informs
https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *Manufacturing & Service Operations Management*

ISSN 1523-4614, EISSN 1526-5498

Fair and Efficient Scheduling with Stratified No-Show Prediction

(Authors' names are not included for peer review)

Abstract. **Problem Definition:** Outpatient clinics face significant challenges in appointment scheduling due to high demand, variability, and the persistent problem of no-shows. Although the development of predictive models allows the design of more advanced scheduling strategies, they also raise ethical concerns by potentially exacerbating disparities in healthcare access. Our research examines the trade-offs between operational efficiency and equitable service delivery. We explore the integration of stratified predictive show-up probabilities into scheduling decisions and evaluate its impact on metrics such as provider overtime, patient waiting time, and group and individual unfairness. **Methodology/Results:** Using numerical and theoretical analyses, we systematically assess the interaction between efficiency and fairness in different systems. We consider a comprehensive range of scheduling levers, including slot-length design, overbooking strategies, and patient sequencing, while maintaining a practical focus on implementable policies. Our findings challenge the necessity of stratified prediction, showing that well-designed scheduling strategies can achieve strong performance without exacerbating group disparities. We also show that, while it is possible to simultaneously reduce overtime and individual unfairness, achieving shorter average waiting times often comes at the expense of increased individual unfairness. **Managerial Implications:** Our results are useful for the design of scheduling policies that strike a balance between efficiency and fairness in healthcare delivery.

Key words: Appointment Scheduling, Fairness, Stratified Prediction

1. Introduction

Appointment systems in outpatient clinics are a crucial component of healthcare access and have attracted substantial attention from the operations management community. Researchers have developed a wide range of analytical and numerical methodologies to address the challenges inherent in these systems, which stem from uncertainty, high demand relative to capacity, and the importance of timely access to care. These efforts aim to efficiently balance patient wait times and provider overtime while improving access and service quality (Ahmadi-Javid et al. 2017).

Compounding these challenges is the persistent problem of no-shows in outpatient clinics. No-shows not only disrupt clinical workflows but also exacerbate inefficiencies, straining already limited resources. No-shows cost the U.S. healthcare system over $150 billion annually and cost individual physicians an average of $200 per unused time slot (Forbes 2019). In response, overbooking and various other appointment scheduling strategies have been developed to address inefficiencies caused by no-shows (Ahmadi-Javid et al. 2017). Advances in machine learning, together with the growing availability of data, have facilitated the development of more accurate no-show prediction models (Alaeddini et al. 2011, Huang and Hanauer 2014, Liu et al. 2022), which can be applied to optimize appointment scheduling and various other operations management strategies in healthcare.

However, relying on predicted no-show rates raises ethical concerns as it risks accelerating existing disparities. Research provides ample evidence that no-show rates vary significantly between racial and socioeconomic groups, with disadvantaged patients more likely to miss appointments due to factors such as transportation problems, inflexible work schedules, and lack of childcare options (Samuels et al. 2015, Dantas et al. 2018, Parsons et al. 2021). Predictive models that rely on patient demographic information can perpetuate systemic biases, disproportionately disadvantaging those already facing substantial barriers to care. For example, a case study by Murray et al. (2020) of a predictive model developed by Epic Systems, Inc. identified potential explicit discrimination. Models that include personal attributes as features, such as ethnicity, financial status, religion, and body mass index, could lead to overbooking practices that systematically reduce access to care for already marginalized individuals who are predicted to have higher no-show rates.

More broadly, healthcare delivery and outcomes in the United States are marred by stark disparities between different racial and socioeconomic groups. The annual National Healthcare Quality and Disparities Reports, compiled by the Agency for Healthcare Research and Quality (2024) for 21 consecutive years, provide a comprehensive overview of trends in healthcare access and quality, stating that there are racial and socioeconomic disparities in receiving timely appointments. Black and Hispanic patients face significant barriers to timely access to primary care (Wisniewski and Walker 2020), and clinic times are significantly longer for racial and ethnic minorities, people with lower levels of education, and unemployed patients (Ray et al. 2015). These findings underscore the ethical imperative to address these inequities, especially in healthcare, where timely access to care can critically affect people's well-being.

Our study investigates outpatient appointment scheduling, focusing on the critical interplay between predictive analytics and equitable service delivery. Specifically, we examine the implications of using stratified no-show probability predictions in scheduling decisions. We focus on

understanding the trade-offs between traditional operational performance measures, such as provider overtime and average patient waiting time, and in-clinic service experience disparities.

Building on existing appointment scheduling literature, our study extends the scope by considering a comprehensive range of scheduling levers, including slot length design, overbooking strategies, and patient sequencing. However, we deliberately limit our analysis to scheduling policies that are straightforward to implement in practice. This pragmatic focus reduces the parameter search space while restricting attention to approaches that remain feasible for adoption in practice.

We evaluate the trade-offs between operational efficiency and fairness across a wide spectrum of scheduling policies and system configurations. We critically reassess the role of stratified no-show predictions in optimizing operational performance. A key contribution of our work is to challenge the assumption that such stratification is essential for improvement. Our goal is to bridge the gap between operational efficiency and equitable service delivery, contributing to the development of more inclusive and effective scheduling frameworks.

We rely on both numerical and analytical methods. Given the inherent analytical complexity of appointment scheduling, we first adopt a numerical approach to explore various trade-offs and derive key managerial insights. This approach allows us to investigate a wide range of scenarios and uncover nuanced patterns that are challenging to capture analytically. Then, to complement our numerical findings and enhance generalizability, we substantiate key observations with the theoretical analysis of a fluid model. In general, our results lead to a deeper understanding of the intricate relationships between efficiency and fairness in scheduling. By highlighting the conditions under which certain trade-offs emerge, our study offers valuable guidance to decision-makers in healthcare and paves the way for designing more equitable and efficient systems.

The main contributions of our work can be summarized as follows.

First, we consider fairness as a performance indicator of appointment scheduling, going beyond traditional operational metrics such as patient waiting time and provider overtime. We measure fairness by quantifying the differences in patient waiting times for a given schedule. A schedule is more fair when the patient waiting times are closer. We consider two measures of fairness, at the individual and group levels. Individual unfairness captures the variability in waiting times between all patients, and group unfairness captures the variability in average waiting times between patient groups, where patient groups are defined based on their stratified show-up probability predictions. By systematically analyzing trade-offs among various measures, we uncover the inherent tensions

between operational efficiency and equitable access to care. Group unfairness can be mitigated simply by not relying on stratified show-up probability predictions when scheduling patients. However, it is unclear how individual unfairness and operational efficiency are dependent on the schedule. A key insight is that it is possible to simultaneously reduce overtime and individual unfairness. However, achieving shorter average waiting times often comes at the expense of increased individual unfairness.

Second, we challenge the necessity of using stratified show-up probability predictions to achieve operational improvements. Our findings demonstrate that with carefully designed scheduling strategies, it is possible to maintain competitive operational performance without exacerbating group unfairness. In particular, we find that scheduling policies that do not rely on stratified show-up probability predictions (by randomly sequencing patients) tend to achieve superior performance, that is, they efficiently balance operational and individual fairness objectives. This insight provides an alternative perspective for healthcare systems where predictive tools may be unavailable, resource-intensive, or ethically controversial.

Finally, our framework is designed with practical implementation in mind, addressing pragmatic considerations that are often overlooked in theoretical studies. We consider different scheduling levers, including appointment slot length design, overbooking strategies, and patient sequencing. Additionally, we evaluate system configurations across diverse scenarios, capturing the complexity of real-world healthcare settings. These scenarios include systems of varying sizes (small and large), differing levels of workload (moderately and heavily overloaded), variations in no-show probabilities and patient group compositions, and both deterministic and stochastic service times.

The remainder of the paper is organized as follows. In Section 2, we provide a brief review of the related literature. In Section 3, we provide details of the scheduling problem that we study, including scheduling policies and performance metrics that we considered. In Section 4, we simulate the system under different scheduling rules and parameter settings to study the trade-offs between different performance metrics. In Section 5, we present a fluid-based approximation of the system and present theoretical results that support the numerical findings of Section 4. Finally, in Section 6, we conclude, discuss modeling limitations, and propose future research directions. Technical proofs and additional numerical results are provided in the Online Appendix.

2. Literature

Our research builds on a comprehensive body of literature, which we categorize into three main areas: racial disparity in healthcare, patient no-show behaviors, and appointment scheduling.

First, we discuss the literature that highlights the pervasive inequities experienced by racial and ethnic minorities in healthcare, as well as the role of technology in either exacerbating or mitigating these disparities. Nelson (2002) and Hostetter and Klein (2018) show that ethnic and racial minorities in the United States are less likely than white individuals to receive preventive healthcare and are more likely to experience a lower overall quality of care, even when socioeconomic factors such as income, neighborhood location, comorbidities, and health insurance are considered. Moreover, modern technologies serve as potent mechanisms for perpetuating racial inequality. Benjamin (2016) argues that both unconscious and deliberate biases are embedded in technologies such as artificial intelligence. Gianfrancesco et al. (2018) find that biases and deficiencies in the data used by machine learning algorithms contribute to socioeconomic disparities in healthcare. Obermeyer et al. (2019) provide a striking example, demonstrating how predictive models based on future health costs disproportionately disadvantage sicker Black patients. To address this issue, they propose a race-unaware approach that focuses on predicting future health outcomes instead of costs, thereby reducing bias. In advocating for a fairer use of machine learning, Rajkomar et al. (2018) emphasize the importance of proactively designing systems to advance health equity. They argue that incorporating principles of distributive justice into model design, deployment, and evaluation is critical. Nevertheless, Murray et al. (2020) reveal that eliminating socioeconomic factors alone does not resolve racial disparities. They show that other features of a patient's previous history remain strongly correlated with race, contributing to persistent inequities.

In the specific context of appointment scheduling, an important factor is patient no-show. Racial disparities in no-show probabilities have been well documented. For example, Huang and Hanauer (2014) report that African Americans, who represent 5.3% of their dataset, exhibit the lowest attendance rates for general pediatric appointments. Similarly, Miller et al. (2015) investigate repeat no show cases and find that younger, black and low-income patients face significant barriers to accessing care. Similarly, Hamilton et al. (2002) confirm that patients from lower socioeconomic backgrounds are less likely to keep their appointments, while Dantas et al. (2018) observe that minority groups are consistently linked to higher rates of missed appointments. Kaplan-Lewis and Percac-Lima (2013) demonstrate that black and Hispanic patients in underserved populations experience disproportionately high no-show rates. In the broader literature on resource allocation, it has also been established that prioritizing fairness often comes at the expense of other performance metrics (Bertsimas et al. 2011). In this paper, we critically reassess the need to utilize stratified predictive show-up probabilities to achieve improved operational efficiency. As highlighted above, such

predictive models can, whether intentionally or unintentionally, correlate with patients' racial and socioeconomic backgrounds, raising significant concerns about potential inequities in healthcare delivery.

A substantial body of research explores scheduling strategies that account for no-shows (e.g., Li et al. 2019, Samorani and LaGanga 2015, Zacharias and Pinedo 2014, Feldman et al. 2014, Hassin and Mendel 2008). See also Cayirli and Veral (2003) and Gupta and Denton (2008) for comprehensive surveys. Two common strategies in the literature to mitigate the adverse effects of no-shows are overbooking and adjusting the length of appointment slots. Using single-server queueing models, Liu and Ziya (2014) examine the optimal level of overbooking. LaGanga and Lawrence (2012) provide numerical evidence supporting the effectiveness of overbooking in diverse service environments and cost structures, and identify a well-performing front-loaded overbooking pattern. Zacharias and Pinedo (2014) observe a similar pattern, particularly in scenarios where overtime costs dominate. Armony et al. (2019) study both the slot length design and the overbooking strategy. They show that overbooking at the end of the session asymptotically minimizes the combined costs of customer waiting time and provider overtime in a large-population, overloaded limit. Kong et al. (2013) formulate the scheduling problem as a convex conic optimization problem with a tractable semidefinite relaxation. They find that when overbooking is necessary, assigning these bookings at the beginning or end of the session is optimal. For appointment slot length design, Wang (1993) presents numerical results that indicate that optimal appointment intervals for homogeneous patients exhibit a dome-shaped pattern. This pattern features appointment intervals that gradually increase toward the middle of the session and then decrease, assuming that service times are independent, identically distributed, and exponential. Supporting evidence for this dome-shaped pattern is also provided by Denton and Gupta (2003) and Robinson and Chen (2003). Klassen and Yoogalingam (2009) extend this understanding by numerically demonstrating that, with integer-valued appointment slots, the optimal scheduling pattern exhibits a plateau-dome structure.

When utilizing stratified or personalized patient information (e.g., show-up probabilities, service times), there are different sequencing rules developed to optimize operational performance (Denton et al. 2007). Wang (1999) studies a system in which patient service times follow exponential distributions with varying rates and finds that sequencing patients in descending order of service rates is optimal. Mak et al. (2015) investigate sequencing when the mean and variance of service times are known. Their results demonstrate that, under certain conditions, following a smallest-variance-first rule yields optimal outcomes. Due to the inherent complexity of these problems,

relatively few studies look at multiple aspects of appointment scheduling: slot length design, overbooking, and sequencing. Most existing research in this domain, including (Weiss 1990, Denton et al. 2007), focuses on systems with only two patients, primarily to address the significant analytical challenges involved. In contrast, our study considers all three dimensions of appointment scheduling simultaneously. To maintain practical feasibility and reduce the computational search space, we focus on policies that are easy to implement, e.g., policies with equal slot lengths and where overbooking is restricted to the first and last slots.

Most of the appointment scheduling literature discussed above focuses on minimizing operational performance costs, such as a weighted sum of patient waiting times, provider overtime, and idle time. Very few studies focus on fairness metrics. A common observation is that, according to current scheduling practice, patients scheduled later in a session tend to experience longer in-clinic delays (Cayirli and Veral 2003, Kong et al. 2020, Qi 2017). Some measures have been proposed to mitigate this unfairness. The paper closest to ours is the recent study by Samorani et al. (2022), which is among the first to examine racial disparities when utilizing stratified show-up probabilities for scheduling. They focus on a specific class of scheduling policies previously studied by Zacharias and Pinedo (2014). Their findings reveal that these policies can result in substantial unfairness between racial groups, prompting the authors to propose incorporating racial disparity into the objective function when optimizing scheduling policies. In contrast, our study explores a broader class of scheduling policies and demonstrates that, with careful optimization of slot length and overbooking strategies, it is possible to achieve strong operational performance metrics without compromising fairness. In addition, we investigate the trade-offs between different operational performance metrics and fairness measures. For instance, we show that while there is often a trade-off between average waiting times and individual fairness, improvements in overtime and individual fairness tend to align.

3. Modeling Framework

In this section, we introduce our modeling framework and scheduling policies. We focus on policies that are practically relevant. Our analysis evaluates performance not only through standard operational metrics, such as waiting time and overtime, but also by examining fairness, both at the individual level and across different patient groups.

3.1. Problem Formulation

We consider an outpatient clinic in which patients may or may not attend their scheduled appointments. The scheduling period, which spans a total length T, must accommodate Λ patients, indexed

by $i = 1, 2, \ldots, \Lambda$. A scheduling policy specifies the start time for the appointment of each patient. Let Δ_i denote the scheduled interarrival time between the $(i-1)$th and ith patients, with the assumption that $\Delta_i \geq 0$. In particular, the indexing of patients reflects the sequencing decision.

Appointment scheduling generally involves a two-stage planning process (Patrick and Aubin 2013, Demeulemeester et al. 2013, Mak et al. 2015). The first stage determines the number of patients to assign to a fixed scheduling period. The second stage allocates various time slots to individual patients. In this work, we explore varying levels of patient demand for a fixed scheduling period and, for each system configuration, investigate how to allocate time slots to patients.

We assume that the patient population consists of two groups, L and H, which are heterogeneous in their show-up probabilities, denoted as p_L and p_H. We assume $p_L < p_H$ so that L patients are less likely than H patients to attend their scheduled appointments. Let Λ_L, Λ_H be the numbers of L and H patients. Then, $\Lambda = \Lambda_L + \Lambda_H$. Let $\gamma = \Lambda_L/\Lambda$ be the fraction of L patients. We also define $p = \gamma p_L + (1-\gamma) p_H$ to be the average show-up probability of the patient population.

Let $\{I_i : 1 \leq i \leq \Lambda\}$ be a sequence of independent Bernoulli random variables with mean $p_i < 1$, indicating whether patient i shows up or not. When a patient shows up, we assume that they are punctual, i.e., they show up exactly at the scheduled appointment time. The provider serves patients in ascending order of their indices, following a first-come, first-served (FCFS) policy. The service times, denoted by $\{S_i : 1 \leq i \leq \Lambda\}$, are independently and identically distributed with mean 1, and are mutually independent of the show-up indicators, I_i's. No service preemptions are allowed.

3.2. Preliminary Analysis

We denote by W_i the waiting time of patient i. The waiting time is the duration between the patient's arrival and the start of their service. The sequence $\{W_i : i \geq 0\}$ forms a Lindley process, with the initial conditions $W_0 = 0$, $S_0 = 0$, $\Delta_1 = 0$. For $i \geq 1$,

$$W_i = (W_{i-1} + S_{i-1} I_{i-1} - \Delta_i)^+. \tag{1}$$

However, $\{W_i : i \geq 0\}$ is not a typical Lindley process since the $(S_{i-1} I_{i-1} - \Delta_i)$'s are not identically distributed. The indicators I_{i-1} are determined by patient sequencing, for example, whether H patients are scheduled before L patients, and Δ_i's are determined by the scheduling policy. For example, Δ_i can equal zero if the appointment slot is overbooked. Thus, analyzing the Lindley process in (1), which is needed to characterize performance or derive the optimal scheduling policy, can be prohibitively complex. Moreover, with heterogeneous patients, optimizing the appointment

sequence is crucial but often presents significant analytical challenges (Kong et al. 2016). This complexity is further exacerbated by the interdependence between the slot length design and the patient sequence, as determining the optimal slot length often requires extensive computations (Denton and Gupta 2003, Gupta 2007). Although a handful of studies have addressed these challenges analytically, primarily leveraging robust optimization (e.g., Mak et al. 2015), there has been no analytical work that simultaneously considers slot length design (determining the start times of appointments), sequencing (deciding on the order of heterogeneous patients), and overbooking (deciding on the number of patients per slot), as we do in this paper. Next, we present preliminary results on how these decisions affect patient waiting times.

PROPOSITION 1. *For constant service time or exponential service time with mean 1, for each i, there exists* $\alpha_i > 0$*, such that if* $\Delta_i \geq p_{i-1} + \alpha_i$*, then* $\mathbb{E}[W_i] \leq \mathbb{E}[W_{i-1}]$*; otherwise,* $\mathbb{E}[W_i] > \mathbb{E}[W_{i-1}]$.

