1} f(\xi_i)[u(x_{i+1}) - u(x_i)] + R(f, u, I_n, \xi)$$

where

$$\begin{aligned} R(f, u, I_n, \xi) &\le H\left[\tfrac{\nu(h)}{2} + \max_{i=0,\ldots,n-1}\left|\xi_i - \frac{x_i + x_{i+1}}{2}\right|\right]^p V_a^b(u) \\ &\le H(\nu(h))^p V_a^b(u), \end{aligned}$$

where $V_a^b(u)$ is the total variation of u over $[a, b]$.

Proof of Lemma 3.3.1. We first show that $\{a(t), t \in\}$ is a CDF. It suffices to show that $a(t)$ is right continuous. Using the Moore-Osgood theorem on exchanging limits, we have

$$\lim_{t\downarrow s} a(t) = \lim_{t\downarrow s}\lim_{n\to\infty} \bar{A}^n(t) = \lim_{n\to\infty}\lim_{t\downarrow s} \bar{A}^n(t) = \lim_{n\to\infty} \bar{A}^n(s) = a(s),$$

where the first and last equalities follow from the condition that $a^n(t) \to a(t)$ for each t, the second equality is from Moore-Osgood theorem, and the third equality is from the right continuity of $\bar{A}^n(t)$. This shows that $\{a(t), t \in\}$ is a CDF, and consequently, $\{H(t), t \in\}$ is also a CDF.

To prove (3.9), we define

$$H^n(t) = \int_0^T F(t-s,s)d\bar{A}^n(s) = \frac{1}{n}\sum_{i=1}^{n} F(t-a_i^n, a_i^n),\ t \in . \tag{5.8}$$

Then we have the following decomposition.

$$\bar{E}^n(t) - H(t) = [\bar{E}^n(t) - H^n(t)] + [H^n(t) - H(t)],\ t \in .$$

We first focus on $H^n(t) - H(t)$. Fix $\epsilon > 0$. Recall that for any CDF $\tilde{F}$, there exists a simple function $\tilde{G}$ such that $\sup_{t\in} |\tilde{F}(t) - \tilde{G}(t)| \le \epsilon$. Thus for each $s \in [0,T]$, there exists a simple function $G(t,s) = \sum_{k=1}^{K(s)} c_k(s) 1_{B_k(s)}(t), t \in$ such that $\sup_{t\in}(F(t,s) - G(t,s)) \le \epsilon/2$, where $K(s)$ is a positive integer, $c_k(s), k = 1,\dots,K(s)$ are positive constants and $B_k(s), k = 1,\dots,K(s)$ are disjoint subsets of . Now noting that $F(t,s)$ is piecewise continuous in s uniformly for t over each piece, there exists $\delta > 0$ such that when $|s_1 - s_2| \le \delta$ and s_1 and s_2 are in any partition interval,

$$\sup_{t\in} |F(t,s_1) - G(t,s_2)| \le \sup_{t\in} |F(t,s_1) - F(t,s_2)| + \sup_{t\in} |F(t,s_2) - G(t,s_2)| \le \epsilon.$$

This yields that there exists a finite collection of $0 = s_0 < s_1 < \cdots < s_m = T$ such that for $i = 1,\dots,m$,

$$\sup_{s\in[s_{i-1},s_i)} \sup_{t\in} |F(t,s) - G(t,s_{i-1})| \le \epsilon. \tag{5.9}$$

Thus we have

$$\begin{aligned}
\sup_{t\in}|H^n(t)-H(t)| &= \sup_{t\in}\left|\int_0^T F(t-s,s)d[\bar{A}^n(s)-a(s)]\right| \\
&= \sup_{t\in}\left|\sum_{i=1}^m\int_{s_{i-1}}^{s_i} F(t-s,s)d[\bar{A}^n(s)-a(s)]\right| \\
&= \sup_{t\in}\left|\sum_{i=1}^m\int_{s_{i-1}}^{s_i} [F(t-s,s)-G(t-s,s_{i-1})]d[\bar{A}^n(s)-a(s)]\right| \\
&\quad + \sup_{t\in}\left|\sum_{i=1}^m\int_{s_{i-1}}^{s_i} G(t-s,s_{i-1})d[\bar{A}^n(s)-a(s)]\right| \\
&\le \sup_{t\in}\sum_{i=1}^m\int_{s_{i-1}}^{s_i}\sup_{s\in[s_{i-1},s_i)}\sup_{t\in}|F(t,s)-G(t,s_{i-1})|d[\bar{A}^n(s)+a(s)] \\
&\quad + \sum_{i=1}^m\sum_{k=1}^{K(s_i)} c_k(s_i)\sup_{s\in[s_{i-1},s_i)}|\bar{A}^n(s)-a(s)| \\
&\le \epsilon T(\bar{A}^n(T)+a(T)) + \sum_{i=1}^m\sum_{k=1}^{K(s_i)} c_k(s_i)\sup_{s\in[0,T]}|\bar{A}^n(s)-a(s)|,
\end{aligned}$$

which yields that

$$\lim_{n\to\infty}\sup_{t\in}|H^n(t)-H(t)| = \lim_{\epsilon\to 0}\lim_{n\to\infty}\sup_{t\in}|H^n(t)-H(t)| = 0 \tag{5.10}$$

We now consider

$$\bar{E}^n(t)-H^n(t) = \frac{1}{n}\sum_{i=1}^n[1_{\{U_i+a_i^n\le t\}} - F(t-a_i^n,a_i^n)],\ t\in .$$

For $\epsilon>0$, we can find a partition of : $-\infty<t_0<t_1<\cdots<t_{L-1}<t_L<\infty$ such that

$$H(t_{j+1}-)-H(t_j)\le\epsilon,\ j=0,\ldots,L-1,$$

where L is a positive integer that depends on ϵ. For each $t \in$, there exists $k = 0, \ldots, L-1$ such that $t \in [t_k, t_{k+1})$. We then have

$$\begin{aligned}
&\bar{E}^n(t) - H^n(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n \le t\}} - F(t - a_i^n, a_i^n)] \\
&\ge \frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n \le t_k\}} - F((t_{k+1} - a_i^n)-, a_i^n)] \\
&= \frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n \le t_k\}} - F(t_k - a_i^n, a_i^n)] - \frac{1}{n}\sum_{i=1}^{n}[F((t_{k+1} - a_i^n)-, a_i^n) - F(t_k - a_i^n, a_i^n)]. \qquad (5.11)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\bar{E}^n(t) - H^n(t) \\
&\le \frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n < t_{k+1}\}} - F(t_k - a_i^n, a_i^n)] \\
&= \frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n < t_{k+1}\}} - F((t_{k+1} - a_i^n)-, a_i^n)] + \frac{1}{n}\sum_{i=1}^{n}[F((t_{k+1} - a_i^n)-, a_i^n) - F(t_k - a_i^n, a_i^n)].
\end{aligned} \qquad (5.12)$$

We note that $\{1_{\{U_i + a_i^n < t_{k+1}\}} - F((t_{k+1} - a_i^n)-, a_i^n)\}_{i=1}^n$ and $\{1_{\{U_i + a_{i,n} \le t_k\}} - F(t_k - a_i^n, a_i^n)\}_{i=1}^n$ are independent sequences with common mean 0 and are uniformly bounded by 2. From the strong law of large numbers (SLLN) for triangular arrays,

$$\frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n < t_{k+1}\}} - F((t_{k+1} - a_i^n)-, a_i^n)] \to 0, \quad \text{almost surely.} \qquad (5.13)$$

Similarly, we have

$$\frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i + a_i^n \le t_k\}} - F(t_k - a_i^n, a_i^n)] \to 0, \quad \text{almost surely.} \qquad (5.14)$$

We next note that the same second term in (5.12) and (5.11) has the following estimate.

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}[F((t_{k+1}-a_i^n)-,a_i^n)-F(t_k-a_i^n,a_i^n)]\right| \\
&=\left|\frac{1}{n}\sum_{i=1}^{n}[\lim_{\delta\downarrow 0}F(t_{k+1}-a_{i,n}-\delta,a_i^n)-F(t_k-a_{i,n},a_i^n)]\right| \\
&=\left|\lim_{\delta\downarrow 0}H^n(t_{k+1}-\delta)-H^n(t_k)\right| \\
&\le \lim_{\delta\downarrow 0}|(H^n(t_{k+1}-\delta)-H(t_{k+1}-\delta))|+|(H^n(t_k)-H(t_k))|+\left|\lim_{\delta\downarrow 0}H(t_{k+1}-\delta)-H(t_k)\right| \\
&\le 2\sup_{t\in}|H^n(t)-H(t)|+[H(t_{k+1}-)-H(t_k)] \\
&\le 2\sup_{t\in}|H^n(t)-H(t)|+\epsilon. \qquad (5.15)
\end{aligned}$$

Using (5.12) and (5.11), we have

$$\begin{aligned}
&\sup_{t\in}|\bar{E}^n(t)-H^n(t)| \\
&\le \max_{0\le k\le L-1}\left|\frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i+a_i^n<t_{k+1}\}}-F((t_{k+1}-a_i^n)-,a_i^n)]\right. \\
&\left.+\frac{1}{n}\sum_{i=1}^{n}[F((t_{k+1}-a_i^n)-)-F(t_k-a_i^n,a_i^n)]\right| \\
&+\max_{0\le k\le L-1}\left|\frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i+a_i^n\le t_k\}}-F((t_k-a_i^n))]-\frac{1}{n}\sum_{i=1}^{n}[F((t_{k+1}-a_i^n)-,a_i^n)-F(t_k-a_i^n,a_i^n)]\right| \\
&\le \max_{0\le k\le L-1}\left|\frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i+a_i^n<t_{k+1}\}}-F((t_{k+1}-a_i^n)-,a_i^n)]\right| \\
&+\max_{0\le k\le L-1}\left|\frac{1}{n}\sum_{i=1}^{n}[1_{\{U_i+a_i^n\le t_k\}}-F(t_k-a_i^n,a_i^n)]\right| \\
&\quad+4\sup_{t\in}|H^n(t)-H(t)|+2\epsilon
\end{aligned}$$

From (5.13), (5.14), and (5.10), we finally have

$$\lim_{n\to\infty}\sup_{t\in}|\bar{E}^n(t)-H^n(t)|=\lim_{\epsilon\to 0}\lim_{n\to\infty}\sup_{t\in}|\bar{A}^n(t)-H^n(t)|\le\lim_{\epsilon\to 0}2\epsilon=0.$$

This completes the proof. □

Proof of Proposition 3.3.1. We now consider the fluid scaled queue length process $\bar{Q}^n(t), t \in$, and note that for $t \in [0, T]$,

$$\begin{aligned}\bar{Q}^n(t) &= \bar{E}^n(t) - \bar{S}^n(B^n(t)) \\ &= [\bar{E}^n(t) - H(t)] - [\bar{S}^n(B^n(t)) - \mu B^n(t)] + H(t) - \mu B^n(t) \\ &= [\bar{E}^n(t) - H(t)] - [\bar{S}^n(B^n(t)) - \mu B^n(t)] + [H(t) - \mu t] + \mu I^n(t),\end{aligned}$$

where $I^n(0) = 0$, $I^n(t)$ is nondecreasing in t, and it increases only when $\bar{Q}^n(t) = 0$. Thus $(\bar{Q}^n, \mu I^n)$ is the unique solution to the one-dimensional Skorokhod problem associated with

$$X^n(t) = [\bar{E}^n(t) - H(t)] - [\bar{S}^n(B^n(t)) - \mu B^n(t)] + [H(t) - \mu t], t \geq 0.$$

We also note that for $t \geq T$, letting $s = t - T$,

$$\begin{aligned}\bar{Q}^n(s+T) &= \bar{E}^n(T) - \bar{S}^n(B^n(s+T)) \\ &= \bar{Q}^n(T) - [\bar{S}^n(B^n(s+T)) - \bar{S}^n(B^n(T))] \\ &= \bar{Q}^n(T) - [\bar{S}^n(B^n(s+T)) - \bar{S}^n(B^n(T)) - \mu(B^n(s+T) - B^n(T))] - \mu s \\ &\quad + \mu(I^n(s+T) - I^n(T)).\end{aligned}$$

Define for $s \geq 0$,

$$\tilde{X}^n(s) = \bar{Q}^n(T) - [\bar{S}^n(B^n(s+T)) - \bar{S}^n(B^n(T)) - \mu(B^n(s+T) - B^n(T))] - \mu s,$$

and

$$\tilde{Q}^n(s) = \bar{Q}^n(s+T), \quad \tilde{I}^n(s) = I^n(s+T) - I^n(T).$$

Then $\tilde{I}^n(0) = 0$, and $\tilde{I}(s)$ is decreasing in s and increases only when $\tilde{Q}^n(s) = 0$, which yields that $(\tilde{Q}^n, \mu\tilde{I}^n)$ is the unique solution to the one-dimensional Skorokhod problem associated with $\tilde{X}^n$. At last, for $t < 0$,

$$\bar{Q}^n(t) = \bar{E}^n(t).$$

From Lemma 3.3.1 and the SLLN for renewal processes, we have for $\tau > 0$,

$$\sup_{t\in[0,\tau]} |X^n(t) - [H(t) - \mu t]| \to 0, \text{ almost surely,} \tag{5.16}$$

$$\sup_{t\in[0,\tau]} |\tilde{X}^n(t) - \bar{Q}^n(T) + \mu t| \to 0, \text{ almost surely.} \tag{5.17}$$

Using the Lipschitz continuity of the one dimensional Skorokhod map, we have the following fluid limit of $(\bar{Q}^n, I^n)$.

□

Proof of Proposition 3.4.1. Under a given $H = \{H(t); t \le T\}$, the fluid queue length process is given as

$$q(t) = \begin{cases} H(t), & t \le 0, \\ H(t) - \mu t + \mu i(t), & t \in [0, T], \\ q(T) - \mu(t - T), & t \in [T, T + q(T)/\mu], \\ 0, & t \ge T + q(T)/\mu, \end{cases}$$

where the boundary reflection process $i(0) = 0$, and $i(t)$ is nondecreasing and increasing only when $q(t) = 0$ for $t \in [0, T]$. From Lemma 1 in [Armony, Atar, and Honnappa, 2017], consider

$$\tilde{H}(t) = \begin{cases} H(t), & t \le 0, \\ H(t) + \mu i(t), & t \in [0, T]. \end{cases} \tag{5.18}$$

Then under the control $\tilde{H} = \{\tilde{H}(t); t \ge T\}$,

$$\tilde{q}(t) = \begin{cases} H(t), & t \le 0, \\ \tilde{H}(t) - \mu t + \mu\tilde{i}(t), & t \in [0, T], \\ \tilde{q}(T) - \mu(t - T), & t \in [T, T + \tilde{q}(T)/\mu], \\ 0, & t \ge T + \tilde{q}(T)/\mu, \end{cases}$$

However, we note that since $\tilde{H}(t) - \mu t = H(t) + \mu i(t) - \mu t \geq 0$ for all $t \in [0,T]$, $\tilde{i}(t) = 0$ for $t \in [0,T]$, which yields that

$$\tilde{q}(t) = q(t),\ t \in,\ \text{and } 0 = \tilde{i}(t) \leq i(t),\quad t \in [0,T], \tag{5.19}$$