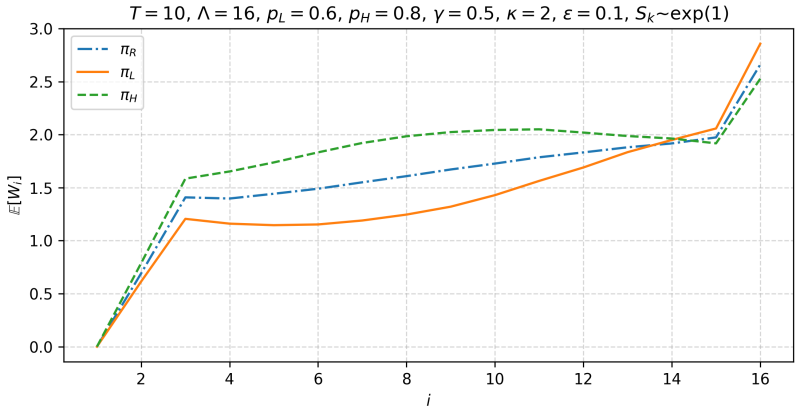
As a special case of Proposition 1, we note that if $\Delta_i = p_{i-1}$, then $\mathbb{E}[W_i] \geq \mathbb{E}[W_{i-1}]$. This suggests that setting the slot length equal to $\mathbb{E}[S_{i-1}I_{i-1}] = \mathbb{E}[S_{i-1}]\,\mathbb{E}[I_{i-1}] = p_{i-1}$ results in patients scheduled later during the scheduling period experiencing longer delays.

In Figure 1, we present a numerical example of expected patient waiting times under different sequencing strategies: scheduling H patients first (π_H), scheduling L patients first (π_L), and random sequencing (π_R). We fix $\Delta_i = p + 0.1$, with 3 patients scheduled in the first slot, 2 patients scheduled in the last slot, and exactly 1 patient scheduled in each of the remaining slots. Depending on the sequencing rule, overbooked patients can be H or L patients, or randomly chosen between the two. Figure 1 illustrates how the specific sequencing rule impacts waiting times. For example, the blue dash-dot curve, corresponding to random sequencing (in this case, $\alpha_i > 0.1$ and $p_{i-1} = p$, that is, $\Delta_i = p + 0.1 < p_{i-1} + \alpha_i$, for $i = 4, \ldots 15$), shows a steady increase in waiting times as the scheduling position increases. In contrast, the solid orange curve, corresponding to scheduling L patients first and the green dashed curve, corresponding to scheduling H patients first, exhibit some downward segments (since for those patients, $\Delta_i = p + 0.1 > p_L + \alpha_i$). Furthermore, overtime, which is equal to the sum of the expected waiting time of the last patient in this example and p_Λ, varies between the sequencing rules, further illustrating the impact of the sequencing rule on overall system performance.

3.3. Scheduling policies

We study a family of scheduling policies designed to efficiently balance provider overtime, patient waiting times, and fairness. We focus on policies that are practical, i.e., easy to implement in real-world settings. The scheduling policies are characterized by three key decisions: (1) *Appointment*

Figure 1 **Average patient wait times for different indices under different sequencing rules.**



slot length: we focus on slots of equal duration to simplify implementation. (2) *Patient allocation per slot:* overbooking is allowed only at the beginning or end of the schedule. (3) *Patient sequence:* patients can be sequenced based on their show-up probabilities. Next, we provide more details about each of these three key decisions.

3.3.1. Slot sizing and overbooking. Many studies have shown that optimal scheduling policies, under various performance objectives, often exhibit a "dome-shaped" pattern in terms of the slot length distribution (Denton and Gupta 2003, Robinson and Chen 2003, Wang 1993, Klassen and Yoogalingam 2009). In this approach, patients scheduled in the middle of the period are assigned longer slots, while those at the beginning or end are assigned shorter slots. Despite their theoretical appeal, dome-shaped scheduling policies are rarely adopted in practice (Kuiper et al. 2021). This is attributed to the analytical complexity of determining optimal slot lengths and the practical challenges of implementing schedules with varying slot durations.

In this paper, we consider a simplified and more practical class of dome-shaped scheduling policies. All slot lengths within the scheduling period, except the last, are set identically to $(p+\varepsilon)$, for some constant $\varepsilon \geq 0$. The last slot is set equal to $(T \mod (p+\varepsilon))$. Overbooking - assigning multiple patients to the same time slot - is allowed only in the first and last slots, i.e., at times 0 and $T-(T \mod (p+\varepsilon))$. Let $\kappa \geq 0$ denote the number of patients who are overbooked at time 0, i.e., the total number of patients scheduled in the first slot is $1+\kappa$. When all slots are fully occupied, any remaining patients are scheduled in the last slot. Under the assumption that patients who do show up are punctual, the interarrival times follow a plateau dome pattern (Klassen and Yoogalingam 2009). In particular, the interarrival times are zero for patients who are overbooked in the first and last slots. That is, if $\kappa > 0$, $\Delta_i = 0$ for $i = 1, \ldots, 1+\kappa$; if $\Lambda > \lfloor T/(p+\varepsilon) \rfloor + \kappa + 1$, $\Delta_i = 0$ for

$i = \lfloor T/(p+\varepsilon) \rfloor + \kappa + 2, \ldots, \Lambda$. For the remaining patients, the interarrival times are equal to $p+\varepsilon$, i.e., $\Delta_i = p+\varepsilon$ for $2+\kappa \le i \le \lfloor T/(p+\varepsilon) \rfloor + \kappa + 1$.

This class of plateau-dome policies is flexible and includes, as special cases, two scheduling rules that have been shown to be optimal under specific conditions. The first scheduling rule is the "L-shaped" front-loading schedule, which sets the slot length to $p+\varepsilon = \mathbb{E}[S_i] = 1$, i.e., $\varepsilon = 1-p$, and only allows overbooking at time 0. With homogeneous patients, when the slot lengths are fixed at 1, front loading has been shown to be optimal when the overtime cost dominates the cost of patient waiting (LaGanga and Lawrence 2012, Zacharias and Pinedo 2014). The second scheduling rule sets the slot length to p, i.e., $\varepsilon = 0$, and only allows overbooking in the last slot. Armony et al. (2019) shows that this scheduling policy asymptotically minimizes the sum of customer waiting times and provider overtime costs in a large population, overloaded limit.

In numerical experiments, we relax the restriction of overbooking to the beginning and end of the scheduling period, i.e., overbooking is allowed at any intermediate slot. Our results show that restricting the scheduling design to this special class of plateau-dome policies leads to only minimal performance loss. Specifically, the suboptimality gaps are less than 5% in most scenarios when considering weighted sum optimization problems and constrained optimization problems across different performance metrics (see online Appendix D for details). By restricting the design space to this special class of plateau-dome policies, near-optimal solutions can be achieved with significantly reduced computational complexity, which highlights the practicality of our framework.

3.3.2. Patient sequencing. We consider patient sequencing rules that leverage stratified show-up probabilities. Zacharias and Pinedo (2014) demonstrate that, under a front-loading schedule, scheduling patients with low show-up probabilities first minimizes the expected total waiting time. Wang (1999) shows that the optimal patient sequence is in descending order of service rates, assuming exponentially distributed service times. Another stream of literature focuses on the smallest-variance-first sequencing rules (Mak et al. 2015, Kemper et al. 2014). In our setting, under the assumption that service times are independent and identically distributed, the smallest-variance-first sequence is determined by the show-up probabilities. For example, when the service times are constant and $p_H > p_L > 0.5$, the smallest-variance-first sequence would schedule H patients first. Smallest-variance-first sequencing is not always optimal. For example, when physician overtime is highly costly, the largest-variance-first sequence might have better performance (Qi 2017).

Motivated by these findings, we focus on the following three sequencing rules. (1) Random sequencing (π_R): Patients are scheduled in random order, ignoring stratified show-up probability

predictions. (2) Low-first sequencing (π_L): Patients with low show-up probabilities are scheduled before those with high show-up probabilities. (3) High-first sequencing (π_H): Patients with high show-up probabilities are scheduled before those with low show-up probabilities.

It is important to note that the use of stratified show-up probabilities in sequencing decisions can raise concerns about fairness. Specifically, prioritizing patients based on their likelihood of showing up could inadvertently disadvantage certain groups, particularly if show-up probabilities are correlated with socioeconomic factors, access to transportation, or other systemic barriers. This approach may result in patient groups experiencing longer wait times, exacerbating existing inequities. Next, we define our performance metrics, including both operational and fairness metrics.

3.4. Performance Metrics

We let $0/0 = 0$ in what follows. We also let $\mathbf{L}$ and $\mathbf{H}$ denote the sets of H patients and L patients.

- The *average waiting time over the entire population*, $\mathbb{E}\left[\overline{W}\right]$, where

$$\overline{W} = \frac{\sum_{i=1}^{\Lambda} W_i I_i}{\sum_{i=1}^{\Lambda} I_i}.$$

- The *average waiting times for groups* $\mathbf{L}$ *and* $\mathbf{H}$, $\mathbb{E}\left[\overline{W}_L\right]$ and $\mathbb{E}\left[\overline{W}_H\right]$, where

$$\overline{W}_L = \frac{\sum_{i\in\mathbf{L}} W_i I_i}{\sum_{i\in\mathbf{L}} I_i} \text{ and } \overline{W}_H = \frac{\sum_{i\in\mathbf{H}} W_i I_i}{\sum_{i\in\mathbf{H}} I_i}.$$

- The *group unfairness*, GF, which is defined as

$$GF = \frac{\left|\mathbb{E}[\overline{W}_L] - \mathbb{E}[\overline{W}_H]\right|}{\mathbb{E}[\overline{W}]}.$$

  The measure GF quantifies the disparity in waiting times between patient groups $\mathbf{L}$ and $\mathbf{H}$, relative to the average waiting time. A higher value of GF indicates greater unfairness.

- The *individual unfairness*, IF, which is defined as

$$IF = \frac{\mathbb{E}[\max_{k=1,\ldots,\Lambda} W_k I_k - \min_{k=1,\ldots,\Lambda} W_k I_k]}{\mathbb{E}[\overline{W}]} = \frac{\mathbb{E}[\max_{k=1,\ldots,\Lambda} W_k I_k]}{\mathbb{E}[\overline{W}]}.$$

  Note that $W_1 = 0$ because the first scheduled patient does not wait. Thus, $\min_{k=1,\ldots,\Lambda} W_k I_k = 0$. The measure IF quantifies the relative range of the waiting times, with a higher value indicating greater unfairness. This is a commonly used unfairness metric in the literature (Xinying Chen and Hooker 2023, Cowell 2011, Wagstaff et al. 1991).

- The *provider overtime*, $\mathbb{E}[V]$, where

$$V = \left(\sum_{k=1}^{\Lambda} \Delta_k + W_\Lambda + S_\Lambda I_\Lambda - T \right)^+ .$$

  Note that $\sum_{k=1}^{\Lambda} \Delta_k$ is the time when patient Λ is scheduled to show up. The variable V quantifies the additional time that the provider has to work beyond T.

Among the performance metrics listed above, $\mathbb{E}[\overline{W}]$ and $\mathbb{E}[V]$ are commonly studied in the appointment scheduling literature and often serve as key criteria to evaluate scheduling efficiency and operational performance (Begen et al. 2012, Feldman et al. 2014). These metrics address critical stakeholder conflicts: patient wait time is widely used to evaluate patient experience and satisfaction, whereas overtime reflects the provider's working experience and system utilization. In contrast, GF and IF, which measure unfairness, have been discussed in Qi (2017), Turkcan et al. (2011), but remain relatively understudied. When a scheduling policy consistently favors one group of patients over another, the less favored group may experience longer waiting times, reduced access to timely care, and lower overall satisfaction. This issue becomes especially pressing when disparities align with larger societal inequalities, such as socioeconomic status or clinical needs, making it a topic of increasing importance in the pursuit of equitable healthcare delivery (Samorani et al. 2022, Mackenbach et al. 2008). Group fairness quantifies inequities between patient groups, while individual unfairness focuses on disparities at the level of individual patients, emphasizing the need to minimize inequities in waiting times and access to care between individuals (Wagstaff et al. 1991, Schlotheuber and Hosseinpoor 2022).

4. Simulation Study

In this section, we describe the results of a numerical study exploring different trade-offs in appointment scheduling. In Section 4.1, we explore the trade-offs between different performance measures. In Section 4.2, we explore the optimal scheduling policy under different objectives.

Each set of performance measures is estimated by averaging over 10^4 replications. We explore various system configurations, including different lengths of the scheduling period ($T = 10$ and $T = 30$) and different panel sizes ($\Lambda = 1.2T/p, 1.5T/p$, and $1.8T/p$). We focus on overloaded systems where effective scheduling policies are especially important. We set $(p_L, p_H) = (0.6, 0.8)$ following the estimates from Samorani et al. (2022). We also consider $(p_L, p_H) = (0.3, 0.7)$ and $(p_L, p_H) = (0.2, 0.3)$. We let $\gamma = 0.25, 0.5, 0.75$. We consider both exponentially distributed and deterministic service times. In total, we experimented with 108 different system configurations.

In terms of scheduling policies, we consider different slot lengths ($\varepsilon = 0, 0.1, ..., 1-p$), different numbers of overbooked patients in the first slot ($\kappa = 0, 1, \ldots, \lfloor \Lambda - T/(p+\varepsilon) \rfloor$), and different sequencing rules (π_R, π_L, and π_H). We define $\kappa_{\max}(\varepsilon) = \lfloor \Lambda - T/(p+\varepsilon) \rfloor$ as the maximum level of overbooking possible and note that $\kappa_{max}(\varepsilon) - \kappa$ patients are overbooked in the last slot. In total, our study evaluates 99,780 unique systems, ensuring that the findings are not restricted to a narrow set of assumptions and demonstrating the robustness of our results under various conditions. By systematically varying these parameters, we provide a characterization of fairness and operational trade-offs across a broad spectrum of operational scenarios, including relatively small and large systems, moderately and heavily overloaded systems, high and low no-show probabilities, different partitions of patient groups, deterministic and random service times, under various scheduling policies.

4.1. Trade-offs Among Different Performance Metrics

We begin by studying performance under different sequencing rules. We use a superscript to mark the dependence of a performance measure on the sequencing rule at hand, e.g., $\mathbb{E}[\overline{W}^{\pi_R}]$, $\mathbb{E}[V^{\pi_R}]$, IF^{π_R}, and GF^{π_R} correspond to π_R. Although π_R, which does not take advantage of the stratified show-up probability, achieves optimal group fairness, it is not clear how much improvement in operational performance can be gained by using this stratified show-up probability.

We explore this in Figure 2, where we compare the four performance metrics, $\mathbb{E}[\overline{W}]$, $\mathbb{E}[V]$, IF, and GF, under the three sequencing rules, π_R, π_H, and π_H. In this figure, we use exponentially distributed service times, and set $T = 10$ and $\Lambda = 17$. We fix $\varepsilon = 0.2$ while varying κ from 0 to 5 on the horizontal axis. We observe from the figure that for fixed ε and κ, compared to π_R, π_H tends to improve overtime and individual fairness, while π_L tends to improve the waiting time. We make similar observations for other system configurations (see Appendix A for additional numerical results). We formally summarize our observations as follows:

Observation 1. *For fixed κ and ε, the overtime follows the ranking*

$$\mathbb{E}\left[V^{\pi_H}\right] \leq \mathbb{E}\left[V^{\pi_R}\right] \leq \mathbb{E}\left[V^{\pi_L}\right].$$

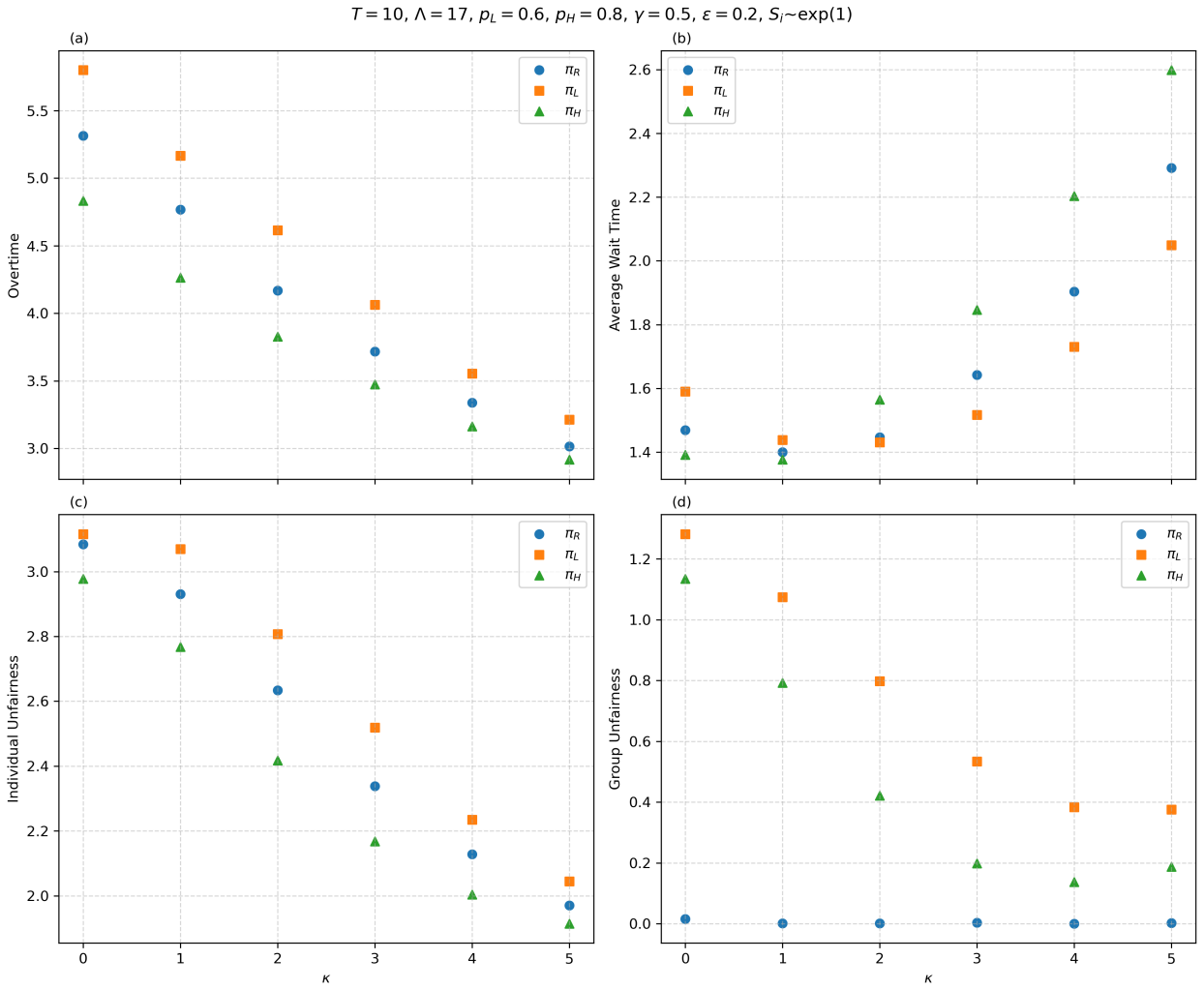
For fixed ε, there exists κ_w such that for all $\kappa \geq \kappa_w$, the waiting time averages are ranked as

$$\mathbb{E}\left[\overline{W}^{\pi_L}\right] \leq \mathbb{E}\left[\overline{W}^{\pi_R}\right] \leq \mathbb{E}\left[\overline{W}^{\pi_H}\right].$$

Additionally, for fixed ε, there exist thresholds $\underline{\kappa}_f$ and $\overline{\kappa}_f$ such that for $\underline{\kappa}_f \leq \kappa \leq \overline{\kappa}_f$, the individual unfairness are ranked as

$$IF^{\pi_H} \leq IF^{\pi_R} \leq IF^{\pi_L}.$$

Figure 2 **System performance across all three sequencing rules for varying κ values.**



For example, when $\varepsilon = 0, 0.1$, $\kappa_w = 1$; when $\varepsilon = 0.2$, $\kappa_w = 2$; and when $\varepsilon = 0.3$, $\kappa_w = 3$. Also, when $\varepsilon = 0, 0.1, 0.2$, $\underline{\kappa}_f = 0$ and $\overline{\kappa}_f = \kappa_{\max}(\varepsilon)$; when $\varepsilon = 0.3$, $\underline{\kappa}_f = 1$ and $\overline{\kappa}_f = \kappa_{\max}(0.3)$.