From (5.18) and (5.19), we conclude that

$$J(\tilde{H};\mu) \geq J(H;\mu). \tag{5.20}$$

Therefore, the optimal control is to make the idle time process equal to zero during the time interval $[0,T]$. We now only consider the controls $H = \{H(t); t \leq T\}$ under which $i(t) = 0, t \in [0,T]$, i.e., $H(t) \geq \mu t$ for $t \in [0,T]$. We note that

$$\begin{aligned}
J(H;\mu) &= r(q(T) + \mu T) - c_w \int_0^\infty q(t)dt - c_o q(T)/\mu \\
&= (r - c_o/\mu)q(T) - c_w \int_0^\infty q(t)dt + r\mu T \\
&= (r - c_o/\mu)q(T) - c_w \int_0^T q(t)dt - c_w \int_T^{T+q(T)/\mu} [q(T) - \mu(t-T)]dt + r\mu T \\
&= (r - c_o/\mu)q(T) - c_w \int_0^T q(t)dt - \frac{c_w}{2\mu} q(T)^2 + r\mu T \\
&= (r - c_o/\mu)(H(T) - \mu T) - c_w \int_0^T (H(t) - \mu t)dt - \frac{c_w}{2\mu}(H(T) - \mu T)^2 + r\mu T \\
&= (r - c_o/\mu)H(T) - c_w \int_0^T H(t)dt - \frac{c_w}{2\mu}(H(T)^2 - 2\mu T H(T)) + C_0 \\
&= (r - c_o/\mu + c_w T)H(T) - \frac{c_w}{2\mu} H(T)^2 - c_w \int_0^T H(t)dt + C_0,
\end{aligned} \tag{5.21}$$

where C_0 is a constant. From (5.21), when $r \leq c_o/\mu$, the optimal $q^*(t) = 0$ and $H^*(t) = \mu t$ for $t \in [0,T]$.

Generally, the FCP is now formulated to choose $H = \{H(t); t \leq T\}$ satisfying $H(t) \geq \mu t$ for $t \in [0,T]$ to maximize

$$\tilde{J}(H;\mu) = (r - c_o/\mu + c_w T)H(T) - \frac{c_w}{2\mu} H(T)^2 - c_w \int_0^T H(t)dt. \tag{5.22}$$

From Lemma 1 in [Armony, Atar, and Honnappa, 2017], the FCP is equivalent to choose $\{H(t); t \leq T\}$ satisfying $H(t) \geq \mu t$ for $t \in [0, T]$ to maximize (5.21). Assuming $r \leq c_o/\mu$, we note that the objective function in (5.21) is less than or equal to 0, and it is equal to 0 when $H(t) = \mu t$ for $t \in [0, T]$. Thus if $r \leq c_o/\mu$, an optimal solution is given by $H(t) = \mu t$ for $t \in [0, T]$ and the optimal state process $q(t) = 0$ for $t \in [0, T]$.

Now assume $r > c_o/\mu$. We consider the equivalent objective function in (5.22). Introduce the terminal cost/reward function

$$\phi(x) = (r - c_o/\mu + c_w T)x - \frac{c_w}{2\mu}x^2.$$

It is easily seen that

$$\arg\max_{x \geq \mu T} \phi(x) = \mu T + \frac{\mu(r - c_o/\mu)}{c_w}. \tag{5.23}$$

For an arbitrary control process $\{H(t); t \in [0, T]\}$ satisfying $H(t) \geq \mu t$ for $t \in [0, T]$, we observe that

$$\tilde{J}(H^*; \mu) - \tilde{J}(H; \mu) = \left[\phi\left(\mu T + \frac{\mu(r - c_o/\mu)}{c_w}\right) - \phi(H(T))\right] - c_w \int_0^T (\mu t - H(t))dt \geq 0,$$

on noting that the term in the brackets is nonnegative because of (5.23), and the integral is nonpositive because $H(t) \geq \mu t$ for $t \in [0, T]$. This shows that H^* is optimal when $r > c_O/\mu$. □

Proof of Corollary 1. This result easily follows from the proof of Proposition 3.4.1 that if $H(t) \geq \mu t$ with $r \leq \frac{c_o}{\mu}$, then the objective function in (5.21) is less than or equal to 0 and it is equal to 0 if and only if $H(t) = \mu t$. Similarly, for $H(t) < \mu t$, nonzero idle cost will be accrued leading to a negative objective value. □

Proof of Lemma 3.5.1. For $k = 0$, $\hat{i}(0) = \frac{1}{\mu}\max\{-H(0), 0\} = 0$ and thus $\hat{q}(0) = \hat{H}(0)$. Thus, $q_0 = \hat{q}(0) = \hat{H}(0)$ and $I_0 = \hat{i}(0) = 0$.

Now, assuming $q_k = \hat{q}(k)$ and $\sum_{j=0}^{k} I_j = \hat{i}(k)$ for $k = 0, 1, \ldots, n-1 < K$, we must now show that this holds for n.

To show that $q_n = \hat{q}(n)$, we note that by the induction hypothesis,

$$\begin{aligned} q_n &= \max\{0, \hat{q}(n-1) + a_n - \mu\tfrac{T}{K}\} \\ &= \max\{0, \hat{H}(t_{n-1}) - \mu\tfrac{(n-1)T}{K} + \mu\hat{i}(n-1) + \hat{H}(t_n) - \hat{H}(t_{n-1}) - \mu\tfrac{T}{K}\} \\ &= \max\{0, \hat{H}(t_n) - \mu\tfrac{nT}{K} + \mu\hat{i}(n-1)\} \\ &= \max\left\{0, \hat{H}(t_n) - \mu\tfrac{nT}{K} + \max_{i\in\{0,\ldots,n\}} \max\left\{0, \mu\frac{iT}{K} - \hat{H}(t_i)\right\}\right\} \\ &= \hat{H}(t_n) - \mu\tfrac{nT}{K} + \mu\hat{i}(n) \\ &= \hat{q}(n). \end{aligned}$$

To show that $\sum_{j=0}^{n} I_j = \hat{i}(n)$, we note that by the induction hypothesis,

$$\begin{aligned} \textstyle\sum_{j=0}^{n} I_j &= \hat{i}(n-1) + \tfrac{1}{\mu}\max\{0, \mu\tfrac{T}{K} - (\hat{H}(t_n) + \hat{H}(t_{n-1})) - q_n\} \\ &= \tfrac{1}{\mu}\left(\max_{i\in\{0,\ldots,n-1\}} \max\left\{0, \mu\frac{iT}{K} - \hat{H}(t_i)\right\} + \max\left\{0, \mu\frac{T}{K} - (\hat{H}(t_n) + \hat{H}(t_{n-1})) - q_n\right\}\right) \\ &= \tfrac{1}{\mu}\max_{i\in\{0,\ldots,n\}} \max\left\{0, \mu\frac{iT}{K} - \hat{H}(t_i)\right\} \\ &= \hat{i}(n). \end{aligned}$$

□

Proof of Proposition 3.5.1. Since patients are not accepted into the system if arriving after time T, we have

$$\int_0^\infty q(t)dt = \int_0^T q(t)dt + \frac{c_w}{2\mu}q(T)^2.$$

Using the triangle inequality, we have

$$\begin{aligned} |J(a^\star) - \hat{J}(\mathbf{p}^\star)| &\leq r|H(T) - \hat{H}(T) + c_w \left| \tfrac{T}{K} \textstyle\sum_{i=0}^{K-1} q_i - \int_0^T q(t)dt \right| \\ &\quad + \tfrac{c_w}{2\mu} \left| q_K^2 - q(T)^2 \right| + c_I \left| i(T) - \textstyle\sum_{i=0}^{K-1} I_i \right| \\ &\quad + \tfrac{c_o}{\mu} \left| q_k - q(T) \right|. \end{aligned}$$

We will examine the individual terms:

1. $\left| H(T) - \hat{H}(T) \right|$,
2. $\left| i(T) - \sum_{i=0}^{K-1} I_i \right|$,
3. $|q_K - q(T)|$,
4. $\left| \frac{T}{K} \sum_{i=0}^{K-1} q_i - \int_0^T q(t)dt \right|$, and
5. $\left| q_K^2 - q(T)^2 \right|$.

For the first term, we refer to Theorem 5.3.2. We first note that F is a 1-$\bar{f}$-Hölder type mapping where $\bar{f}$ is the maximum value the derivative of F achieves on $[0, T]$. Further, $a(t)$ is a nondecreasing function on $[0, T]$ and is thus of bounded variation with $V_a^b(a) = \bar{a}(T) < \infty$ which is guaranteed if 0 is within the convex hull. Thus,

$$\left| H(T) - \hat{H}(T) \right| \leq \frac{T}{K} r \bar{f} \bar{a}(T).$$

For the second term, by Lemma 3.5.1, we have

$$\begin{aligned} \left| i(T) - \textstyle\sum_{i=0}^{K-1} I_i \right| &= \left| i(T) - \hat{i}(K) \right| \\ &= \mu^{-1} \left| \sup_{0 \leq s \leq T} \max\{\mu s - H(s), 0\} - \max_{i \in \{0,1,\ldots,K\}} \max\left\{ \mu t_i - \hat{H}(t_i) \right\} \right| \\ &\leq \tfrac{T}{K} \left(1 + \tfrac{\bar{f}\bar{a}(T)}{\mu} \right). \end{aligned}$$

For the third term, we have

$$\begin{aligned} |q_K + q(T)| &= \left|\hat{H}(T) + \mu\hat{i}(K) - (H(T) + \mu i(T))\right| \\ &\leq \left|\hat{H}(T) - H(T)\right| + \mu\left|\hat{i}(K) - i(T)\right| \\ &\leq \frac{T}{K}\left(\bar{f}\bar{a}(T) + 1 + \frac{\bar{f}\bar{a}(T)}{\mu}\right) \\ &= \frac{T}{K}\left(1 + \frac{(\mu+1)\bar{f}\bar{a}(T)}{\mu}\right) \end{aligned}$$

by using our results for the first 2 terms.

For the fourth term, we have

$$\begin{aligned} \left|\frac{T}{K}\sum_{k=0}^{K-1} q_i - \int_0^T q(t)dt\right| &= \left|\frac{T}{K}\sum_{k=0}^{K-1}\left(\hat{H}(t_k) - \mu t_k + \mu\hat{i}(k)\right) - \int_0^T (H(t) - \mu t + \mu i(t))\,dt\right| \\ &\leq \left|\frac{T}{K}\sum_{k=0}^{K-1}\hat{H}(t_k) - \int_0^T H(t)dt\right| + \left|\frac{T}{K}\sum_{k=0}^{K-1}\mu t_k - \int_0^T \mu t dt\right| \\ &\quad + \left|\frac{T}{K}\sum_{k=0}^{K-1}\mu\hat{i}(k) - \int_0^T \mu i(t)dt\right|. \end{aligned}$$

We will examine these 3 components below and using the approximation error of a left Riemann-sum, we have:

$$\begin{aligned} \left|\frac{T}{K}\sum_{k=0}^{K-1}\hat{H}(t_k) - \int_0^T H(t)dt\right| &= \left|\frac{T}{K}\sum_{k=0}^{K-1}\hat{H}(t_k) - \int_0^T \hat{H}(t)dt + \int_0^T \hat{H}(t)dt - \int_0^T H(t)dt\right| \\ &\leq \left|\frac{T}{K}\sum_{k=0}^{K-1}\hat{H}(t_k) - \int_0^T \hat{H}(t)dt\right| + \left|\int_0^T \hat{H}(t)dt - \int_0^T H(t)dt\right| \\ &\leq \bar{h}\frac{T^2}{K} + \bar{f}\frac{T^2}{K}\bar{a}(T) \\ &= \frac{T}{K}\left(\bar{h}T + \bar{f}T\bar{a}(T)\right), \end{aligned}$$

$$\left|\frac{T}{K}\sum_{k=0}^{K-1}\mu t_k - \int_0^T \mu t dt\right| = \frac{T}{K}\mu T,$$

and

$$\left| \frac{T}{K} \sum_{k=0}^{K-1} \mu \hat{i}(k) - \int_0^T \mu i(t) dt \right| \leq \frac{T^2}{K}.$$

For the final and fifth term, we have

$$\begin{aligned}
\left| q_K^2 - q(T)^2 \right| &= |(q_K - q(T))(q_K + q(T))| \\
&\leq |q_K + q(T)| \left(\left| \hat{H}(T) - H(T) \right| + \mu \left| \hat{i}(K) - i(T) \right| \right) \\
&\leq |q_K + q(T)| \, \tfrac{T}{K} \left(\bar{f} \bar{a}(T) + 1 + \tfrac{(\mu+1)\bar{f}\bar{a}(T)}{\mu} \right) \\
&\leq \tfrac{T}{K} \left(1 + \tfrac{(2\mu+1)\bar{f}\bar{a}(T)}{\mu} \right) \left(2\bar{a}(T) + 1 + \tfrac{(\mu+1)\bar{f}\bar{a}(T)}{\mu} \right).
\end{aligned}$$

Combining the five terms gives us our desired result. □

APPENDIX 4: SUPPLEMENTS TO CHAPTER 4

5.4 Mathematical Details

Proof of Proposition 4.3.1. First, we will prove the case for $k = 1$ test when starting in state $\mathbf{e}_S$. We need to show that a one-step look-ahead policy is optimal. By (Ross, 1983), we need only determine that the $\theta = 0$ process is stable and that the set of states where it is optimal to use the test immediately rather than move ahead by single step before using the test exists as a closed set. Note the belief state space is countable, as we start in state $\mathbf{e}_S$ and evolve according to a DTMC. Since our costs are bounded, we have a stable system by (Ross, 1983).

Suppose we have never observed symptoms and have not used the test yet. We define $\mathbf{b}_t^{\theta=0}$ as the belief state vector at time t given that $\theta = 0$ has been observed at times $t \in \{0, 1, \ldots, t\}$. Let $\mathbf{b}_0^{\theta=0} = \mathbf{e}_S$, we can derive the following time-evolution equations for each component in $\mathbf{b}_t^{\theta=0}$ for $t \geq 0$:

$$\begin{aligned}
&b_t^{\theta=0}(S) = \tfrac{1}{\prod_{i=0}^{t-1}\beta(\mathbf{b}_i^{\theta=0})}\,(1-p)^t\,,\\
&b_t^{\theta=0}(I_a) = \tfrac{1}{\prod_{i=0}^{t-1}\beta(\mathbf{b}_i^{\theta=0})}\alpha p\left[\textstyle\sum_{i=0}^{t-1}(1-p)^{t-1-i}(1-q)^i\right]\mathbb{I}_{t\geq 1},\\
&b_t^{(\theta=0)}(I_s) = 0,\\
&b_t^{\theta=0}(R) = \tfrac{1}{\prod_{i=0}^{t-1}\beta(\mathbf{b}_i^{\theta=0})}\alpha p q\left[\textstyle\sum_{i=0}^{t-2}\sum_{j=0}^{i}(1-p)^i(1-q)^{i-j}\right]\mathbb{I}_{t\geq 2},
\end{aligned} \tag{5.24}$$

where $\beta(\mathbf{b}) = b(S)(1-(1-\alpha)p) + b(I_a) + b(R) = P(\theta|\mathbf{b}, c)$.