The intuition behind Observation 1 is as follows. Under π_H, high show-up probability patients are scheduled first. In this case, the backlog caused by overbooked patients at the beginning of the scheduling period clears slowly, leading to a greater propagation of delays throughout the scheduling period. Thus, this approach reduces idle time during the scheduling period, which in turn reduces overtime. In addition, π_H reduces individual unfairness by balancing patient delays throughout the scheduling period. In contrast, under π_L, patients with low show-up probability are scheduled first. Although this approach increases overtime due to more idling when early patients do not show up, it also helps reduce overall waiting times by clearing the initial backlog more quickly.

The contrasting trends between overtime and waiting time underscore a fundamental trade-off in appointment scheduling. Overtime primarily concerns the provider, as it affects both the working experience and resource management. In contrast, waiting time is critical from the patient's point of view, as prolonged waits can negatively impact the overall patient experience and quality of care. Balancing these two objectives –minimizing overtime for the provider while ensuring reasonable patient wait times – poses a significant challenge, since scheduling policies tend to prioritize one objective at the expense of the other.

Having compared the sequencing rules, we now focus on performance under π_R, and study the effects of other parameters of the scheduling policy. In Figure 3, we plot individual fairness, IF, as a function of the expected overtime, $\mathbb{E}[V]$, and the expected average waiting time, $\mathbb{E}[\overline{W}]$. We vary κ and ε. In this figure, we set $T = 10$, $\Lambda = 21$, and $S_i = 1$. Each labeled dot (e.g., 'Rx_y') in the figure represents a specific scheduling policy, where 'R' indicates the π_R sequencing rule, $\kappa = x$, and $\varepsilon = y/100$, e.g., "R2_20" corresponds to the π_R policy with $\kappa = 2$ and $\varepsilon = 0.2$. Recall that $\kappa_{max}(\varepsilon) - \kappa$ patients are overbooked in the last slot, so increasing κ amounts to decreased overbooking in the last slot. In what follows, we also use the superscript to mark the dependence of the performance metrics on the scheduling policy, e.g., $V^{R_{\varepsilon,\kappa}}$ and $\overline{W}^{R_{\varepsilon,\kappa}}$.

As κ increases, scheduling becomes more front-loaded, with less overbooking towards the end of the session. We observe from the figure that this shift typically results in greater congestion early in the session, leading to increased average waiting times but reduced overtime. Interestingly, individual unfairness generally trends in the same direction as overtime, but in the opposite direction of the average waiting time. We make similar observations in other system configurations (see Appendix A for additional numerical results), and formally summarize our observations as follows:

Observation 2. *For fixed $\varepsilon \geq 0$, there exists κ_0, such that for $\kappa_2 \geq \kappa_1 \geq \kappa_0$, the overtime, average waiting time, and individual fairness are ranked as:*

$$\mathbb{E}\left[V^{R_{\varepsilon,\kappa_1}}\right] \geq \mathbb{E}\left[V^{R_{\varepsilon,\kappa_2}}\right],$$

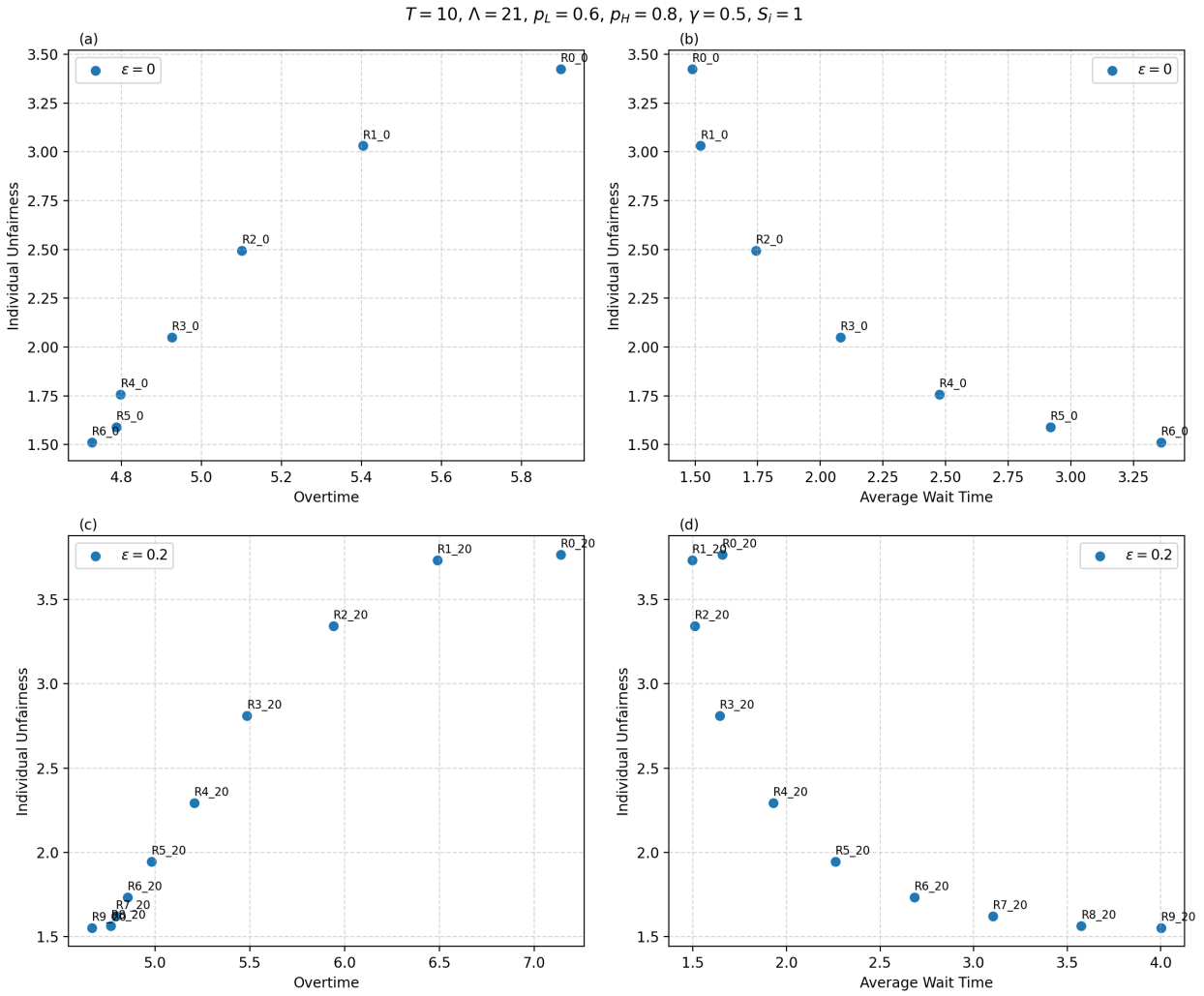
$$\mathbb{E}\left[\overline{W}^{R_{\varepsilon,\kappa_1}}\right] \leq \mathbb{E}\left[\overline{W}^{R_{\varepsilon,\kappa_2}}\right],$$

$$IF^{R_{\varepsilon,\kappa_1}} \geq IF^{R_{\varepsilon,\kappa_2}}.$$

In addition, individual unfairness is a convex function of the waiting time.

For example, when $\varepsilon = 0$, $\kappa_0 = 0$ as Figure 3(a)(b) show and when $\varepsilon = 0.2$, $\kappa_0 = 1$ as Figure 3(c)(d) show. Since individual unfairness is a convex function of the waiting time, indicating that in

Figure 3 Relationships between individual unfairness, overtime, and wait time, for π_R under different κ values.



systems with relatively short waiting times, achieving further reductions in waiting time may come at the expense of a substantial increase in individual unfairness.

The intuition behind Observation 2 is as follows. Patients typically experience the longest waiting times toward the end of the session. When overtime is high, e.g., as κ decreases so that there is more overbooking in the last slot, waiting times for patients served during the overtime period also increase. This results in greater variability in individual waiting times and, consequently, greater individual unfairness. On the other hand, as κ increases, the average waiting time can increase due to the substantial backlog earlier in the session.

One caveat to the relationship between individual unfairness and average waiting time is that both can increase as κ decreases when κ is very small, e.g., for $\kappa = 0$ in Figure 3(d). This occurs because, with minimal overbooking at the beginning of the scheduling period and, consequently,

more overbooking in the last slot, the waiting times of patients seen during overtime can be so large that they disproportionately affect the average waiting time.

Finally, we study individual fairness as a function of overtime and waiting time, for different scheduling policies. In Figure 4, we present one such example in which we also marked the corresponding Pareto frontiers. Based on this figure and the numerical results of many other system configurations (see Appendix A for additional numerical results), we make the following observation.

Observation 3. *Random sequencing can effectively balance individual unfairness with waiting time and overtime. In particular, there are points corresponding to π_R that lie on or near the efficiency frontier.*

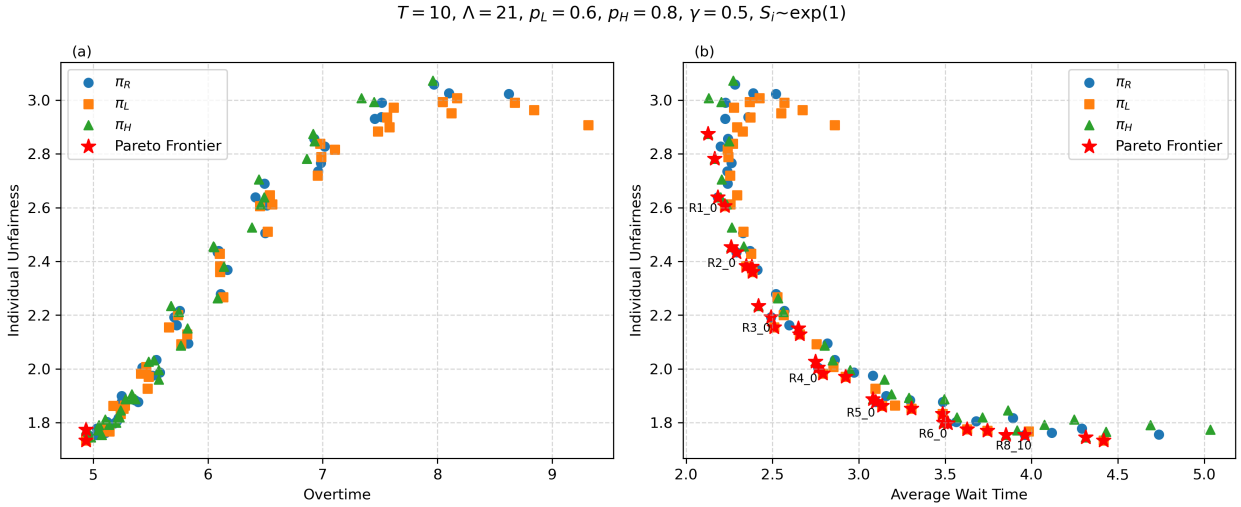
Consistent with Observation 2, the efficiency frontier demonstrates that reduced overtime is typically associated with decreased individual unfairness, while shorter average waiting times often result in increased individual unfairness. In other words, it is possible to simultaneously reduce overtime and individual unfairness. However, achieving shorter average waiting times often comes at the expense of increased individual unfairness. Thus, a provider can simultaneously minimize overtime and individual unfairness, but a trade-off must be made between fairness and average waiting time. This trade-off becomes particularly important when patient satisfaction is a key priority. If the primary goal is to minimize individual unfairness, that is, ensuring that no patient waits significantly longer than others, then patients may need to accept generally longer waiting times, on average. However, if the objective is to reduce the average waiting time, individual unfairness can increase, as some patients will experience disproportionately longer waits.

Observations 1, 2, and 3 provide valuable insights into navigating various trade-offs, and can be useful in designing scheduling policies that strike the desired balance between overtime, average waiting time, and individual fairness. Furthermore, it should be emphasized that individual unfairness, overtime, and average waiting times can be effectively balanced without sacrificing group fairness. We further illustrate this point through some case studies in the next subsection.

4.2. Case Studies

In this section, we present case studies to explore optimal scheduling rules in various scenarios. Since we have multiple objectives - minimizing waiting time, reducing overtime, and ensuring fairness - we consider two approaches: (a) assigning different weights to different performance metrics and optimizing the resulting weighted sum and (b) optimizing a single performance metric while

Figure 4 **Pareto frontier of trade-offs between overtime, wait time, and fairness across the different sequencing rules.**



treating the others as constraints. Through these analyses, our goal is to improve our understanding of the optimal policy in practical contexts, providing actionable strategies and managerial insights.

We begin with approach (a). For each of 108 system configurations $(T, \Lambda, p_L, p_H, \gamma, S_i)$, we fix the weight for the average waiting time at 1, i.e., $w_w = 1$, while varying the weights for the average overtime, w_0, individual fairness, w_{if}, and group fairness, w_{gf}. Specifically, we set $w_o \in \{0.1, 0.5, 1, 2, 10\}$ and $w_{if}, w_{gf} \in \{0, 2, 10\}$. The corresponding optimization problem is given by:

$$\Phi = \min \left(w_w \, \mathbb{E}\left[\overline{W}\right] + w_o \, \mathbb{E}[V] + w_{if} IF + w_{gf} GF \right). \tag{2}$$

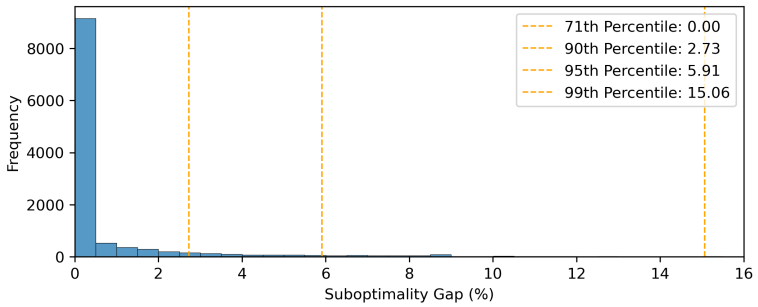
In approach (b), we minimize the average waiting time while imposing constraints on overtime and fairness. Specifically, we consider the following constrained optimization problem:

$$\begin{aligned} \min \quad & \mathbb{E}\left[\overline{W}\right], \\ \text{s.t.} \quad & \mathbb{E}[V] \le c_o, \quad \mathbb{E}[IF] \le c_{if}, \quad \text{and} \quad \mathbb{E}[GF] \le c_{gf}, \end{aligned} \tag{3}$$

where c_o, c_{if}, and c_{gf} are nonnegative parameters. We vary the values of these constants by setting them equal to numerical estimates of the 25th, 50th, and 75th percentiles, and the maximum of the corresponding performance metrics calculated across all simulated policies.

For the problems defined by (2) and (3) with different weights or constraints, and for various system configurations $(T, \Lambda, p_L, p_H, \gamma, S_i)$, we solve 11,772 problem scenarios in total. To search for the optimal scheduling policies for each problem scenario, we consider $\varepsilon = 0, 0.1, ..., 1 - p$ with the corresponding values of $\kappa \in \{0, \ldots, \kappa_{\max}(\varepsilon)\}$, and the three sequencing rules: π_R, π_L, and π_H.

Figure 5 Histogram of optimality gap for the best performing π_R policies across all scenarios.



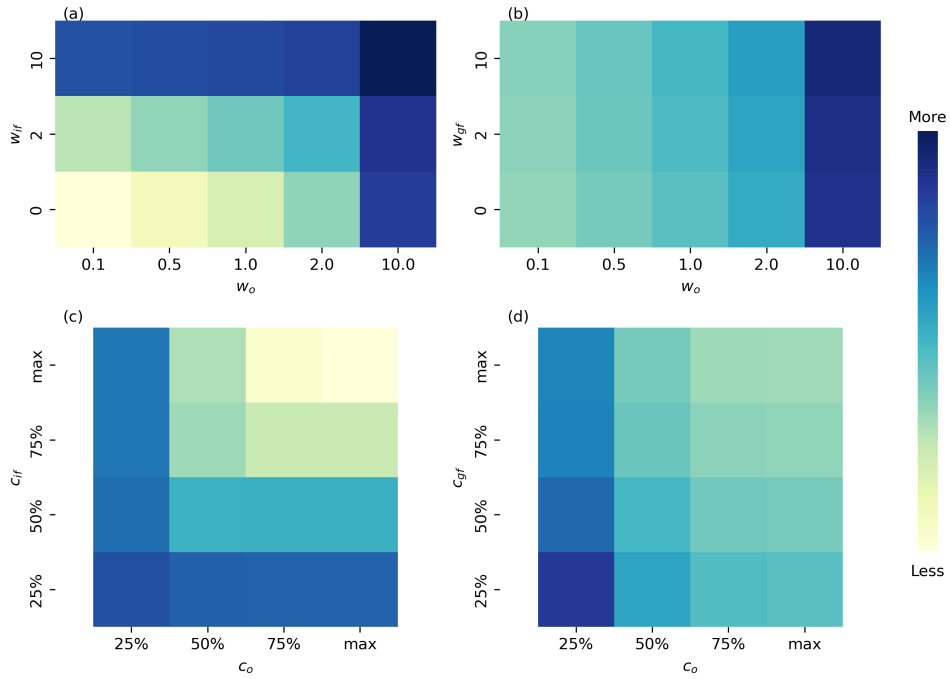
In each scenario, we find the top 10 best-performing policies, which all perform nearly optimally, with objective values differing only marginally. Within the set of top-performing policies for a given system configuration and problem formulation, the values of ε and the sequencing rule may vary, while the values of κ remain similar; see details in Appendix A.