Note that we only need the equations for $b_t^{\theta=0}(S)$ and $b_t^{\theta=0}(I_a)$ as they are the only states associated with nonzero costs.

To verify closedness, it is sufficient to verify that if $b_t^{\theta=0}(S) \leq \frac{q}{\alpha p} b_t^{\theta=0}(I_a)$ holds true for some $t \geq 0$, then $b_{t+1}^{\theta=0}(S) \leq \frac{q}{\alpha p} b_{t+1}^{\theta=0}(I_a)$ follows. Using the Equation System (5.24), we assume our sufficient condition is true and rewrite it as

$$(1-p)^t \leq q\left[\sum_{i=0}^{t-1}(1-p)^{t-1-i}(1-q)^i\right].$$

Multiply by $(1-p)$ on both sides gives us

$$\begin{aligned}(1-p)^{t+1} &\leq q(1-p)\left[\sum_{i=0}^{t-1}(1-p)^{t-1-i}(1-q)^i\right] \\ &= q\sum_{i=0}^{t-1}(1-p)^{t-i}(1-q)^i \\ &\leq q\sum_{i=0}^{t}(1-p)^{t-i}(1-q)^i.\end{aligned}$$

We note that $(1-p)^{t+1} \leq q\sum_{i=0}^{t}(1-p)^{t-i}(1-q)^i$ is simply

$$b_{t+1}^{\theta=0}(S) = \frac{1}{\prod_{i=0}^{t}\beta(\mathbf{b}_i^{\theta=0})}(1-p)^{t+1} \leq \frac{q}{\alpha p}\left(\frac{\alpha p}{\prod_{i=0}^{t}\beta(\mathbf{b}_i^{\theta=0})}\right)\sum_{i=0}^{t}(1-p)^{t-i}(1-q)^i = \frac{q}{\alpha p}\mathbf{b}_{t+1}^{\theta=0}(I_a).$$

We must now show that $(1-p)^t \leq q\sum_{i=0}^{t-1}(1-p)^{t-1-i}(1-q)^i$ occurs for some finite $t \geq 0$. We write

$$\begin{aligned}q\sum_{i=0}^{t-1}(1-p)^{t-1-i}(1-q)^i - (1-p)^t &= q(1-p)^{t-1}\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i - (1-p)^t \\ &= q(1-p)^{t-1}\left[\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i - \frac{1-p}{q}\right].\end{aligned}$$

We want the above term to be greater than 0 for some finite number t. This requires us to only look at the term inside the square brackets. That is, if

$$\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i - \frac{1-p}{q} > 0 \tag{5.25}$$

for some finite t, we are done. We note that $\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i$ is a geometric series that, if we take the limit as $t \to \infty$, will diverge if $p \geq q$, guaranteeing (5.25) to be true. Suppose $p < q$ which means the infinite geometric series converges to a finite sum. In order for (5.25) to hold true for finite t, we need

$$\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i = \frac{1}{1-\frac{1-q}{1-p}} = \frac{1-p}{q-p} > \frac{1-p}{q}.$$

We can see this is clearly true when simplified.

Now suppose we have $k \geq 2$ tests. We wish to show that the set of states $\left\{\mathbf{b}_t^{\theta=0} : b_t^{\theta=0}(S) \leq \frac{b_t^{\theta=0}(I_a)}{pV^\star(\mathbf{e}_S,k-1)}, \mathbf{b}_0 = \mathbf{e}_S\right\}$ is a closed set of states. Recall Equation System (5.24) and the $k = 1$ case where we showed if $b_t^{\theta=0}(S) \leq \frac{q}{p} b_t^{\theta=0}(I_a)$ holds true, then $b_{t+1}^{\theta=0}(S) \leq \frac{q}{\alpha p} b_{t+1}^{\theta=0}(I_a)$ also holds true. If we substitute $\frac{1}{V^\star(\mathbf{e}_S)}$ for $\frac{\alpha}{q}$, the induction still holds and we have shown that $\left\{\mathbf{b}_t : b_t(S) \leq \frac{b_t(I_a)}{pV^\star(\mathbf{e}_S,k-1)}, \mathbf{b}_0 = \mathbf{e}_S\right\}$ is a closed set of states.

Suppose $k \geq 2$. To show that the optimal stopping criterion will be met in a finite amount of time, we need

$$(1-p)^t \leq \frac{1}{V^\star(\mathbf{e}_S, k-1)} \sum_{i=0}^{t-1} (1-p)^{(t-1-i)}(1-q)^i$$

for some finite t. We rewrite the above expression as

$$(1-p)^t \left[\frac{1}{1-p} \sum_{i=0}^{t-1} \frac{(1-q)^i}{(1-p)^i} - V^\star(\mathbf{e}_S, k-1)\right] \geq 0,$$

which will hold true if

$$\frac{1}{1-p} \sum_{i=0}^{\infty} \frac{(1-q)^i}{(1-p)^i} > V^\star(\mathbf{e}_S, k-1).$$

If $p \geq q$, this is trivially true as the sum diverges. Suppose $p < q$, then

$$\frac{1}{1-p} \sum_{i=0}^{\infty} \frac{(1-q)^i}{(1-p)^i} = \frac{1}{q-p} > \frac{\alpha}{q-p} > \frac{\alpha}{q} \geq V^\star(\mathbf{e}_S, k-1).$$

□

Proof of Corollary 4.3.2. Using Equation System (5.24) and the condition for using a test from Equation (4.3), we find

$$\tau_k = \inf\left\{t \in \mathbb{Z}^+ : \frac{\alpha}{1-p} \sum_{i=0}^{t-1} \left(\frac{1-q}{1-p}\right)^i \geq V^\star(\mathbf{e}_S, k-1)\right\}.$$

Using the partial sum identity

$$\sum_{k=0}^{n-1} r^k = \begin{cases} n, & \text{if } r = 1 \\ \frac{1-r^n}{1-r}, & \text{otherwise,} \end{cases}$$

allows us to simplify to the expression in the corollary. Also, the expression for $V^\star(\mathbf{e}_S, k)$ is simply the expected cost of following the τ_k policy for the first test and performing optimally onwards. $\square$

Proof of Corollary 4.3.1. Suppose the agent starts from $\mathbf{e}_S$ with $i+1$ tests. They can wait one day and immediately use the test, then follow the optimal policy when i tests remain. This results in an expected cost of $(1-p)V^\star(\mathbf{e}_S, i)$. If we call this policy π, then we have $V^\star(\mathbf{e}_S, i+1) \leq V^\pi(\mathbf{e}_S, i+1) < V^\star(\mathbf{e}_S, i)$. The ordering of the τ_k values easily follows from the closed-form expression for τ_k. $\square$

Proof of Proposition 4.3.3. We need to show that $\phi_s(t) \leq \frac{q}{\alpha p_t}\phi_a(t)$ implies $\phi_s(t+1) \leq \frac{q}{\alpha p_{t+1}}\phi_a(t+1)$ for all $t \geq t_0$. Rewriting the first inequality and multiplying by $(1-p_t)$, we get

$$\begin{aligned} \phi_s(t)(1-p_t) &= \phi_s(t+1) \\ &\leq \tfrac{q(1-p_t)}{\alpha p_t}\phi_a(t). \end{aligned}$$

If we can show $\frac{q(1-p_t)}{\alpha p_t}\phi_a(t) \leq \frac{q)}{\alpha p_{t+1}}\phi_a(t+1)$ under our condition, we are done. We consider the difference

$$\frac{1-p_t}{p_t}\phi_a(t) - \frac{1}{p_{t+1}}\phi_a(t+1)$$

$$= \frac{1-p_t}{p_t}\left(p_0(1-q)^{t-1} + (1-p_0)p_1(1-q)^{t-2} + \cdots + p_{t-1}\prod_{j=0}^{t-2}(1-p_j)\right)$$

$$-\frac{1}{p_{t+1}}\left(p_0(1-q)^{t} + (1-p_0)p_1(1-q)^{t-1} + \cdots + p_{t-1}(1-q)\prod_{j=0}^{t-2}(1-p_j)\right)$$

$$-\frac{p_t}{p_{t+1}}\prod_{j=0}^{t-1}(1-p_j)$$

$$= \left(\frac{1-p_t}{p_t}p_0(1-q)^{t-1} - \frac{1-q}{p_{t+1}}p_0(1-q)^{t-1}\right)$$

$$+\left(\frac{1-p_t}{p_t}p_1(1-p_0)(1-q)^{t-2} - \frac{1-q}{p_{t+1}}p_1(1-p_0)(1-q)^{t-2}\right)$$

$$+\cdots+\left(\frac{1-p_t}{p_t}p_{t-1}\prod_{j=0}^{t-2}(1-p_j) - \frac{1-q}{p_{t+1}}p_{t-1}\prod_{j=0}^{t-2}(1-p_j)\right)$$

$$-\frac{p_t}{p_{t+1}}\prod_{j=0}^{t-1}(1-p_j)$$

$$= \left(\frac{1-p_t}{p_t} - \frac{1-q}{p_{t+1}}\right)\phi_a(t) - \frac{p_t}{p_{t+1}}\phi_s(t),$$

which is nonpositive under our condition. □

Proof of Corollary 2.

$$\left(\frac{1-p_t}{p_t} - \frac{1-q}{p_{t+1}}\right)\phi_a(t) - \frac{p_t}{p_{t+1}}\phi_s(t)$$

is strictly negative if

$$\left(\frac{1-p_t}{p_t} - \frac{1-q}{p_{t+1}}\right) \leq 0,$$

since $\frac{p_t}{p_{t+1}}\phi_s(t)$ is strictly positive. □

Proof of Proposition 4.3.4. We note that Equation (4.8) is a one-step look-ahead (OSLA) policy for the optimal stopping problem associated with a single test. Thus, we will follow the same

schematic based upon (Ross, 1983) as we did in the proof of Proposition 4.3.1, we must prove that the set of OSLA states is closed and can be reached in a finite amount of time. We will redefine System (5.24 to handle starting in a possibly asymptomatic-infected belief state. Let the initial belief state be $\mathbf{b}_0$, then

$$\begin{aligned} b_t^{\theta=0}(S) = & \quad \frac{b_0(S)}{\prod_{i=1}^{t-1}\beta(\mathbf{b}_i^{\theta=0})}(1-p)^t, \\ b_t^{\theta=0}(I_a) = & \quad \frac{b_0(I_a)(1-q)^t + b_0(S)p(1-p)^{t-1}\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i}{\prod_{i=0}^{t-1}\beta(\mathbf{b}_i^{\theta=0})}. \end{aligned} \tag{5.26}$$

Note that we will not explicitly state beliefs for state I_s and R as they are not associated with costs. Suppose we have a single test. The OSLA condition is to stop at time t if

$$\frac{\alpha}{q}b_t^{\theta=0}(S) + \frac{\alpha\zeta}{q}b_t^{\theta=0}(I_a) \leq \left(\frac{q+\alpha\zeta(1-q)}{q}\right)b_t^{\theta=0}(I_a) + \frac{\alpha}{q}\left(1-p(1-\zeta)\right)b_t^{\theta=0}(S),$$

which can be simplified to

$$b_t^{\theta=0}(S) \geq \frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}b_t^{\theta=0}(I_a).$$

Substituting in our values for $b_t^{\theta=0}(S)$ and $b_t^{\theta=0}(I_a)$ from System (5.26), we have

$$b_0(S)(1-p)^t \leq \left(\frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}\right)\left(b_0(I_a)(1-q)^t + b_0(S)p(1-p)^{t-1}\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i\right),$$

which can be rewritten as

$$b_0(S)\left(\frac{1-p}{1-q}\right)^t\left(1 - \frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\sum_{i=1}^{t-1}\left(\frac{1-q}{1-p}\right)^i\right) \leq b_0(I_a)\left(\frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}\right). \tag{5.27}$$

This expression allows us to examine the left side of the inequality as the right side is constant in t and must be a nonnegative quantity. We wish to show the set of states for which this inequality holds is closed; that is, once this inequality holds for some time $t = t_0$, it also holds for all $t > t_0$ and that this inequality holds true after a finite amount of time. If $p \geq q$, this is trivially true as the LHS of 5.27 is strictly decreasing towards $-\infty$.

If $p < q$, we note that

$$\lim_{t\to\infty}\left(\frac{1-p}{1-q}\right)^t = \infty$$

and

$$\begin{aligned}\lim_{t\to\infty}\left(1-\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i\right) &= \left(1-\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\left(\frac{1}{1-\left(\frac{1-q}{1-p}\right)}\right)\right)\\ &= \frac{\alpha(1-\zeta)(q-p)-q(1-\alpha\zeta)}{\alpha(1-\zeta)(q-p)}\\ &< 0.\end{aligned}$$

This only shows that the inequality holds true after a finite amount of time, but does not guarantee closedness. For closedness, we assume an induction hypothesis that the inequality holds true at time t and consider $t+1$.

$$\begin{aligned}&b_0(S)\left(\frac{1-p}{1-q}\right)^{t+1}\left(1-\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\left(\frac{1-\left(\frac{1-q}{1-p}\right)^{t+1}}{1-\left(\frac{1-q}{1-p}\right)}\right)\right)\\ &= \left(\frac{1-p}{1-q}\right)\left(b_0(S)\left(\frac{1-p}{1-q}\right)^t\left(1-\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\sum_{i=0}^{t-1}\left(\frac{1-q}{1-p}\right)^i-\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\left(\frac{1-q}{1-p}\right)^t\right)\right)\\ &\le \left(\frac{1-p}{1-q}\right)\left(b_0(I_a)\frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}-b_0(S)\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\right).\end{aligned}$$

We note that

$$\left(\frac{1-p}{1-q}\right)\left(b_0(I_a)\frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}-b_0(S)\frac{q(1-\alpha\zeta)}{(1-p)\alpha(1-\zeta)}\right) \le b_0(I_a)\frac{q(1-\alpha\zeta)}{\alpha p(1-\zeta)}$$

if and only if

$$b_0(I_a) \le \frac{p}{q-p}b_0(S),$$

which is our required condition. □

BIBLIOGRAPHY

Ahmadi-Javid, A., Z. J. and Klassen, K. J. (2017). Outpatient appointment systems in healthcare: a review of optimization studies. *European Journal of Operational Research*, 258(1):3–34.

Aldila, D., M. S. S. H. A. K. M. A. F. S. A. I. Y. R. A. and Samiadji, B. M. (2022). Optimal control problem arising from covid-19 transmission model with rapid-test. *Results in Physics*, 37:105501. https://doi.org/10.1016/j.rinp.2022.105501.

Alexopoulos, C., D. G. J. F. D. K. and Wilson, J. R. (2008). Modeling patient arrivals in community clinics. *Omega*, 36(1):33–43.

Alvarado, M. M. and Ntaimo, L. (2018). Chemotherapy appointment scheduling under uncertainty using mean-risk stochastic integer programming. *Health Care Management Science*, 21(1):87–104.

Armony, M., R. A. and Honnappa, H. (2019). Asymptotically optimal appointment schedules. *Mathematics of Operations Research*, 44(4):1345–1380.

Asadi, Y., S. S. Z. and Yaghoobi, M. A. (2019). A stochastic mixed integer programming model for outpatient appointment scheduling considering late cancellation and physician lateness. *International Journal of Hospital Research*, 8(2).