A key observation is that random sequencing policies π_R consistently appear among the top performing policies. We also report the best performing π_R policies, along with the percentage optimality gap compared to the overall optimal policy (see some examples in Table 1 in Appendix A). In Figure 5, we plot the histogram of the distribution of the percentage optimality gaps for the best performing π_R policies across 11,772 problem scenarios. We see that π_R policies perform optimally in 71% scenarios and, in most cases, π_R policies are either optimal or near-optimal with a sub-optimality gap of less than 5%. Exceptions occur in scenarios where the weights are heavily skewed toward a single performance metric, favoring policies with π_H or π_L sequencing.

We can glean a key insight based on our numerical study: It is in general not necessary to rely on stratified show-up predictions to achieve optimal or near-optimal performance. By appropriately adjusting the scheduling structure design, specifically through κ (overbooking at time 0) and ε (slot length adjustment), one can effectively balance efficiency and individual fairness while leveraging only population-level information.

In addition, for every problem defined by equations (2) and (3), we report the average values of κ in all system configurations among the top 10 performing policies. The results are visualized in the heat map presented in Figure 6. The heat map reveals a clear trend: As the weights assigned to overtime and unfairness increase, or when constraints (imposed on overtime and unfairness) become more restrictive, the average value of κ increases. This suggests that when greater emphasis is placed on reducing overtime and unfairness, it becomes optimal to schedule more patients at the beginning of the session. This behavior aligns with the underlying trade-offs in balancing performance metrics discussed in Observation 2.

Figure 6 **Heat map of average κ of best performing policies across all scenarios.**



5. Fluid Analysis

To provide theoretical support to the findings of the numerical experiments, this section presents analyses based on a deterministic fluid approximation of the problem. Our objective is to deepen our understanding of the key trade-offs between different performance metrics.

The fluid approximation relies on two key relaxations to simplify the analysis. First, all random quantities are treated as deterministic, i.e., we replace the random service duration and the indicator of whether a patient shows up with their corresponding average values. Second, we assume that fluid processes are continuous in time; e.g., arriving patients are treated as continuous quanta of fluid that leave the system continuously (provided that there is remaining fluid). In Section 5.4, we illustrate that the fluid approximation is most accurate when T is large relative to the service time.

5.1. The Fluid Model

Consistent with the scheduling design assumed in Section 3, the fluid model restricts overbooking to times 0 and T. With a slight abuse of notation, we continue to use κ to denote the amount of overbooking at time 0 (which is equivalent to the amount booked in the fluid model). We define the patient scheduled arrival rate as $1/(p+\varepsilon)$, where we require $\varepsilon \in [0, 1-p]$. We assume that the system is overloaded and that, the overbooking at a specific time point (0 or T) belongs to the same patient group. We also require that the amount of overbooking at time 0 does not exceed the total amount that is required to be overbooked. We formalize these assumptions in Assumption 1.

ASSUMPTION 1. *In the fluid model, we assume:*

1. *Show-up probabilities:* $p_L < p_H < 1$.
2. *Patient population size:* $T < \Lambda_L p_L + \Lambda_H p$, $\Lambda_L < T$, $\Lambda_H < T$.
3. *Slot length and overbooking:* $0 \le \varepsilon \le 1 - p$. *For any fixed* $\varepsilon \ge 0$, $\kappa \le \Lambda - T/(p+\varepsilon)$.

We use lowercase letters to represent fluid performance measures. That is, for a sequencing rule $j \in \{\pi_R, \pi_L, \pi_H\}$, we let v^j, $\bar{w}^j$, if^j, and gf^j denote overtime, waiting time, individual and group unfairness, respectively. Let $x^j(t)$ denote the fluid level of patients in the system at time t.

For π_R, $x^{\pi_R}(t)$ at times 0 and T are given by $x^{\pi_R}(0) = \kappa p$ and $x^{\pi_R}(T) = x^{\pi_R}(T-) + (\Lambda - T/(p+\varepsilon) - \kappa)p$. For $t \in (0,T)$ and $t \in (T, T+\Lambda)$, $x^{\pi_R}(t) \ge 0$ evolves according to

$$\frac{dx^{\pi_R}(t)}{dt} = \left(\frac{p}{p+\varepsilon} - 1\right) \mathbb{1}\{t < T\} - \mathbb{1}\{t > T\} + l^{\pi_R}(t),$$

where $l^{\pi_R}(0) = 0$, $dl^{\pi_R}(t) \ge 0$, and $\int_0^{T+\Lambda} x^{\pi_R}(t) dl^{\pi_R}(t) = 0$. Note that the fluid continuously arrives at a constant rate $p/(p+\varepsilon)$ on $(0,T)$, and is served at rate 1 if the system is not empty. $(x^{\pi_R}(t), l^{\pi_R}(t))$ on $(0,T)$ and $(T, T+\Lambda)$ is a Skorohold problem.

For π_L, we have $x^{\pi_L}(0) = \kappa p_L$ and $x^{\pi_L}(T) = x^{\pi_L}(T-) + (\Lambda - T/(p+\varepsilon) - \kappa)p_H$ since κ fluid of type L is scheduled at time 0, and the remaining $(\Lambda - T/(p+\varepsilon) - \kappa)$ fluid of type H is scheduled at time T. For $t \in (0,T)$ and $t \in (T, T+\Lambda)$, $x^{\pi_L}(t) \ge 0$ evolves according to

$$\begin{aligned}\frac{dx^{\pi_L}(t)}{dt} = &\left(\frac{p_L}{p+\varepsilon} - 1\right) \mathbb{1}\{t \le (\Lambda_L - \kappa)(p+\varepsilon)\} \\ &+ \left(\frac{p_H}{p+\varepsilon} - 1\right) \mathbb{1}\{(\Lambda_L - \kappa)(p+\varepsilon) < t < T\} - \mathbb{1}\{t > T\} + l^{\pi_L}(t),\end{aligned}$$

where $l^{\pi_L}(0) = 0$, $dl^{\pi_L}(t) \ge 0$, and $\int_0^{T+\Lambda} x^{\pi_L}(t) dl^{\pi_L}(t) = 0$. Note that κ fluid of type L is overbooked at time 0 and the rest of type L shows up continuously at a rate $p_L/(p+\varepsilon)$ until time $(\Lambda_L - \kappa)(p+\varepsilon)$. Thereafter, fluid of type H is scheduled.

Similarly to π_L, for π_H we have $x^{\pi_H}(0) = \kappa p_H$ and $x^{\pi_H}(T) = x^{\pi_L}(T-) + (\Lambda - T/(p+\varepsilon) - \kappa)p_L$. For $t \in (0,T)$ and $t \in (T, T+\Lambda)$, $x^{\pi_H}(t) \ge 0$ and evolves according to

$$\begin{aligned}\frac{dx^{\pi_H}(t)}{dt} = &\left(\frac{p_H}{p+\varepsilon} - 1\right) \mathbb{1}\{t \le (\Lambda_H - \kappa)(p+\varepsilon)\} \\ &+ \left(\frac{p_L}{p+\varepsilon} - 1\right) \mathbb{1}\{(\Lambda_H - \kappa)(p+\varepsilon) < t < T\} - \mathbb{1}\{t > T\} + l^{\pi_H}(t),\end{aligned}$$

where $l^{\pi_H}(0) = 0$, $dl^{\pi_H}(t) \ge 0$, and $\int_0^{T+\Lambda} x^{\pi_H}(t) dl^{\pi_H}(t) = 0$.

We can derive closed-form characterizations of $x^j(t)$'s in a relatively straightforward manner. For example,

$$x^{\pi_R}(t) = \begin{cases} \left(\kappa p - \frac{\varepsilon}{p+\varepsilon} t\right)^+, \text{ if } 0 \le t < T, \\ \left(\left(\kappa p - \frac{\varepsilon}{p+\varepsilon} T\right)^+ + (\Lambda - T/(p+\varepsilon) - \kappa)p - (t-T)\right)^+, \text{ if } t \ge T. \end{cases} \tag{4}$$

In Appendix B, we provide the explicit expressions for x^{π_L} and x^{π_H}.

Given $x^j(t)$, we have the following expressions for the various fluid performance measures:

$$v^j = x^j(T), \quad \overline{w}^j = \frac{1}{\Lambda p} \int_0^{T+\Lambda} x^j(t)dt,$$

$$\text{if}^j = \frac{\max_{0 \le t \le T+\Lambda} x^j(t)}{\overline{w}^j}, \quad \text{and} \quad \text{gf}^j = \frac{\left|\overline{w}_L^j - \overline{w}_H^j\right|}{\overline{w}^j},$$

where

$$\overline{w}_L^{\pi_R} = \frac{1}{\Lambda_L p_L} \int_0^{T+\Lambda} \frac{\Lambda_L p_L}{\Lambda p} x^{\pi_R}(t)dt = \frac{1}{\Lambda p} \int_0^{T+\Lambda} x^{\pi_R}(t)dt = \overline{w}_H^{\pi_R},$$

$$\overline{w}_L^{\pi_L} = \frac{1}{\Lambda_L p_L} \int_0^{(\Lambda_L - \kappa)(p+\varepsilon)} x^{\pi_L}(t)dt, \quad \overline{w}_H^{\pi_L} = \frac{1}{\Lambda_H p_H} \int_{(\Lambda_L - \kappa)(p+\varepsilon)}^{T+\Lambda} x^{\pi_L}(t)dt,$$

and

$$\overline{w}_L^{\pi_H} = \frac{1}{\Lambda_L p_L} \int_{(\Lambda_H - \kappa)(p+\varepsilon)}^{T+\Lambda} x^{\pi_H}(t)dt, \quad \overline{w}_H^{\pi_H} = \frac{1}{\Lambda_H p_H} \int_0^{(\Lambda_H - \kappa)(p+\varepsilon)} x^{\pi_H}(t)dt.$$

5.2. Performance Comparison With and Without Stratified Show-up Probabilities

In this section, we present analytical results comparing the performance measures - overtime, waiting time, and individual unfairness - across different sequencing rules. Recall that $\text{gf}^{\pi_R} = 0$ is minimal since $\overline{w}_L^{\pi_R} = \overline{w}_H^{\pi_R}$. However, relying on stratified show-up probabilities, with π_L and π_H, may improve other performance metrics. Let $\bar{\kappa}_\varepsilon = \Lambda - T/(p+\varepsilon)$ be the maximum level of overbooking at time zero, for a given $\varepsilon \ge 0$, since we do not allow for idle slots before T. Because v^j depends on κ, with a slight abuse of notation, we write $v^j(\kappa)$ to treat it explicitly as a function of κ.

Proposition 2. *For the fluid scheduling problem, under Assumption 1,*

1. *For any fixed $\varepsilon \ge 0$ and $\kappa \in [0, \bar{\kappa}_\varepsilon]$, $v^{\pi_H} \le v^{\pi_R} \le v^{\pi_L}$.*
2. *For any fixed $\varepsilon \ge 0$, there exist $\kappa_1, \kappa_2 \in [0, \bar{\kappa}_\varepsilon)$, such that for any $\kappa \in [\kappa_1, \bar{\kappa}_\varepsilon]$, $\bar{w}^{\pi_R} < \bar{w}^{\pi_H}$, and for any $\kappa \in [\kappa_2, \bar{\kappa}_\varepsilon]$, $\bar{w}^{\pi_L} < \bar{w}^{\pi_R}$.*
3. *For any fixed $\varepsilon \ge 0$, define $\delta_H, \delta_L > 0$ such that $\max\{(\bar{\kappa}_\varepsilon - \delta_H)p, p\bar{\kappa}_{\varepsilon=0}\} = \delta_H p_H + [(p_H - p - \varepsilon)(\Lambda_H - \delta_H)]^+$ and $\max\{(\bar{\kappa}_\varepsilon - \delta_L)p_H, p\bar{\kappa}_{\varepsilon=0}\} = \delta_L p$. Then,*

(i) When $\kappa_1 < \delta_H$, there exist $0 \le \underline{\kappa}_1 < \overline{\kappa}_1 \le \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\underline{\kappa}_1, \overline{\kappa}_1]$, $if^{\pi_H} < if^{\pi_R}$.

(ii) When $\kappa_2 < \delta_L$, there exist $0 \le \underline{\kappa}_2 < \overline{\kappa}_2 \le \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\underline{\kappa}_2, \overline{\kappa}_2]$, $if^{\pi_R} < if^{\pi_L}$.

Proposition 2 supports Observation 1 of the simulation study presented in Section 4.1. First, scheduling H patients before L patients reduces overtime because more work is scheduled upfront. In fact, scheduling H patients first tends to minimize idle time, thus decreasing overtime. In a similar vein, scheduling L patients before H leads to longer overtime compared to random sequencing because less work is scheduled upfront.

Second, for κ sufficiently large, scheduling L patients before H patients reduces the average waiting time. This is because significant overbooking at the beginning of the schedule creates a backlog of work that π_L tends to mitigate more effectively. In contrast, scheduling H patients first increases the average waiting time, e.g., compared to random sequencing. However, when κ is small, the above relationships may not hold. For example, the numerical example in Figure 7(b) shows that when $\kappa \ge 2$, we have $\bar{w}^{\pi_L} < \bar{w}^{\pi_R} < \bar{w}^{\pi_H}$, but for $\kappa < 2$, the relationship is more complex.

Third, for moderate values of κ, scheduling H patients first reduces individual unfairness, whereas scheduling L patients first increases individual unfairness, compared to random sequencing. This is because scheduling L patients upfront is associated with reduced waiting time and increased idleness at the beginning of the schedule, and increased waiting time in later slots. In contrast, π_H tends to better balance the waiting times of patients overbooked at the beginning, $t = 0$, and at the end, $t = T$. However, for very small or large κ, these relationships may not hold. For example, the numerical example in Figure 7(c) shows that when $1 < \kappa < 6$, we have $\text{if}^{\pi_H} < \text{if}^{\pi_R} < \text{if}^{\pi_L}$, but the relationship becomes more complex when $\kappa < 1$ or $\kappa > 6$.

5.3. Trade-offs between Overtime, Wait Time, and Fairness

To shed more light on the trade-offs between different performance metrics for a given sequencing rule, we now focus on random sequencing.

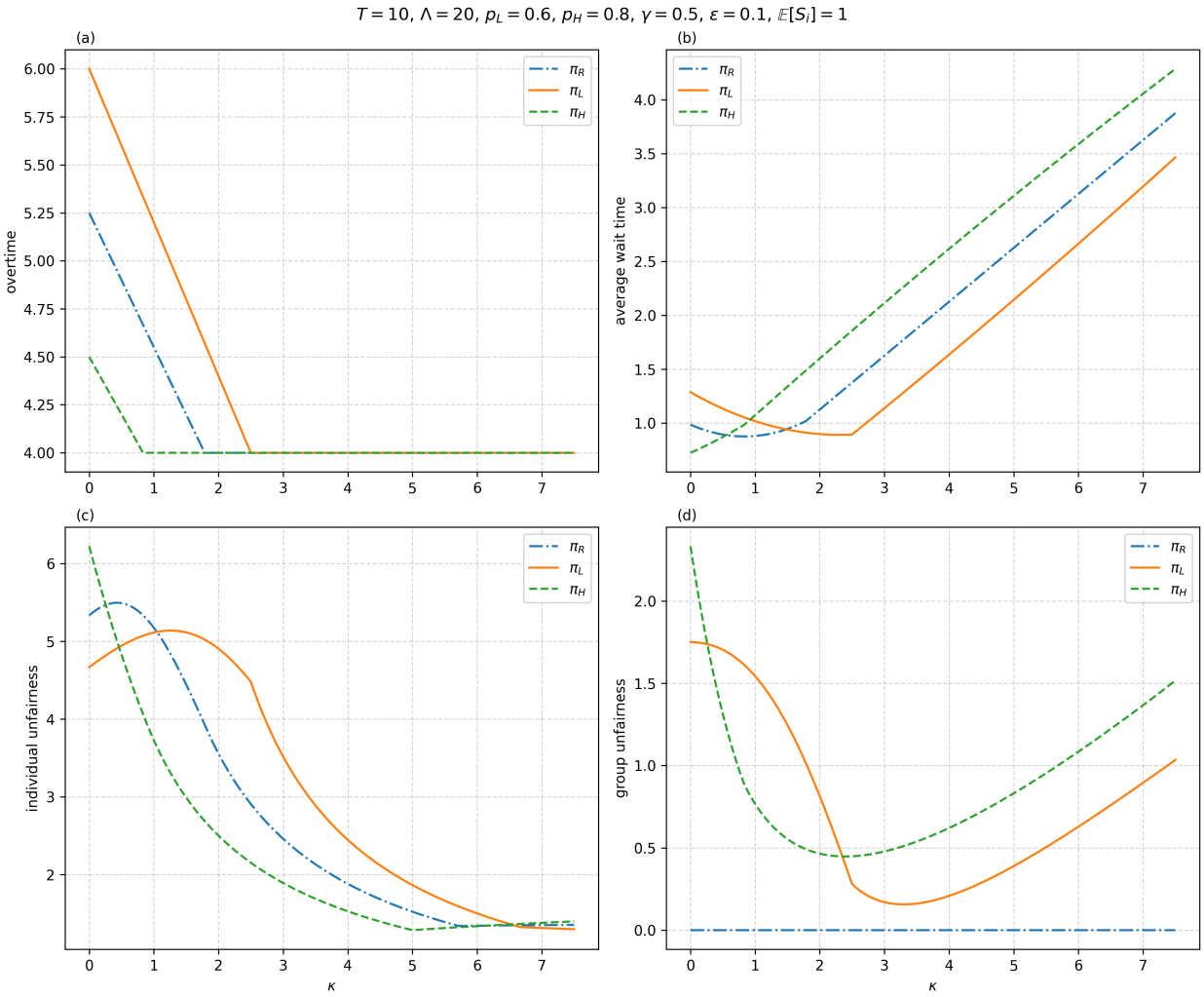
PROPOSITION 3. *For any ε, there exist $0 \le \tilde{\kappa}_l < \tilde{\kappa}_u \le \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\tilde{\kappa}_l, \tilde{\kappa}_u]$, we have*

$$\frac{dv^{\pi_R}}{d\kappa} \le 0, \quad \frac{d\bar{w}^{\pi_R}}{d\kappa} > 0, \quad \frac{dif^{\pi_R}}{d\kappa} < 0, \quad \text{and} \quad \frac{d^2 if^{\pi_R}}{d(\bar{w}^{\pi_R})^2} > 0.$$

For any fixed $\kappa > 0$, there exists $\tilde{\varepsilon} > 0$, such that for any $\varepsilon \in [0, \tilde{\varepsilon}]$, we have

$$\frac{dv^{\pi_R}}{d\varepsilon} \ge 0, \quad \frac{d\bar{w}^{\pi_R}}{d\varepsilon} < 0, \quad \frac{dif^{\pi_R}}{d\varepsilon} > 0, \quad \text{and} \quad \frac{d^2 if^{\pi_R}}{d(\bar{w}^{\pi_R})^2} > 0.$$