Bailey, N. T. (1952). A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society*, 14:185–199.

Berg, B. P., B. T. D. S. A. E. T. R. and Huschka, T. (2014). Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research*, 50:24–37.

Brahimi, M. and Worthington, D. J. (1991). Queueing models for out-patient appointment systems: a case study. *Journal of the Operational Research Society*, 42(9):733–746.

Buhat, C. A. H., J. C. C. D. E. F. O. F. J. F. R. and Mamplata, J. B. (2021). Optimal allocation of covid-19 test kits among accredited testing centers in the philippines. *Journal of Healthcare Informatics Research*, 5:54–69.

Castaing, J., A. C. B. T. D. and Weizer, A. (2016). A stochastic programming approach to reduce patient wait times and overtime in an outpatient infusion center. *IIE Transactions on Healthcare Systems Engineering*, 6(3):111–125.

Cayirli, T. E., E. A. V. and Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9:47–58.

Cayirli, T. E. and Veral, H. R. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12:519–549.

Cayirli, T. E. and Yang, K. K. (2019). Altering the environment to improve appointment system performance. *Service Science*, 11(2):138–154.

Chau, N. V. V., V. T. L. N. T. D. L. M. Y. N. N. Q. M. L. M. H. N. M. N. N. T. D. D. N. H. M. L. A. N. L. T. H. N. L. N. T. N. N. T. H. N. N. T. T. H. E. K. N. T. P. D. T. C. X. T. T. H. N. T. P. T. N. G. T. R. B. G. T. T. T. N. T. T. N. T. B. T. C. T. G. T. and Tan, L. V. (2020). The natural history and transmission potential of asymptomatic severe acute respiratory syndrome coronavirus 2 infection. *Clinical Infectious Diseases*, 71(10):2679–2687.

Chen, P.-S., H.-W. C. M. D. M. A. R. M. B. G. and Latina, C. G. E. (2023). Patient unpunctuality's effect on appointment scheduling: a scenario-based analysis. *Healthcare*, 11(2):231. https://doi.org/10.3390/healthcare11020231.

Chen, X., L. W.-J. D. and Thomas, N. (2016). Patient flow scheduling and capacity planning in a smart hospital environment. *IEEE Access*, 4:135–148.

Cheong, S., R. R. B. and Fontanesi, J. (2013). Modeling scheduled patient punctuality in an infusion center. *Lecture Notes in Management Science*, 5:45–46.

Choi, W. and Shim, E. (2021). Optimal strategies for social distancing and testing to control covid-19. *Journal of Theoretical Biology*, 512:110568. https://doi.org/10.1016%2Fj.jtbi.2020.110568.

Dai, J. G. and Williams, R. (2017). Skorokhod problems. *Lecture Notes at University of California San Diego*, pages 49–72.

Deceuninck, M., D. F. and Vuyst, S. D. (2018). Outpatient scheduling with unpunctual patients and no-shows. *European Journal of Operational Research*, 265:195–207.

Deceuninck, M., S. D. V. and Fiems, D. (2019). An efficient control variate method for appointment scheduling with patient unpunctuality. *Simulation Modelling Practice and Theory*, 90:116–129.

Denton, B. and Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 23:1003–1016.

Dexter, F., A. M. R. D. T. M. H. and Lubarsky, D. A. (1999). An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patient preferences for surgical waiting time. *Anesthesia and Analgesia*, 89:7–20.

Dexter, F. (1999). Design of appointment systems to minimize patient waiting times: a review of computer simulation and patient survey studies. *Anesthesia and Analgesia*, 89:925–931.

Dogru, A. K. and Melouk, S. H. (2019). Adaptive appointment scheduling for patient-centered medical homes. *Omega*, 85:166–181.

Dragomir, S. S. (2000). On the ostrowski's inequality for riemann-stieltjes integral. *Korean Journal of Computational & Applied Mathematics*, 7:611–627.

Ely, J., A. G. O. J. and Steiner, J. (2021). Optimal test allocation. *Journal of Economic Theory*, 193:105236.

Erdogan, S. A., A. G. and Denton, B. (2015). On-line appointment sequencing and scheduling. *IIE Transactions*, 47(11):1267–1286.

Gautam, N. (2012). Analysis of queues: Methods and applications. 1st ed. CRC Press, Taylor & Francis Group, Boca Raton, FL.

Gerchak, Y., D. G. and Henig, M. (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334.

Ghosh, S., A. S. J. C. C. H. and Ghosh, D. (2021). Optimal test-kit-based intervention strategy of epidemic spreading in heterogeneous complex networks. *Chaos*, 31(7):071101.

Gocgun, Y. (2018). Dynamic scheduling with cancellations: an application to chemotherapy appointment booking. *International Journal of Optimization and Control: Theories & Applications*, 8(2):161–169.

Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: challenges and opportunities. *IIE Transactions*, 40:800–819.

Harper, P. R. and Gamlin, H. M. (2003). Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, 25:207–222.

Heaney, D. J., J. G. H. and Porter, A. M. (1991). Factors influencing waiting times and consultation times in general practice. *British Journal of General Practice*, 41:315–319.

Ho, C.-J. and Lau, H.-S. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science*, 38(12):1750–1762.

Ho, C.-J. and Lau, H.-S. (1999). Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operational Research*, 112:542–553.

Honnappa, H., R. J. A. R. W. (2015). A queuing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems*, 80:71–103.

Hribar, M. R., A. E. H. I. H. G. L. G. R. A. K. A. R. L. D. J. K. L. W. and Chiang, M. F. (2019). Data driven scheduling for improving patient efficiency in ophthalmology clinics. *Ophthalmology*, 126(3):347–354.

Jafarnia-Jahromi, M. and Jain, R. (2020). Non-indexability of the stochastic appointment scheduling problem. *Automatica*, 118.

Jansson, B. (1966). Choosing a good appointment system: a study of queues of the type (d, m, 1). *Operations Research*, 14:292–312.

Jiang, B., J. T. and Yan, C. (2019). A stochastic programming model for outpatient appointment scheduling considering unpunctuality. *Omega*, 82:70–82.

Jiang, R., M. R. and Xu, G. (2019). Data-driven distributionally robust appointment scheduling over wasserstein balls. arXiv:1907.03219v1.

Kemper, B., C. A. K. and Mandjes, M. (2014). Optimized appointment scheduling. *European Journal of Operational Research*, 239(1):243–255.

Kim, S. and Giachetti, R. E. (2006). A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on Systems, Man, and Cybernetics–Part A*, 36(6):1211–1219.

Kim, S.-H., W. W. and Cha, W. C. (2018). A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS Journal on Computing*, 30(1):181–199.

Klassen, K. J. and Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, 14:83–101.

Klassen, K. J. and Yoogalingam, R. (2014). Strategies for appointment policy design with patient unpunctuality. *Decision Sciences*, 45(5):881–991.

Kocas, C. (2015). An extension of osuna's model to observable queues. *Journal of Mathematical Psychology*, 66:53–58.

Kolisch, R. and Sickinger, S. (2008). Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum*, 30:375–395.

Kong, Q., C. Y. L. C. P. T. and Zheng, Z. (2016). Appointment sequencing: why the smallest-variance-first rule may not be optimal. *European Journal of Operational Research*, 255(3):809–821.

LaGanga, L. R. and Lawrence, S. R. (2007). Clinical overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276.

Lee, D. K. K. and Zenios, S. A. (2009). Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations Research*, 57(4):852–865.

Lee, C., X. L. Y. L. and Zhang, L. (2021). Optimal control of a time-varying double-ended production queueing model. *Stochastic Systems*, 11(2):140–173.