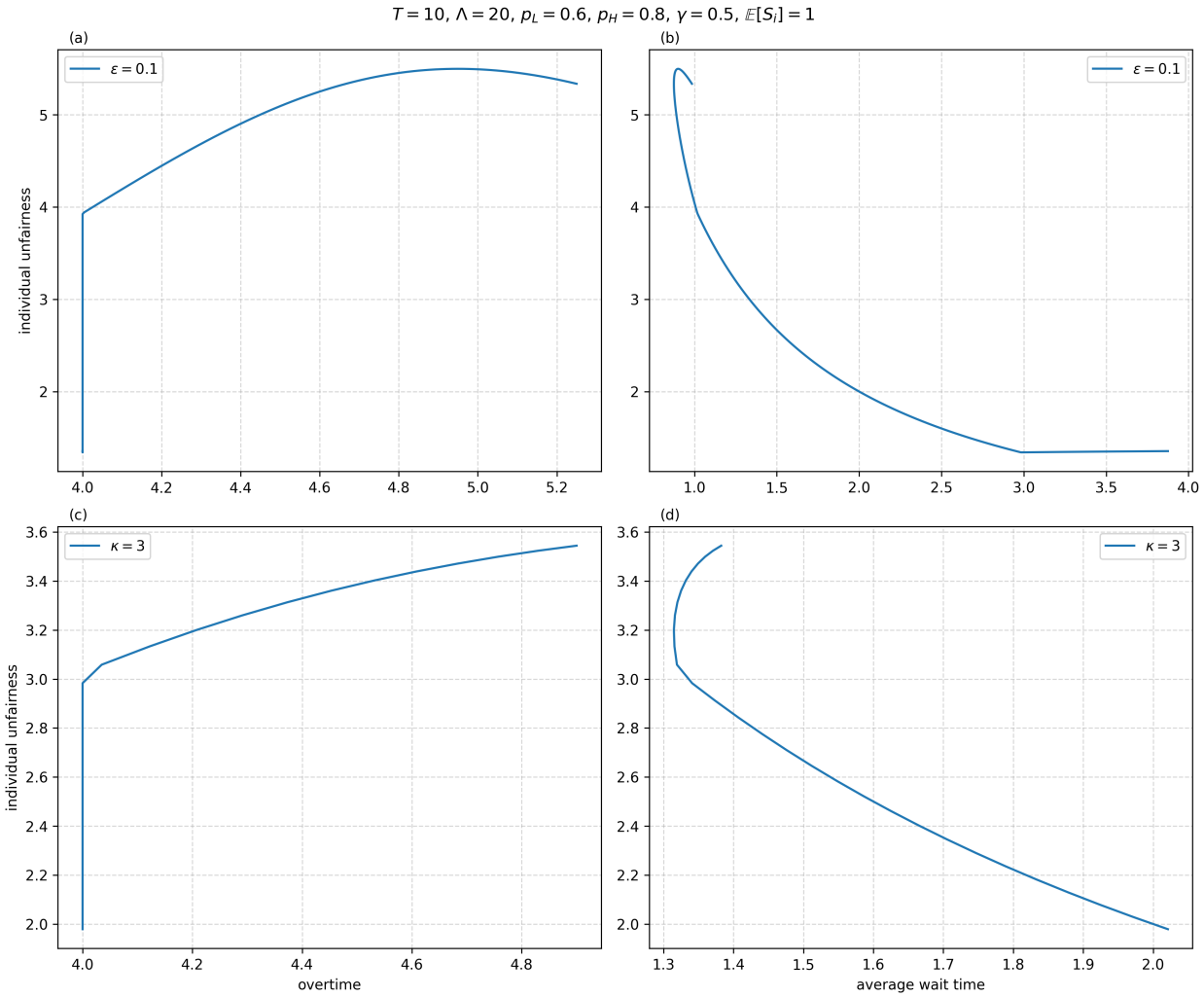
Figure 7 Fluid performance for all three sequencing rules.



Proposition 3 examines how κ and ε affect system performance. Specifically, for moderate κ and ε, increasing the amount of overbooking κ while keeping the slot length ε fixed, or decreasing the slot length through ε while keeping κ fixed, both result in reduced overtime, increased average waiting time, and reduced individual unfairness. In particular, overtime and individual unfairness trend in the same direction, while the average waiting time trends in the opposite direction. In addition, there is a convex relationship between individual unfairness and average waiting time. These findings align closely with Observation 2 in our simulation study.

In Figure 8, we present a numerical example quantifying the trade-offs between overtime, average waiting time, and individual unfairness. In Figure 8(a)(b), the slot length is fixed while the initial overbooking κ varies. In Figure 8(c)(d), the overbooking at time 0 is fixed while ε varies along the curve. In addition to confirming the findings of Proposition 3, these graphs also highlight that the

Figure 8 **Trade-offs between fluid performance metrics.**



ranges of κ and ε characterized in Proposition 3 correspond to scheduling policies on the efficiency frontier of the trade-off curves. In contrast, parameter values outside these ranges may represent strategies that deviate from the efficiency frontier, making them less practically relevant.

5.4. Connecting Stochastic and Fluid Models

In this section, we study the connection between the stochastic model and the fluid model numerically, analyzing the conditions under which these two models yield similar performance. As illustrated in Figure 13 in Appendix C, we observe that as the size of the system (T and Λ) increases, the fluid model becomes a more accurate approximation to the stochastic model. Additional numerical results supporting this observation are provided in Appendix C. Furthermore, while the stochastic model captures finer variations, the fluid model remains useful even in smaller systems. In particular, qualitative insights, such as trade-offs between different system performance

metrics, remain consistent in both models. For example, in small systems, Figures 2 and 7 reveal differences in magnitude but convey the same insights regarding the impact of utilizing stratified show-up predictions in stochastic and fluid models. The same applies to Figures 3 and 8, which show how individual unfairness is traded off with other metrics of operational performance.

6. Conclusion

This study examines the use of stratified show-up probability predictions in outpatient scheduling systems and evaluates their impact on both fairness and operational efficiency. We also explore the trade-offs between operational performance metrics and fairness metrics, offering insights into how these competing objectives can be balanced in appointment scheduling. Our results reveal that stratified show-up probability predictions, although widely considered essential to improve operational performance, may not be necessary. Furthermore, while it is possible to simultaneously reduce overtime and individual unfairness, reducing average waiting time often results in increased individual unfairness. This highlights a key trade-off in the design of scheduling policies.

We investigate a comprehensive set of scheduling decisions, including appointment slot length design, overbooking strategies, and patient sequencing, while ensuring that the policies remain practical and implementable. Our numerical study covers a wide range of system configurations in various scenarios, providing robust insights. To further support our findings, we complement numerical analysis with analytical results based on fluid-based approximations, enhancing both the depth and generalizability of our results.

Our study has several limitations that open interesting avenues for future research. First, we do not provide an analytical solution for the appointment scheduling problem, e.g., the case studies considered in Section 4.2, leaving an opportunity for future work to derive closed-form solutions or approximate analytical insights. Second, while our study focuses on stratified prediction, an extension to fully personalized, individual-level predictive information could offer new possibilities for optimizing scheduling strategies. Investigating how these more granular predictions impact both operational efficiency and fairness would be a valuable contribution. Finally, our work centers on offline scheduling; future research could examine online sequential scheduling systems that incorporate individual predictive no-show information in real-time decision making.

References

Agency for Healthcare Research and Quality (2024) National healthcare quality and disparities reports. Content last reviewed July 2024. Agency for Healthcare Research and Quality, Rockville, MD., available at:, URL `https://www.ahrq.gov/research/findings/nhqrdr/index.html`.

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

Alaeddini A, Yang K, Reddy C, Yu S (2011) A probabilistic model for predicting the probability of no-show in hospital appointments. *Health care management science* 14:146–157.

Armony M, Atar R, Honnappa H (2019) Asymptotically optimal appointment schedules. *Mathematics of Operations Research* 44(4):1345–1380, URL http://dx.doi.org/10.1287/moor.2018.0973.

Begen MA, Levi R, Queyranne M (2012) Technical note—a sampling-based approach to appointment scheduling. *Operations Research* 60(3):675–681, URL http://dx.doi.org/10.1287/opre.1120.1053.

Benjamin R (2016) Innovating inequity: If race is a technology, postracialism is the genius bar. *Ethnic and racial studies* 39(13):2227–2234.

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Operations research* 59(1):17–31.

Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. *Production and operations management* 12(4):519–549.

Cowell F (2011) *Measuring Inequality*. London School of Economics Perspectives in Economic Analysis (OUP Oxford), ISBN 9780191625121, URL https://books.google.com/books?id=0-V4wIGDxhIC.

Dantas LF, Fleck JL, Oliveira FLC, Hamacher S (2018) No-shows in appointment scheduling–a systematic literature review. *Health Policy* 122(4):412–421.

Demeulemeester E, Beliën J, Cardoen B, Samudra M (2013) Operating room planning and scheduling. *Handbook of Healthcare Operations Management: Methods and Applications* 121–152.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* 35(11):1003–1016, URL http://dx.doi.org/10.1080/07408170304395.

Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science* 10:13–24.

Feldman J, Liu N, Topaloglu H, Ziya S (2014) Appointment scheduling under patient preference and no-show behavior. *Operations Research* 62(4):794–811, URL http://dx.doi.org/10.1287/opre.2014.1286.

Forbes (2019) Missed appointments, missed opportunities: Tackling the patient no-show problem. URL https://www.forbes.com/sites/sachinjain/2019/10/06/missed-appointments-missed-opportunities-tackling-the-patient-no-show-problem/, accessed: 2025-01-07.

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* 178(11):1544–1547.

Gupta D (2007) Surgical suites' operations management. *Production and Operations Management* 16(6):689–700.

Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. *IIE transactions* 40(9):800–819.

Hamilton W, Round A, Sharp D (2002) Patient, hospital, and general practitioner characteristics associated with non-attendance: a cohort study. *British Journal of General Practice* 52(477):317–319.

Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management science* 54(3):565–572.

Hostetter M, Klein S (2018) In focus: Reducing racial disparities in health care by confronting racism. *Commonwealth Fund* 10.

Huang Y, Hanauer DA (2014) Patient no-show predictive model development using multiple data sources for an effective overbooking approach. *Applied clinical informatics* 5(03):836–860.

Kaplan-Lewis E, Percac-Lima S (2013) No-show to primary care appointments: why patients do not come. *Journal of primary care & community health* 4(4):251–255.

Kemper B, Klaassen CA, Mandjes M (2014) Optimized appointment scheduling. *European Journal of Operational Research* 239(1):243–255, ISSN 0377-2217, URL http://dx.doi.org/https://doi.org/10.1016/j.ejor.2014.05.027.

Klassen KJ, Yoogalingam R (2009) Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management* 18(4):447–458, URL http://dx.doi.org/https://doi.org/10.1111/j.1937-5956.2009.01021.x.

Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* 61(3):711–726, URL http://dx.doi.org/10.1287/opre.2013.1158.

Kong Q, Lee CY, Teo CP, Zheng Z (2016) Appointment sequencing: Why the smallest-variance-first rule may not be optimal. *European Journal of Operational Research* 255(3):809–821.

Kong Q, Li S, Liu N, Teo CP, Yan Z (2020) Appointment scheduling under time-dependent patient no-show behavior. *Management Science* 66(8):3480–3500.

Kuiper A, de Mast J, Mandjes M (2021) The problem of appointment scheduling in outpatient clinics: A multiple case study of clinical practice. *Omega* 98:102122.

LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and operations management* 21(5):874–888.

Li Y, Tang SY, Johnson J, Lubarsky DA (2019) Individualized no-show predictions: Effect on clinic overbooking and appointment reminders. *Production and Operations Management* 28(8):2068–2086.

Liu D, Shin WY, Sprecher E, Conroy K, Santiago O, Wachtel G, Santillana M (2022) Machine learning approaches to predicting no-shows in pediatric medical appointment. *NPJ digital medicine* 5(1):50.

Liu N, Ziya S (2014) Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management* 23(12):2209–2223.

Mackenbach JP, Stirbu I, Roskam AJR, Schaap MM, Menvielle G, Leinsalu M, Kunst AE (2008) Socioeconomic inequalities in health in 22 european countries. *New England Journal of Medicine* 358(23):2468–2481, URL http://dx.doi.org/10.1056/NEJMsa0707519.

Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Science* 61(2):316–334, URL http://dx.doi.org/10.1287/mnsc.2013.1881.

Miller AJ, Chae E, Peterson E, Ko AB (2015) Predictors of repeated "no-showing" to clinic appointments. *American journal of otolaryngology* 36(3):411–414.

Murray SG, Wachter RM, Cucina RJ (2020) Discrimination by artificial intelligence in a commercial electronic health record—a case study. *Health Affairs Forefront* .

Nelson A (2002) Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association* 94(8):666–668.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

Parsons J, Bryce C, Atherton H (2021) Which patients miss appointments with general practice and the reasons why: a systematic review. *British Journal of General Practice* 71(707):e406–e412.

Patrick J, Aubin A (2013) Models and methods for improving patient access. *Handbook of Healthcare Operations Management: Methods and Applications*, 403–420 (Springer).

Qi J (2017) Mitigating delays and unfairness in appointment systems. *Management Science* 63(2):566–583, URL http://dx.doi.org/10.1287/mnsc.2015.2353.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH (2018) Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169(12):866–872.

Ray KN, Chari AV, Engberg J, Bertolet M, Mehrotra A (2015) Disparities in time spent seeking medical care in the united states. *JAMA internal medicine* 175(12):1983–1986.

Robinson LW, Chen RR (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions* 35(3):295–307, URL http://dx.doi.org/10.1080/07408170304367.

Samorani M, Harris SL, Blount LG, Lu H, Santoro MA (2022) Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management* 24(6):2825–2842, URL http://dx.doi.org/10.1287/msom.2021.0999.

Samorani M, LaGanga LR (2015) Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research* 240(1):245–257.

Samuels RC, Ward VL, Melvin P, Macht-Greenberg M, Wenren LM, Yi J, Massey G, Cox JE (2015) Missed appointments: factors contributing to high no-show rates in an urban pediatrics primary care clinic. *Clinical pediatrics* 54(10):976–982.

Schlotheuber A, Hosseinpoor AR (2022) Summary measures of health inequality: A review of existing measures and their application. *International Journal of Environmental Research and Public Health* 19(6):3697, ISSN 1660-4601, URL http://dx.doi.org/10.3390/ijerph19063697, the authors declare no conflict of interest. The authors are staff members of the World Health Organization. The authors alone are responsible for the views

expressed in this article and they do not necessarily represent the decisions, policy or views of the World Health Organization.

Turkcan A, Zeng B, Muthuraman K, Lawley M (2011) Sequential clinical scheduling with service criteria. *European Journal of Operational Research* 214(3):780–795, ISSN 0377-2217, URL http://dx.doi.org/https://doi.org/10.1016/j.ejor.2011.05.023.

Wagstaff A, Paci P, van Doorslaer E (1991) On the measurement of inequalities in health. *Social Science & Medicine* 33(5):545–557, ISSN 0277-9536, URL http://dx.doi.org/https://doi.org/10.1016/0277-9536(91)90212-U.

Wang P (1999) Sequencing and scheduling n customers for a stochastic server. *European Journal of Operational Research* 119(3):729–738, ISSN 0377-2217, URL http://dx.doi.org/https://doi.org/10.1016/S0377-2217(98)00340-3.

Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics (NRL)* 40(3):345–360, URL http://dx.doi.org/https://doi.org/10.1002/1520-6750(199304)40:3<345::AID-NAV3220400305>3.0.CO;2-N.

Weiss EN (1990) Models for determining estimated start times and case orderings in hospital operating rooms. *IIE transactions* 22(2):143–150.

Wisniewski JM, Walker B (2020) Association of simulated patient race/ethnicity with scheduling of primary care appointments. *JAMA Network Open* 3(1):e1920010–e1920010.

Xinying Chen V, Hooker J (2023) A guide to formulating fairness in an optimization model. *Annals of Operations Research* 326:581–619, URL http://dx.doi.org/https://doi.org/10.1007/s10479-023-05264-y.

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production and Operations Management* 23(5):788–801, URL http://dx.doi.org/https://doi.org/10.1111/poms.12065.

Online Appendix to Fair and Efficient Scheduling with Stratified No-Show Prediction

Appendix A: Numerical Study

Our numerical experiments cover a wide range of system configurations including different lengths of the scheduling period, $T = 10$ and $T = 30$, different panel sizes, $\Lambda = 1.2T/p, 1.5T/p$, and $1.8T/p$, different show-up probabilities, $(p_L, p_H) = (0.6, 0.8)$, $(p_L, p_H) = (0.3, 0.7)$, $(p_L, p_H) = (0.2, 0.3)$, and different compositions of patients, $\gamma = 0.25, 0.5, 0.75$. We also consider both exponentially distributed and deterministic service times.

For each system configuration, to study the trade-off between different performance metrics or to search for the best performing policy with the problem formulations introduced in Section 4.2, we consider different slot lengths ($\varepsilon = 0, 0.1, ..., 1 - p$), different numbers of overbooked patients in the first slot ($\kappa = 0, 1, \ldots, \lfloor \Lambda - T/(p + \varepsilon) \rfloor$), and different sequencing rules (π_R, π_L, and π_H). In total (across different system configurations and policies) we evaluate 99,780 unique systems.

All results are available at `https://tinyurl.com/mwhfw4jr`. The observations in Section 4 are consistent across all system configurations and problem instances. This indicates that our findings are not restricted to a narrow set of assumptions and demonstrate the robustness of our results under various conditions.

In this section, we provide some additional examples. In Figures 9, 10, and 11, we show performance trends and trade-offs for a system with $T = 30$, $\Lambda = 47$, $(p_L, p_H) = (0.6, 0.8)$, $\gamma = 0.25$, $S_i = 1$. In Table 1, we show 5 examples out of 11,772 case study problems discussed in Section 4.2. The detailed system configuration and problem parameters can be found in the table.

Appendix B: Fluid Analysis: Evolution of $x^j(t)$

B.1. Sequencing Rule π_H

In Figure 12, we plot the evolution of $x^{\pi_H}(t)$ with $p + \varepsilon < p_H$, based on a numerical example. At time 0, κ patients are overbooked, and κp_H patients show up. The value of $x^{\pi_H}(t)$ then increases with time at a rate of $\left(\frac{p_H}{p+\varepsilon} - 1\right)$ until $t = (\Lambda_H - \kappa)(p + \varepsilon)$, since all patients up to this point belong to group H. Afterward, $x^{\pi_H}(t)$ decreases at a rate of $\left(1 - \frac{p_L}{p+\varepsilon}\right)$ until time T, when patients from group L are scheduled. At time T, $\bar{\kappa}_\varepsilon - \kappa$ patients are overbooked, and the total number of show-ups is $(\bar{\kappa}_\varepsilon - \kappa)p_L$. Finally, after time T, no more patients are scheduled, and $x^{\pi_H}(t)$ decays at a rate of 1. Here is an explicit characterization of $x^{\pi_H}(t)$.

For the π_H rule, we have

$$x^{\pi_H}(t) = \begin{cases} \left(\kappa p_H + \left(\frac{p_H}{p+\varepsilon} - 1\right) t\right)^+, \text{ if } 0 \le t < (\Lambda_H - \kappa)(p + \varepsilon), \\ \left(\left(\kappa p_H - \left(1 - \frac{p_H}{p+\varepsilon}\right)(\Lambda_H - \kappa)(p + \varepsilon)\right)^+ + \left(\frac{p_L}{p+\varepsilon} - 1\right)(t - (\Lambda_H - \kappa)(p + \varepsilon))\right)^+ \\ = \left(\kappa p_H + \frac{p_H - p_L}{p+\varepsilon}(\Lambda_H - \kappa)(p + \varepsilon) - \left(1 - \frac{p_L}{p+\varepsilon}\right) t\right)^+, \text{ if } (\Lambda_H - \kappa)(p + \varepsilon) \le t < T, \\ (v^{\pi_H} - (t - T))^+, \text{ if } t \ge T. \end{cases} \tag{5}$$

- For $0 \le t < (\Lambda_H - \kappa)(p + \varepsilon)$: This phase represents the initial scheduling of H patients. The expression $\left(\kappa p_H + \left(\frac{p_H}{p+\varepsilon} - 1\right) t\right)^+$ reflects the evolution of the number of high-probability patients in the system. The term κp_H accounts for the initial overbooked patients who show up, while the slope $\left(\frac{p_H}{p+\varepsilon} - 1\right)$ captures the arrival and service rate of high-probability patients.

Figure 9 **System performance across all three sequencing rules for varying κ values.**

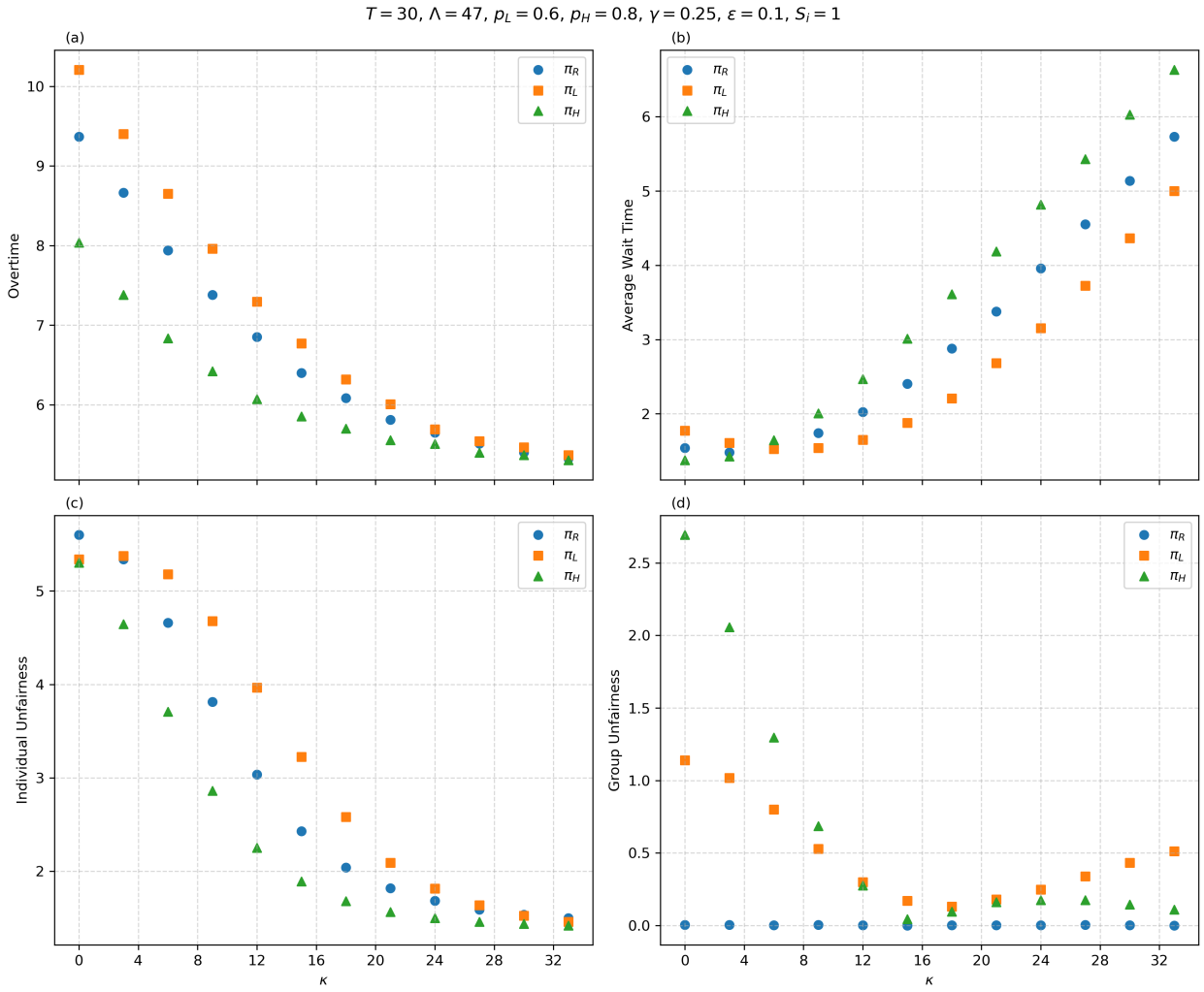


Figure 10 **The relationship between individual unfairness, overtime, and wait time under different scheduling designs and π_R.**

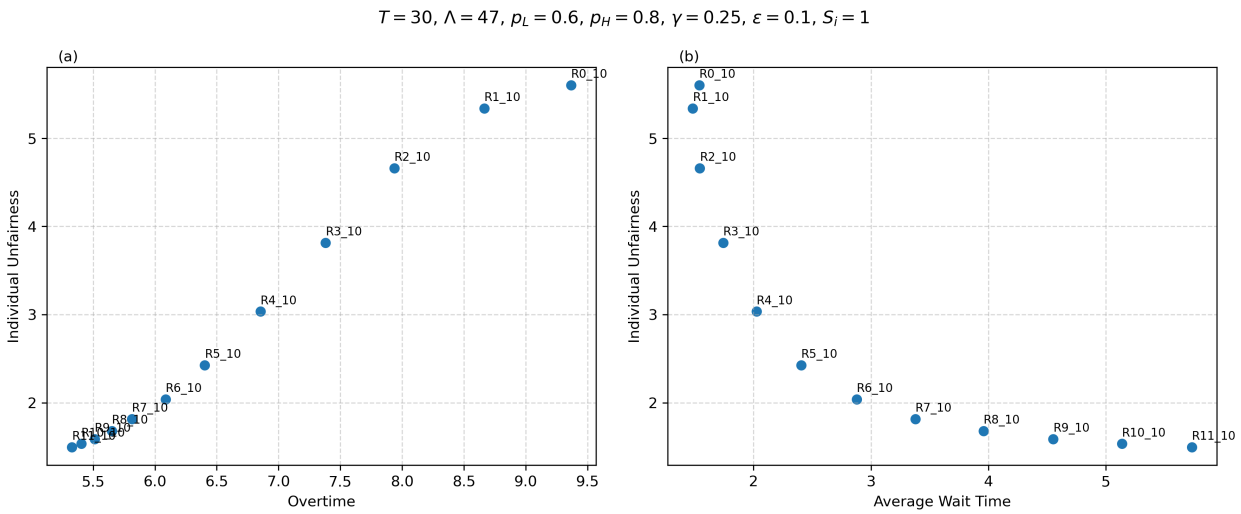


Figure 11 **Pareto frontier of trade-offs between overtime, wait time, and fairness across sequencing rules.**

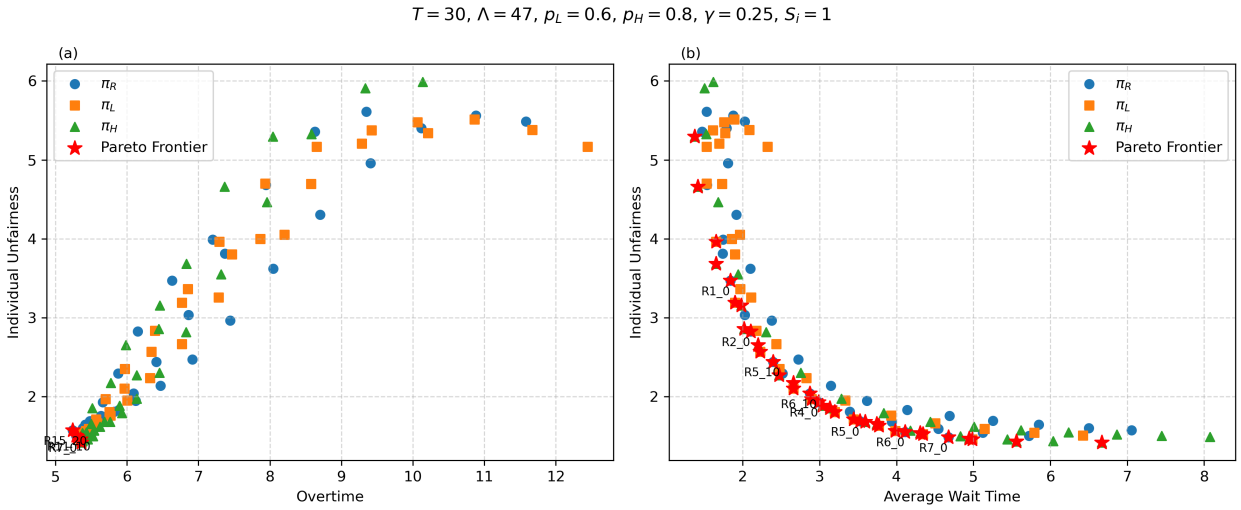
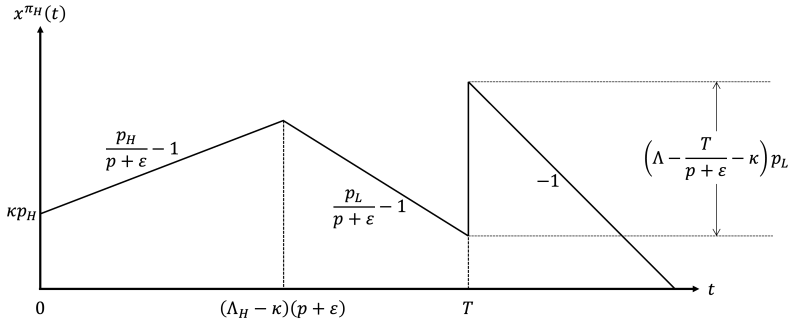


Table 1 **Five examples of the Case Study results.**

	(i)	(ii)	(iii)	(iv)	(v)
T	10	30	10	30	10
p_H	0.8	0.8	0.7	0.7	0.3
p_L	0.6	0.6	0.3	0.3	0.2
γ	0.5	0.25	0.25	0.5	0.75
Λ	17	47	20	20	53
S_i	1	0	0	1	1
w_o	1	2	0	0	0
w_{if}	2	0	0	0	0
w_{gf}	0	2	0	0	0
c_o			50%	75%	max
c_{if}			50%	75%	max
c_{gf}			50%	max	max
top 10 policies	R4_10, H2_0, H4_10, L7_30, L4_10, R5_20, L6_30, H3_10, H5_20, L5_20	R3_0, R4_0, R2_0, R5_0, H3_0, R6_0, H5_10, L7_10, R7_10, R6_10	H3_30, R3_10, R2_0, H2_10, R4_20, H4_40, H1_0, H3_20, R4_10, H4_30	R6_10, H1_20, H0_20, H4_30, H2_20, R7_10, R0_0, H6_40, R1_0, R2_0	H3_10, R6_10, H5_10, R7_10, L3_0, H4_10, R4_10, H1_10, R5_10, L8_10
objective values	9.07, 9.12, 9.15, 9.2, 9.25, 9.25, 9.3, 9.31, 9.34, 9.35	14.27, 14.31, 14.41, 14.49, 14.79, 14.86, 14.9, 14.94, 15.04, 15.06	1.22, 1.25, 1.33, 1.33, 1.37, 1.39, 1.43, 1.45, 1.58, 1.59	3.41, 3.48, 3.48, 3.49, 3.5, 3.53, 3.59, 3.59, 3.59, 3.59	1.72, 1.74, 1.75, 1.76, 1.76, 1.76, 1.76, 1.76, 1.76, 1.77
range	5.32	15.39	2.42	10.5	2.9
best π_R	R4_10	R3_0	R3_10	R6_10	R6_10
best π_R objective	9.07	14.27	1.25	3.41	1.74
gap in %	0	0	3.13	0	1.06

Figure 12 Fluid level in the system under π_H at each time t.



- For $(\Lambda_H-\kappa)(p+\varepsilon)\le t<T$: After H patients have been served, L patients are served next. The slope $\left(\frac{p_L}{p+\varepsilon}-1\right)$ is adjusted for the new arrival rate. With $t=(\Lambda_H-\kappa)(p+\varepsilon)$, the expression $\left(\kappa p_H-\left(1-\frac{p_H}{p+\varepsilon}\right)(\Lambda_H-\kappa)(p+\varepsilon)\right)^+$ captures the amount of fluid at time $(\Lambda_H-\kappa)(p+\varepsilon)$.
- For $t\ge T$: This phase represents the overtime period, where v^{π_H} patients are in the system at T and patients are served at a rate of 1. The explicit expression of v^{π_H} is given in the proof of Proposition 2 in Appendix E.

B.2. Sequencing Rule π_R

For the π_R rule, the overbooked fluid that shows up at time 0 is κp, which decays at a rate of $\frac{p}{p+\varepsilon}-1=\frac{\varepsilon}{p+\varepsilon}$. Therefore, for $0\le t<T$, $x^{\pi_R}(t)=\left(\kappa p-\frac{\varepsilon}{p+\varepsilon}t\right)^+$. At time T, the overbooked fluid that shows up is $(\Lambda-T/(p+\varepsilon)-\kappa)p$, so $x^{\pi_R}(T)=\left(\kappa p-\frac{\varepsilon}{p+\varepsilon}T\right)^+ +(\Lambda-T/(p+\varepsilon)-\kappa)p$. Thus, the expression for $x^{\pi_R}(t)$ in (4) is obtained.

B.3. Sequencing Rule π_L

For the π_L rule, where L patients are scheduled before H patients, the analysis mirrors that for the π_H rule, but with L patients served earlier. The resulting expression for $x^{\pi_L}(t)$:

$$x^{\pi_L}(t)=\begin{cases}\left(\kappa p_L-\left(1-\frac{p_L}{p+\varepsilon}\right)t\right)^+, & \text{if } 0\le t<(\Lambda_L-\kappa)(p+\varepsilon),\\ \left(\left(\kappa p_L-\left(1-\frac{p_L}{p+\varepsilon}\right)(\Lambda_L-\kappa)(p+\varepsilon)\right)^+ +\left(\frac{p_H}{p+\varepsilon}-1\right)(t-(\Lambda_L-\kappa)(p+\varepsilon))\right)^+, & \\ \quad \text{if } (\Lambda_L-\kappa)(p+\varepsilon)\le t<T,\\ (v^{\pi_L}-(t-T))^+, & \text{if } t\ge T.\end{cases} \tag{6}$$

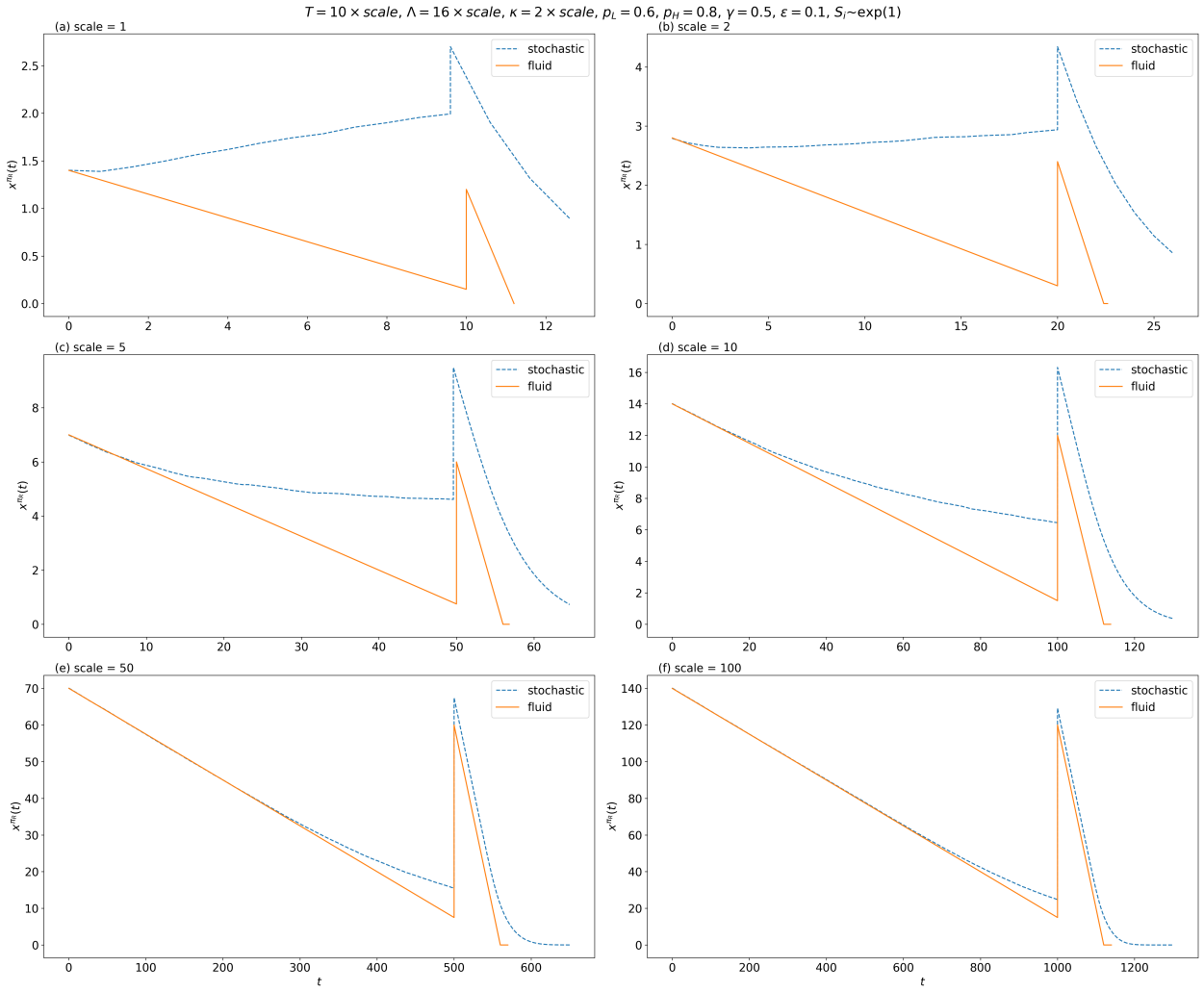
The explicit characterization of v^{π_L} is given in the proof of Proposition 2 in Appendix E.