Li, X., J. W. and Fung, R. Y. K. (2018). Approximate dynamic programming approaches for appointment scheduling with patient preferences. *Artificial Intelligence in Medicine*, 85:16–25.

Li, R., S. P. B. C. Y. S. T. Z. W. Y. and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493.

Lin, J., K. M. and Lawley, M. (2011). Optimal and approximate algorithms for sequential clinical scheduling with no shows. *IIE Transactions on Healthcare Systems Engineering*, 1(1):20–36.

Lippi, G. and Plebani, M. (2020). Asymptomatic covid-19 transmission: the importance of avoiding official miscommunication. *Diagnosis*, 7(4):347–348.

Liu, L. and Liu, X. (1998). Dynamic and static job allocation for multi-server systems. *IIE Transactions*, 30:845–854.

Lu, C.-C., S.-W. L. H.-J. C. and Ying, K.-C. (2017). Optimal allocation of cashiers and pharmacists in large hospitals: a point-wise fluid-based dynamic queueing network approach. *IEEE Access*, 6:2859–2870.

Luo, L., Y. Z.-B. T. H.-Y. S. Q. S. X. H. and Guo, Z. (2016). A simulation model for outpatient appointment scheduling with patient unpunctuality. *International Journal of Simulation and Process Modelling*, 11(3-4):281–291.

Lyng, G. D., N. E. S. C. J. K. D. O. G. and Berke, E. M. (2021). Identifying optimal covid-19 testing strategies for schools and businesses: Balancing testing frequency, individual test technology, and cost. *PLOS ONE*, 16(3):e0248783. https://doi.org/10.1371/journal.pone.0248783.

Ma, Q., J. L. Q. L. L. K. R. L. W. J. Y. W. and Liu, M. (2021). Global percentage of asymptomatic sars-cov-2 infections among the tested population and individuals with confirmed covid-19 diagnosis. *JAMA Network Open*, 4(12).

Madubueze, C. E., S. D. I. O. O. (2020). Controlling the spread of covid-19: Optimal control analysis. *Computational and Mathematical Models in Medicine*, 2020. https://doi.org/10.1155/2020/6862516.

Mak, H. Y., Y. R. and Zhang, J. (2014). Appointment scheduling with limited distributional information. *Management Science*, 61(2):316–334.

Mancilla, C. and Storer, R. (2012). A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, 44(8):655–670.

Mandelbaum, A., P. M. N. T. S. K. R. L. and Bunnell, C. A. (2020). Data-driven appointment-scheduling under uncertainty: the case of an infusion unit in a cancer center. *Management Science*, 66(1):243–270.

Mansourifard, F., P. M. M. Z. and Krishnamachari, B. (2018). A heuristic policy for outpatient surgery appointment sequencing: newsvendor ordering. *Proceedings of the 2018 International Conference on Industrial Engineering and Operations Management*, pages 233–243.

Mehrizi, A., M. K. S. F. and Faradonbeh, M. S. S. (2022). Asymptotic analysis of multi-class advance patient scheduling. http://dx.doi.org/10.2139/ssrn.4055715.

Moradi, S., M. N. S. M. and Zolfagharinia, H. (2022). The utilization of patients' information to improve the performance of radiotherapy centers: a data-driven approach. *Computers & Industrial Engineering*, (A):108547.

Muthuraman, K. and Lawley, M. (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9):820–837.

Oh, H. J., A. M. H. B. K. A. and Ptaszkiewicz, T. (2013). Guidlines for scheduling in primary care under different patient types and stochastic nurse and provider service times. *IIE Transactions of Healthcare Systems Engineering*, 3(4):263–279.

Olivares, A. and Staffetti, E. (2021). Optimal control-based vaccination and testing strategies. *Computer Methods and Programs in Biomedicine*, 211:106411. doi: 10.1016/j.cmpb.2021.106411.

Oran, D. P. and Topol, E. J. (2020). Prevalence of asymptomatic sars-cov-2 infection. *Annals of Internal Medicine*, 173(5):362–367.

Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology*, 29(1):82–105.

Pan, X., N. G. and Xie, X. (2019). A bender's decomposition approach for appointment scheduling of unpunctual patients in a multi-server setting. *2019 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 64–68.

Pan, X., N. G. and Xie, X. (2021a). Appointment scheduling and real-time sequencing strategies for patient unpunctuality. *European Journal of Operational Research*, 295:246–260.

Pan, X., N. G. and Xie, X. (2021b). A stochastic approximation approach for managing appointments in the presence of unpunctual patients, multiple servers and no-shows. *International Journal of Production Research*, 59(10):2996–3016.

Piguillem, F. and Shi, L. (2022). Optimal covid-19 quarantine and testing policies. *The Economic Journal*, 132(647):2534–2562.

Pinedo, M. L. (2012). Scheduling: Theory, algorithms, and systems. 5th ed. Springer, Springer Cham Heidelberg, NY.

Riise, A., C. M. and Burke, E. (2016). Modelling and solving generalised operational surgery scheduling problems. *Computers & Operations Research*, 66:1–11.

Ross, S. M. (1983). Introduction to stochastic dynamic programming. 1st ed. Academic Press, Inc., San Diego, CA.

Samorani, M. and Ganguly, S. (2016). Optimal sequencing of unpunctual patients in high-service-level clinics. *Production and Operations Management*, 25(2):330–346.

Sasmita, N. R., M. I. S. S. and Chongsuvivatwong, V. (2020). Optimal control on a mathematical model to pattern the progression of coronavirus disease 2019 (covid-19) in indonesia. *Global Health Research and Policy*, 38(38). https://doi.org/10.1186/s41256-020-00163-2.

Tai, G. and Williams, P. (2012). Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine*, 108(2):467–479.

Tatsuki, O., H. N. and Igarashi, Y. (2023). Optimal covid-19 testing strategy on limited resources. *PLOS ONE*, 18(2):e0281319. https://doi.org/10.1371/journal.pone.0281319.

Thunstrom, L., S. C. N. D. F. M. A. and Shogren, J. F. (2020). The benefits and costs of using social distancing to flatten the curve for covid-19. *Journal of Benefit-Cost Analysis*, 11(2):179–195.

Tsai, P.-F. J. and Teng, G.-Y. (2014). A stochastic appointment scheduling system on multiple resources with dynamic call-in sequence and patient no-shows for an outpatient clinic. *European Journal of Operational Research*, 239(2):427–436.

Tsay, C., F. L. M. A. S. and Baldea, M. (2020). Modeling, state estimation, and optimal control for the us covid-19 outbreak. *Scientific Reports*, 10:10711. https://doi.org/10.1038/s41598-020-67459-8.

Vatcheva, K. P., J. S. T. O. J. C. M. T. H. and Villalobos, M. C. (2021). Social distancing and testing as optimal strategies against the spread of covid-19 in the rio grande valley of texas. *Infectious Disease Modelling*, 6:729–742.

Wells, C. R., J. P. T. A. P. S. M. M. G. K. B. S. R. H. M. M. C. F. and Galvani, A. P. (2021). Optimal covid-19 quarantine and testing strategies. *Nature Communications*, 12(356). https://doi.org/10.1038/s41467-020-20742-8.

White, M. J. B. and Pike, M. C. (1964). Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care*, 2:133–145.

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

Zacharias, C. and Armony, M. (2017). Joint panel sizing and appointment scheduling in outpatient care. *Management Science*, 63(11):3978–3997.

Zhou, S. and Yue, Q. (2021). Sequencing and scheduling appointments for multi-stage service systems with stochastic service durations and no-shows. *International Journal of Production Research*, 60(5):1–20.

Zhu, H., Y. C. E. L. and Liu, X. (2018). Outpatient appointment scheduling with unpunctual patients. *International Journal of Production Research*, 56(5):1982–2002.