Appendix C: Connecting the Stochastic and Fluid Models

To explore the accuracy of the fluid approximation, for a fixed slot length, we let T increase, while increasing the census level Λ proportionally. In Figure 13, we plot the trajectories for $x^{\pi_R}(t)$ in the fluid and stochastic models, for $T=10$, 20, 50, 100, 500, and 1000. The plots illustrate how the trajectories evolve as T and Λ increase.

In Figure 13, we estimate the stochastic model curves by averaging over 10^4 simulation replications. For $T=10$, the stochastic and fluid trajectories are distinct, reflecting the inherent variability and randomness in smaller systems. However, as T increases, these trajectories begin to converge, revealing similar patterns by $T=50$. As T continues to increase, the curves overlap significantly and by $T=500$, the fluid and stochastic trajectories are nearly indistinguishable. In these larger systems, the fluid model serves as an effective approximation, capturing the overall system dynamics while smoothing out individual variations that are more pronounced in smaller systems.

Figure 13 **System trajectories comparing the numbers of patients in the system over time for the stochastic and fluid models across different session lengths.**

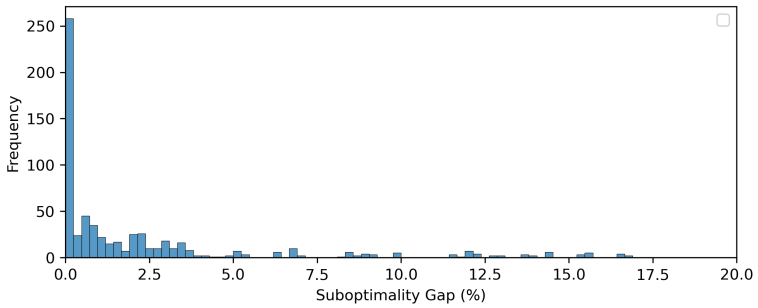


Appendix D: Plateau-Dome Policies

In our main analysis, we focus exclusively on a special class of "plateau-dome" policies. In this section, we expand the search space to also look at other scheduling policies. In particular, we remove the restriction that overbooking occurs only at the beginning and end of the session, while keeping other system characterizations the same. In this case, overbooking is allowed at any slot during the session.

For computational traceability, we investigate system configurations with $T = 5, 6, 7$, $(p_L, p_H) = (0.6, 0.8)$, $\gamma = 0.5$, $\Lambda = 1.2T/p$, and deterministic and exponential service times with mean 1. For each system configuration $(T, \Lambda, p_L, p_H, \gamma, S_i)$, we explore the sequencing rules (π_R, π_H, π_L), vary $\varepsilon = 0, 0.05, 0.1, 0.15, ..., 0.3$, and consider the unrestricted overbooking design described above. In total, we evaluated 12,516 systems. The plateau-dome policies are assessed in this expanded policy set as well. We consider problems (2) and (3) in Section 4.2, proceed similarly to that section, and report the 654 suboptimality gaps for the best performing plateau-dome policies.

Figure 14 Histogram of suboptimality gaps for plateau-dome policies.



We can see in Figure 14 that in most of the scenarios tested, the suboptimality gaps are less than 5%. Thus, by narrowing the design space to the more restricted special class of plateau-dome policies, it is still possible to achieve near-optimal performance.

Appendix E: Proofs of Propositions

Proof of Proposition 1 We discuss the cases with constant service time and exponential service time separately.

Constant service times. Assume that the service time is constant, that is, $S_i = 1$. We consider the following two subcases depending on the value of Δ_i.

(a) $\Delta_i \geq 1$:

$W_0 = 0$, $S_0 = 0$, $\Delta_1 = 0$, so $W_1 = 0$, $W_2 = (I_1 - \Delta_2)^+ = 0$. Therefore, $W_i = 0$ and $\mathbb{E}[W_i] = 0$ for all i.

Meanwhile, since we have $\Delta_i \geq 1 > p_{i-1}$, there exists $\alpha_i > 0$, such that $\Delta_i \geq p_{i-1} + \alpha_i$.

(b) $\Delta_i < 1$:

If $0 \leq W_{i-1} \leq \Delta_i < 1$, then $1 - (\Delta_i - W_{i-1}) > 0$ and $0 \leq \Delta_i - \mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i] \leq \Delta_i < 1$. Therefore, we have

$$\begin{aligned}\mathbb{E}[W_i \mid W_{i-1} > \Delta_i] &= \mathbb{E}[W_{i-1} + I_{i-1} - \Delta_i \mid W_{i-1} > \Delta_i] \\ &= \mathbb{E}[W_{i-1} \mid W_{i-1} > \Delta_i] + (p_{i-1} - \Delta_i).\end{aligned}$$

$$\begin{aligned}\mathbb{E}[W_i \mid W_{i-1} \leq \Delta_i] &= \mathbb{E}[(W_{i-1} + I_{i-1} - \Delta_i)^+ \mid W_{i-1} \leq \Delta_i] \\ &= p_{i-1}\,\mathbb{E}[W_{i-1} + 1 - \Delta_i \mid W_{i-1} \leq \Delta_i,\, I_{i-1} = 1] \\ &= p_{i-1}\,\mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i] + p_{i-1}(1 - \Delta_i).\end{aligned}$$

$$\begin{aligned}\mathbb{E}[W_i] &= \mathbb{E}[W_i \mid W_{i-1} > \Delta_i]\,\mathbb{P}(W_{i-1} > \Delta_i) + \mathbb{E}[W_i \mid W_{i-1} \leq \Delta_i]\,\mathbb{P}(W_{i-1} \leq \Delta_i) \\ &= (\mathbb{E}[W_{i-1} \mid W_{i-1} > \Delta_i] + (p_{i-1} - \Delta_i))\,\mathbb{P}(W_{i-1} > \Delta_i) + (p_{i-1}\,\mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i] + p_{i-1}(1 - \Delta_i))\,\mathbb{P}(W_{i-1} \leq \Delta_i) \\ &= \mathbb{E}[W_{i-1}] - (1 - p_{i-1})\,\mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i]\,\mathbb{P}(W_{i-1} \leq \Delta_i) + (p_{i-1} - \Delta_i) + (1 - p_{i-1})\Delta_i\,\mathbb{P}(W_{i-1} \leq \Delta_i) \\ &= \mathbb{E}[W_{i-1}] + p_{i-1} - \Delta_i + (1 - p_{i-1})(\Delta_i - \mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i])\,\mathbb{P}(W_{i-1} \leq \Delta_i).\end{aligned}$$

$$\mathbb{E}[W_i] \leq \mathbb{E}[W_{i-1}] \Leftrightarrow p_{i-1} \leq \frac{\Delta_i - \mathbb{P}(W_{i-1} \leq \Delta_i)(\Delta_i - \mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i])}{1 - \mathbb{P}(W_{i-1} \leq \Delta_i)(\Delta_i - \mathbb{E}[W_{i-1} \mid W_{i-1} \leq \Delta_i])} =: \frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i},$$

where $0 < \tilde{\alpha}_i \leq \Delta_i < 1$. Moreover, with $\Delta_i - \tilde{\alpha}_i < \Delta_i - \Delta_i\tilde{\alpha}_i \Leftrightarrow \frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i} < \Delta_i$, we let $\alpha_i := \Delta_i - \frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i}$, where $\alpha_i > 0$. That is, $\mathbb{E}[W_i] \leq \mathbb{E}[W_{i-1}] \Leftrightarrow p_{i-1} \leq \Delta_i - \alpha_i \Leftrightarrow \Delta_i \geq p_{i-1} + \alpha_i$.

Exponential service times. Assume that the service time S_i is exponentially distributed with $\mathbb{E}[S_i] = 1$.

We have $\mathbb{E}[S_i | S_i > s] = s + 1$ and

$$\begin{aligned}\mathbb{E}[W_i \mid W_{i-1} = w] &= \mathbb{E}[(w + S_{i-1}I_{i-1} - \Delta_i)^+] \\ &= (1 - p_{i-1})(w - \Delta_i)^+ + p_{i-1}\,\mathbb{E}[(w + S_{i-1} - \Delta_i)^+].\end{aligned}$$

$$\begin{aligned}\mathbb{E}[(w + S_{i-1} - \Delta_i)^+] &= \mathbb{E}[w + S_{i-1} - \Delta_i \mid S_{i-1} > \Delta_i - w]\,\mathbb{P}(S_{i-1} > \Delta_i - w) \\ &= (w + (1 + (\Delta_i - w)^+) - \Delta_i)e^{-(\Delta_i - w)^+}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[W_i] &= \mathbb{E}[\mathbb{E}[W_i \mid W_{i-1}]] \\ &= \mathbb{E}\left[(1 - p_{i-1})(W_{i-1} - \Delta_i)^+ + p_{i-1}(W_{i-1} + (1 + (\Delta_i - W_{i-1})^+) - \Delta_i)e^{-(\Delta_i - W_{i-1})^+}\right] \\ &= \mathbb{E}\left[p_{i-1}e^{-(\Delta_i - W_{i-1})} \mid W_{i-1} \le \Delta_i\right]\mathbb{P}(W_{i-1} \le \Delta_i) + \mathbb{E}\left[W_{i-1} - \Delta_i + p_{i-1} \mid W_{i-1} > \Delta_i\right]\mathbb{P}(W_{i-1} > \Delta_i) \\ &= \mathbb{E}[W_{i-1}] + (p_{i-1} - \Delta_i) \\ &\quad + \left(\Delta_i - \mathbb{E}[W_{i-1} \mid W_{i-1} \le \Delta_i] - p_{i-1} + p_{i-1}\,\mathbb{E}[e^{-(\Delta_i - W_{i-1})} \mid W_{i-1} \le \Delta_i]\right)\mathbb{P}(W_{i-1} \le \Delta_i) \\ &= \mathbb{E}[W_{i-1}] + (p_{i-1} - \Delta_i) + (1 - p_{i-1})(\Delta_i - \mathbb{E}[W_{i-1} \mid W_{i-1} \le \Delta_i])\,\mathbb{P}(W_{i-1} \le \Delta_i) \\ &\quad + p_{i-1}\,\mathbb{E}[e^{-(\Delta_i - W_{i-1})} - (1 - (\Delta_i - W_{i-1})) \mid W_{i-1} \le \Delta_i]\,\mathbb{P}(W_{i-1} \le \Delta_i) \\ &:= \mathbb{E}[W_{i-1}] + (p_{i-1} - \Delta_i) + (1 - p_{i-1})\tilde{\alpha}_i + p_{i-1}\tilde{\beta}_i,\end{aligned}$$

where $\tilde{\beta}_i \ge 0$ as $e^{-z} \ge 1 - z$ for all $z \ge 0$ and $0 \le \tilde{\alpha}_i - \tilde{\beta}_i = \mathbb{E}[1 - e^{-(\Delta_i - W_{i-1})} \mid W_{i-1} \le \Delta_i]\,\mathbb{P}(W_{i-1} \le \Delta_i) < 1$.

Therefore, similar as case 1 above, $\mathbb{E}[W_i] \le \mathbb{E}[W_{i-1}] \Leftrightarrow p_{i-1} \le \frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i + \tilde{\beta}_i}$.

Furthermore, if $\tilde{\alpha}_i = \tilde{\beta}_i$, then we have $\mathbb{E}[W_i] \le \mathbb{E}[W_{i-1}] \Leftrightarrow p_{i-1} \le \Delta_i - \tilde{\alpha}_i$ and we let $\alpha_i := \tilde{\alpha}_i$, where $\tilde{\alpha}_i > 0$. Otherwise, $\frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i + \tilde{\beta}_i} < \Delta_i \Leftrightarrow \Delta_i - \tilde{\alpha}_i < \Delta_i(1 - \tilde{\alpha}_i + \tilde{\beta}_i) \Leftrightarrow \Delta_i < \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - \tilde{\beta}_i}$. We then let $\alpha_i := \Delta_i - \frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i + \tilde{\beta}_i}$ if $\Delta_i < \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - \tilde{\beta}_i}$. That is, $\mathbb{E}[W_i] \le \mathbb{E}[W_{i-1}] \Leftrightarrow p_{i-1} \le \Delta_i - \alpha_i \Leftrightarrow \Delta_i \ge p_{i-1} + \alpha_i$ if $\Delta_i < \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - \tilde{\beta}_i}$. When $\Delta_i \ge \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - \tilde{\beta}_i}$, we have $\frac{\Delta_i - \tilde{\alpha}_i}{1 - \tilde{\alpha}_i + \tilde{\beta}_i} \ge \Delta_i \ge \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - \tilde{\beta}_i} > 1 > p_{i-1}$. That is $\mathbb{E}[W_i] \le \mathbb{E}[W_{i-1}]$, and $\Delta_i > p_{i-1}$, in which case, there exists $\alpha_i > 0$, such that $\Delta_i \ge p_{i-1} + \alpha_i$.

Proof of Proposition 2 For any fixed $\varepsilon \ge 0$, we define

$$\delta := \Lambda - \frac{T}{p},\ \delta_\varepsilon := \Lambda - \frac{T}{p + \varepsilon} = \bar{\kappa}_\varepsilon.\ \text{So we have } 0 \le \kappa \le \delta_\varepsilon,\ \delta = \delta_{\varepsilon=0},\ \delta_\varepsilon = \delta + \frac{\varepsilon T}{p(p + \varepsilon)}. \tag{7}$$

Proof of part 1 of the proposition. Let $y^j(t)$ where $j = \pi_L, \pi_R, \pi_H$ be the remaining of the initial overbooking at time t, which is allowed to be negative. Since $x^{\pi_R}(t) = \left(\kappa p - \frac{\varepsilon}{p + \varepsilon}t\right)^+$ for $0 \le t < T$ by (4), we have $y^{\pi_R}(t) = \kappa p - \frac{\varepsilon}{p + \varepsilon}t$. Moreover,

$$y^{\pi_R}(T) > 0 \Leftrightarrow \kappa p - \frac{\varepsilon}{p + \varepsilon}T > 0 \Leftrightarrow \kappa > \frac{\varepsilon T}{p(p + \varepsilon)} \Leftrightarrow (\delta_\varepsilon - \kappa)p < \delta p \text{ (by (7))}, \text{ and}$$

$$y^{\pi_R}(T) + (\delta_\varepsilon - \kappa)p = \kappa p - \frac{\varepsilon T}{p + \varepsilon} + \left(\delta + \frac{\varepsilon T}{p(p + \varepsilon)} - \kappa\right)p = \delta p.$$

Note that $(\delta_\varepsilon - \kappa)p$ is the show-up amount of overbooking at T. That is, if $y^{\pi_R}(T) > 0$, then $(\delta_\varepsilon - \kappa)p < \delta p$ and $x^{\pi_R}(T) = y^{\pi_R}(T) + (\delta_\varepsilon - \kappa)p = \delta p$; otherwise, if $y^{\pi_R}(T) \le 0$, then $(\delta_\varepsilon - \kappa)p \ge \delta p$ and $x^{\pi_R}(T) = 0 + (\delta_\varepsilon - \kappa)p$. As $\frac{dx^{\pi_R}(t)}{dt} = -1$ for $t \ge T$,

$$v^{\pi_R} = x^{\pi_R}(T) = (\delta_\varepsilon - \kappa)p + [y^{\pi_R}(T)]^+ = \max\{(\delta_\varepsilon - \kappa)p, \delta p\}, \tag{8}$$

in other words, $v^{\pi_R} = (\delta_\varepsilon - \kappa)p$ if $\kappa < \frac{\varepsilon T}{p(p + \varepsilon)} =: \kappa_R = \delta_\varepsilon - \delta$, otherwise $v^{\pi_R} = \delta p$. Note that $0 \le \kappa_R < \delta_\varepsilon$.

Similarly, by (5), $\Lambda p = \Lambda_H p_H + \Lambda_L p_L$, and $\Lambda_L + \Lambda_H = \Lambda$, we have $y^{\pi_H}(T) = \kappa p_L + (p_H - p_L)\Lambda_H - \frac{p+\varepsilon-p_L}{p+\varepsilon}T = \kappa p_L + \Lambda p - T - \Lambda p_L + \frac{T}{p+\varepsilon}p_L = \kappa p_L + \delta p - \delta_\varepsilon p_L$. Then again $(\delta_\varepsilon - \kappa)p_L$ is the show-up amount of overbooking at T, and we have

$$y^{\pi_H}(T) > 0 \Leftrightarrow (\delta_\varepsilon - \kappa)p_L < \delta p \text{ and } y^{\pi_H}(T) + (\delta_\varepsilon - \kappa)p_L = \delta p,$$

$$v^{\pi_H} = \max\{(\delta_\varepsilon - \kappa)p_L, \delta p\}.$$

Similarly, we define κ_H such that $(\delta_\varepsilon - \kappa_H)p_L = \delta p$. Then for $\varepsilon > 0$, $\kappa_H < \kappa_R \Leftrightarrow \delta_\varepsilon - \delta\frac{p}{p_L} < \delta_\varepsilon - \delta \Leftrightarrow \delta > 0$, which is true by assumption. And $\kappa_R = \kappa_H = 0$ for $\varepsilon = 0$. Therefore, for any $\kappa \in [0, \kappa_H)$, $v^{\pi_H} = (\delta_\varepsilon - \kappa)p_L \le (\delta_\varepsilon - \kappa)p = v^{\pi_R}$, for any $\kappa \in [\kappa_H, \kappa_R)$, $v^{\pi_H} = \delta p \le (\delta_\varepsilon - \kappa)p = v^{\pi_R}$, and for any $\kappa \ge \kappa_R$, $v^{\pi_H} = v^{\pi_R} = \delta p$. Hence, we have $v^{\pi_H} \le v^{\pi_R}$.

Follow the similar lines of reasoning, we show $v^{\pi_R} \le v^{\pi_L}$ in the following. With (6) and $\Lambda p = \Lambda_H p_H + \Lambda_L p_L$, $\Lambda_L + \Lambda_H = \Lambda$, we have $y^{\pi_L}(T) = (\kappa p_L - (p+\varepsilon-p_L)(\Lambda_L - \kappa))^+ + (p+\varepsilon-p_H)(\Lambda_L - \kappa) - \left(1 - \frac{p_H}{p+\varepsilon}\right)T \ge \kappa p_L + p_L(\Lambda_L - \kappa_- p_H(\Lambda_L - \kappa) - T + \frac{p_H T}{p+\varepsilon} = \kappa p_H + \delta p - \delta_\varepsilon p_H =: \tilde{y}^{\pi_L}(T)$. Meanwhile, $\tilde{y}^{\pi_L}(T) > 0 \Leftrightarrow (\delta_\varepsilon - \kappa)p_H < \delta p$ and $\tilde{y}^{\pi_L}(T) + (\delta_\varepsilon - \kappa)p_H = \delta p$. Therefore, $v^{\pi_L} \ge \max\{(\delta_\varepsilon - \kappa)p_H, \delta p\} =: \tilde{v}^{\pi_L}$ and $v^{\pi_R} \le v^{\pi_L}$, where we define κ_L such that $(\delta_\varepsilon - \kappa_L)p_H = \delta p$ and we have $\kappa_R < \kappa_L < \delta_\varepsilon$.

Hence, $v^{\pi_H} \le v^{\pi_R} \le v^{\pi_L}$.

Proof of part 2 of the proposition. When $\kappa \ge \kappa_R$, we know by the proof of the first part of Proposition 2 that $v^{\pi_R} = v^{\pi_H} = \delta p$, $y^{\pi_R}(T) = \delta p - (\delta_\varepsilon - \kappa)p \ge 0$, and $y^{\pi_H}(T) = \delta p - (\delta_\varepsilon - \kappa)p_L \ge 0$. Therefore, we have

$$\begin{aligned}
\bar{w}^{\pi_R} &= \frac{1}{\Lambda p}\left(\int_0^T x^{\pi_R}(t)dt + \int_T^{T+\Lambda} x^{\pi_R}(t)dt\right) = \frac{1}{\Lambda p}\left(\int_0^T x^{\pi_R}(t)dt + \frac{1}{2}(v^{\pi_R})^2\right) \\
&= \frac{T}{2\Lambda p}(x^{\pi_R}(0) + y^{\pi_R}(T)) + \frac{1}{2\Lambda p}(v^{\pi_R})^2 \qquad (*) \\
&= \frac{T}{2\Lambda p}(\kappa p + \delta p - (\delta_\varepsilon - \kappa)p) + \frac{1}{2\Lambda p}(v^{\pi_R})^2,
\end{aligned}$$

$$\begin{aligned}
\bar{w}^{\pi_H} &= \frac{1}{\Lambda p}\left(\int_0^T x^{\pi_H}(t)dt + \int_T^{T+\Lambda} x^{\pi_H}(t)dt\right) = \frac{1}{\Lambda p}\left(\int_0^T x^{\pi_H}(t)dt + \frac{1}{2}(v^{\pi_H})^2\right) \\
&\ge \frac{T}{2\Lambda p}(x^{\pi_H}(0) + y^{\pi_H}(T)) + \frac{1}{2\Lambda p}(v^{\pi_H})^2 \qquad (**) \\
&= \frac{T}{2\Lambda p}(\kappa p_H + \delta p - (\delta_\varepsilon - \kappa)p_L) + \frac{1}{2\Lambda p}(v^{\pi_R})^2.
\end{aligned}$$

In step (*), the derivation follows from the trapezoid area formula for $t \in [0, T)$. In step (**), the inequality follows because

$$\left.\frac{dx^{\pi_H}(t)}{dt}\right|_{t<(\Lambda_H-\kappa)(p+\varepsilon)} = \frac{p_H}{p+\varepsilon} - 1 > \frac{p_L}{p+\varepsilon} - 1 = \left.\frac{dx^{\pi_H}(t)}{dt}\right|_{(\Lambda_H-\kappa)(p+\varepsilon)\le t<T},$$

then $x^{\pi_H}(t)$ is concave for $t \in [0, T)$. Therefore, the associated area is larger than the trapezoid area consisting of $x^{\pi_H}(0)$, $y^{\pi_H}(T)$, and T. By assumption, we have $p_L < p < p_H$. Hence, $\bar{w}^{\pi_H} > \bar{w}^{\pi_R}$. That is, there exists $0 \le \kappa_1 \le \kappa_R < \delta_\varepsilon = \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\kappa_1, \bar{\kappa}_\varepsilon]$, $\bar{w}^{\pi_H} > \bar{w}^{\pi_R}$.

By (6), $y^{\pi_L}(T) > \tilde{y}^{\pi_L}(T)$ only when $p_H > p+\varepsilon$ and $\kappa p_L - (p+\varepsilon-p_L)(\Lambda_L - \kappa) < 0$, where the latter is equivalent to $\kappa < \Lambda_L \frac{p+\varepsilon-p_L}{p+\varepsilon} := \tilde{\kappa}_L$. Moreover, $\tilde{\kappa}_L < \delta_\varepsilon \Leftrightarrow \Lambda_L p_L + \Lambda_H(p+\varepsilon) > T$, which is true by the assumption of $\Lambda_L p_L + \Lambda_H p > T$. Therefore, similar to above, when $\kappa \ge \max\{\kappa_L, \tilde{\kappa}_L\}$, we have $x^{\pi_L}(0) = \kappa p_L < \kappa p_H = x^{\pi_R}(0)$, $x^{\pi_L}(T) = x^{\pi_R}(T)$, and $x^{\pi_L}(t)$ is convex for $t \in [0, T)$. Hence, we have $\bar{w}^{\pi_R} > \bar{w}^{\pi_L}$. That is, there exists $0 \le \kappa_2 \le \max\{\kappa_L, \tilde{\kappa}_L\} < \delta_\varepsilon = \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\kappa_2, \bar{\kappa}_\varepsilon]$, $\bar{w}^{\pi_R} > \bar{w}^{\pi_L}$.

Proof of part 3 of the proposition. We consider the following two subcases.

(a) By (8), v^{π_R} is a function of κ which is non-increasing in κ. We denote this function as $v^{\pi_R}(\kappa)$. On the other hand, the function $m(\kappa) := \kappa p_H + [(p_H - p - \varepsilon)(\Lambda_H - \kappa)]^+ = \kappa p_H + (p_H - p - \varepsilon)^+(\Lambda_H - \kappa)$ is increasing in κ. When $p_H > p + \varepsilon$, we have $m(\kappa) = \kappa(p+\varepsilon) + \Lambda_H(p_H - p - \varepsilon)$. Moreover, by the assumption of $\Lambda_L p_L + \Lambda_H p > T$, we have $\Lambda_L p_L + \Lambda_H(p+\varepsilon) > \frac{Tp}{p+\varepsilon} \Leftrightarrow \Lambda_H p_H - \Lambda_H(p+\varepsilon) < \Lambda p - \frac{Tp}{p+\varepsilon} \Leftrightarrow \Lambda_H(p_H - p - \varepsilon) < \Lambda p - \frac{Tp}{p+\varepsilon}$. So $m(0) \le \Lambda_H(p_H - p - \varepsilon) < \delta_\varepsilon p = v^{\pi_R}(0)$. Then we define δ_H, such that $v^{\pi_R}(\delta_H) = m(\delta_H)$, where $\delta_H > 0$ and $v^{\pi_R}(\kappa) > m(\kappa)$ for $\kappa < \delta_H$.

By the characterization of $x^j(t)$ in (5) and (4), we have

$$\mathrm{if}^{\pi_R} = \frac{\max\{x^{\pi_R}(0), x^{\pi_R}(T)\}}{\bar{w}^{\pi_R}} = \frac{\max\{\kappa p, v^{\pi_R}\}}{\bar{w}^{\pi_R}},$$

$$\begin{aligned}\mathrm{if}^{\pi_H} &= \frac{\max\{x^{\pi_H}(0), x^{\pi_H}((\Lambda_H - \kappa)(p+\varepsilon)), x^{\pi_H}(T)\}}{\bar{w}^{\pi_H}} \\ &= \frac{\max\{\kappa p_H, \left(\kappa p_H + \frac{p_H - p_L}{p+\varepsilon}(\Lambda_H - \kappa)(p+\varepsilon) - \left(1 - \frac{p_L}{p+\varepsilon}\right)(\Lambda_H - \kappa)(p+\varepsilon)\right)^+, v^{\pi_H}\}}{\bar{w}^{\pi_H}} \\ &= \frac{\max\{\kappa p_H + [(p_H - p - \varepsilon)(\Lambda_H - \kappa)]^+, v^{\pi_H}\}}{\bar{w}^{\pi_H}} = \frac{\max\{m(\kappa), v^{\pi_H}\}}{\bar{w}^{\pi_H}}.\end{aligned}$$

From the proof of the second part of Proposition 2 above, we know that for all $\kappa \ge \kappa_1$, $\bar{w}^{\pi_R} < \bar{w}^{\pi_H}$. Moreover, for $\kappa \le \delta_H$, we have $v^{\pi_R}(\kappa) \ge m(\kappa)$. From the first part of Proposition 2, $v^{\pi_R} \ge v^{\pi_H}$. Therefore, $\mathrm{if}^{\pi_R} > \mathrm{if}^{\pi_H}$ for $\kappa_1 \le \kappa \le \delta_H$. That is, there exist $0 \le \underline{\kappa}_1 \le \kappa_1 < \delta_H \le \overline{\kappa}_1 \le \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\underline{\kappa}_1, \overline{\kappa}_1]$, $\mathrm{if}^{\pi_H} < \mathrm{if}^{\pi_R}$.

(b) Similarly, we define δ_L such that $\tilde{v}^{\pi_L}(\delta_L) = \delta_L p$. Since $\tilde{v}^{\pi_L}(0) > 0$, we have $\delta_L > 0$ and $\tilde{v}^{\pi_L}(\kappa) > \kappa p$ for $\kappa < \delta_L$. By (6), we have

$$\mathrm{if}^{\pi_L} = \frac{\max\{\kappa p_L, v^{\pi_L}\}}{\bar{w}^{\pi_L}}.$$

Moreover, we know that $v^{\pi_R} \le v^{\pi_L}$ and for all $\kappa \ge \kappa_2$, $\bar{w}^{\pi_L} < \bar{w}^{\pi_R}$. Therefore, with $v^{\pi_R} \ge \tilde{v}^{\pi_L}(\kappa) > \kappa p$, we have $\mathrm{if}^{\pi_L} > \mathrm{if}^{\pi_R}$ for $\kappa_2 \le \kappa \le \delta_L$. That is, there exist $0 \le \underline{\kappa}_2 \le \kappa_2 < \delta_L \le \overline{\kappa}_2 \le \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\underline{\kappa}_2, \overline{\kappa}_2]$, $\mathrm{if}^{\pi_R} < \mathrm{if}^{\pi_L}$.

Proof of Proposition 3 Since $v^{\pi_R} = \max\{(\delta_\varepsilon - \kappa)p, \delta p\}$, we have $\frac{dv^{\pi_R}}{d\kappa} \le 0$.

Proof of part 1 of the proposition. We consider the following two cases.

(a) Case 1: $\Lambda < \frac{2T}{p}$.

(i) $\kappa_R \ge \delta \Leftrightarrow (2T - \Lambda p)\varepsilon \ge (\Lambda p - T)p \Leftrightarrow \varepsilon \ge \frac{(\Lambda p - T)p}{2T - \Lambda p} = \frac{\delta p^2}{2T - \Lambda p}$.

Then for $\kappa \ge \kappa_R \ge \delta$, we have $v^{\pi_R} = \delta p$, $\bar{w}^{\pi_R} = \frac{1}{2\Lambda p}((\kappa p + \delta p - (\delta_\varepsilon - \kappa)p)T + (v^{\pi_R})^2)$, and $\mathrm{if}^{\pi_R} = \frac{\kappa p}{\bar{w}^{\pi_R}}$. Then $\frac{dv^{\pi_R}}{d\kappa} = 0$, $\frac{d\bar{w}^{\pi_R}}{d\kappa} = \frac{T}{\Lambda} > 0$, and $\frac{d\mathrm{if}^{\pi_R}}{d\kappa} = \frac{p}{\bar{w}^{\pi_R}} - \frac{\kappa p T}{(\bar{w}^{\pi_R})^2 \Lambda} < 0 \Leftrightarrow \varepsilon > \frac{\delta^2 p^2}{\Lambda(2T - \Lambda p)}$. Moreover, $\frac{d\mathrm{if}^{\pi_R}}{d\bar{w}^{\pi_R}} = \frac{d\mathrm{if}^{\pi_R}}{d\kappa}\frac{d\kappa}{d\bar{w}^{\pi_R}} < 0$ and $\frac{d^2\mathrm{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} = \frac{d(d\mathrm{if}^{\pi_R}/d\bar{w}^{\pi_R})}{d\kappa}\frac{d\kappa}{d\bar{w}^{\pi_R}} = \frac{2p}{(\bar{w}^{\pi_R})^2}\left(\frac{\kappa}{\bar{w}^{\pi_R}} - \frac{\Lambda}{T}\right) > 0 \Leftrightarrow \varepsilon > \frac{\delta^2 p^2}{\Lambda(2T - \Lambda p)}$, which is true as $\frac{\delta^2 p^2}{\Lambda(2T - \Lambda p)} < \frac{\delta p^2}{2T - \Lambda p}$.

(ii) $\kappa_R < \delta \Leftrightarrow \varepsilon < \frac{\delta p^2}{2T - \Lambda p}$.

For $\kappa \ge \delta$, same as above, we have $\frac{dv^{\pi_R}}{d\kappa} = 0$, $\frac{d\bar{w}^{\pi_R}}{d\kappa} > 0$, and $\frac{d\mathrm{if}^{\pi_R}}{d\kappa} < 0 \Leftrightarrow \varepsilon > \frac{\delta^2 p^2}{\Lambda(2T - \Lambda p)}$. Moreover, $\frac{d\mathrm{if}^{\pi_R}}{d\bar{w}^{\pi_R}} < 0$ and $\frac{d^2\mathrm{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} > 0 \Leftrightarrow \varepsilon > \frac{\delta^2 p^2}{\Lambda(2T - \Lambda p)}$.

And for κ where $\kappa_R \le \kappa < \delta$, we have $v^{\pi_R} = \delta p$, $\bar{w}^{\pi_R} = \frac{1}{2\Lambda p}((\kappa p + \delta p - (\delta_\varepsilon - \kappa)p)T + (v^{\pi_R})^2)$, and $\mathrm{if}^{\pi_R} = \frac{\delta p}{\bar{w}^{\pi_R}}$. Then $\frac{dv^{\pi_R}}{d\kappa} = 0$, $\frac{d\bar{w}^{\pi_R}}{d\kappa} = \frac{T}{\Lambda} > 0$, and $\frac{d\mathrm{if}^{\pi_R}}{d\kappa} < 0$. Moreover, $\frac{d\mathrm{if}^{\pi_R}}{d\bar{w}^{\pi_R}} = \frac{d\mathrm{if}^{\pi_R}}{d\kappa}\frac{d\kappa}{d\bar{w}^{\pi_R}} = -\frac{\delta p}{(\bar{w}^{\pi_R})^2} < 0$ and $\frac{d^2\mathrm{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} = \frac{d(d\mathrm{if}^{\pi_R}/d\bar{w}^{\pi_R})}{d\kappa}\frac{d\kappa}{d\bar{w}^{\pi_R}} > 0$.

(b) Case 2: $\Lambda \geq \frac{2T}{p}$.

(i) $\kappa_R < \delta \Leftrightarrow (2T - \Lambda p)\varepsilon < (\Lambda p - T)p$, which is true as $(2T - \Lambda p)\varepsilon < 0 < (\Lambda p - T)p$. Then for κ where $\kappa_R \leq \kappa < \delta$, we have $v^{\pi_R} = \delta p$, $\bar{w}^{\pi_R} = \frac{1}{2\Lambda p}((\kappa p + \delta p - (\delta_\varepsilon - \kappa)p)T + (v^{\pi_R})^2)$, and $\text{if}^{\pi_R} = \frac{\delta p}{\bar{w}^{\pi_R}}$. Then $\frac{dv^{\pi_R}}{d\kappa} = 0$, $\frac{d\bar{w}^{\pi_R}}{d\kappa} = \frac{T}{\Lambda} > 0$, and $\frac{d\text{if}^{\pi_R}}{d\kappa} < 0$. Moreover, $\frac{d\text{if}^{\pi_R}}{d\bar{w}^{\pi_R}} = \frac{d\text{if}^{\pi_R}}{d\kappa}\frac{d\kappa}{d\bar{w}^{\pi_R}} = -\frac{\delta p}{(\bar{w}^{\pi_R})^2} < 0$ and $\frac{d^2\text{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} = \frac{d(d\text{if}^{\pi_R}/d\bar{w}^{\pi_R})}{d\kappa}\frac{d\kappa}{d\bar{w}^{\pi_R}} > 0$.

(ii) For $\kappa \geq \delta$, similar as in Case 1, $\frac{d\text{if}^{\pi_R}}{d\kappa} < 0 \Leftrightarrow (2T - \Lambda p)\varepsilon > \frac{\delta^2 p^2}{\Lambda}$, which can not be true as $\Lambda \geq \frac{2T}{p}$.

Therefore, there exist $0 \leq \tilde{\kappa}_l < \tilde{\kappa}_u \leq \bar{\kappa}_\varepsilon$, such that for any $\kappa \in [\tilde{\kappa}_l, \tilde{\kappa}_u]$, we have

$$\frac{dv^{\pi_R}}{d\kappa} \leq 0, \quad \frac{d\bar{w}^{\pi_R}}{d\kappa} > 0, \quad \frac{d\text{if}^{\pi_R}}{d\kappa} < 0, \quad \text{and} \quad \frac{d^2\text{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} > 0.$$

Proof of part 2 of the proposition. As $v^{\pi_R} = \max\{(\delta_\varepsilon - \kappa)p, \delta p\}$, we have that for $\varepsilon < \frac{\kappa p^2}{T - \kappa p} := \varepsilon_R$, $v^{\pi_R} = \delta p$, for $\varepsilon \geq \varepsilon_R$, $v^{\pi_R} = (\delta_\varepsilon - \kappa)p$. Therefore, $\frac{dv^{\pi_R}}{d\varepsilon} \geq 0$.

Moreover, when $\kappa > 0$, for $\varepsilon \leq \varepsilon_R$, $\bar{w}^{\pi_R} = \frac{1}{2\Lambda p}\left(2\kappa p - \frac{T\varepsilon}{p+\varepsilon}\right)T + \frac{(v^{\pi_R})^2}{2\Lambda p}$. Then $\frac{d\bar{w}^{\pi_R}}{d\varepsilon} = -\frac{T^2}{2\Lambda(p+\varepsilon)^2} < 0$. Furthermore, for $\varepsilon \leq \varepsilon_R$, $\text{if}^{\pi_R} = \frac{\max\{\delta p, \kappa p\}}{\bar{w}^{\pi_R}}$. Therefore, $\frac{d\text{if}^{\pi_R}}{d\varepsilon} > 0$. And $\frac{d\text{if}^{\pi_R}}{d\bar{w}^{\pi_R}} = -\frac{\max\{\delta p, \kappa p\}}{(\bar{w}^{\pi_R})^2}$, $\frac{d^2\text{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} = \frac{d(d\text{if}^{\pi_R}/d\bar{w}^{\pi_R})}{d\varepsilon}\frac{d\varepsilon}{d\bar{w}^{\pi_R}} > 0$.

Hence, for any fixed $\kappa > 0$, there exists $\tilde{\varepsilon} \geq \varepsilon_R > 0$, such that for any $\varepsilon \in [0, \tilde{\varepsilon}]$, we have

$$\frac{dv^{\pi_R}}{d\varepsilon} \geq 0, \quad \frac{d\bar{w}^{\pi_R}}{d\varepsilon} < 0, \quad \frac{d\text{if}^{\pi_R}}{d\varepsilon} > 0, \quad \text{and} \quad \frac{d^2\text{if}^{\pi_R}}{d(\bar{w}^{\pi_R})^2} > 0.$$