# Appointment Scheduling with Delay Tolerance Heterogeneity

Shuming Wang,[a,b,*] Jun Li,[c] Marcus Ang,[d] Tsan Sheng Ng[c]

a School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China; ∗ Corresponding author

b MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation, University of Chinese Academy of Sciences, Beijing 100190, China;

c Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576, Singapore;

d Lee Kong Chian School of Business, Singapore Management University, Singapore 178899, Singapore

Published in INFORMS Journal on Computing (2024). DOI: 10.1287/ijoc.2023.0025

**Abstract**: In this study, we investigate an appointment sequencing and scheduling problem with heterogeneous user delay tolerances under service time uncertainty. We aim to capture the delay tolerance effect with heterogeneity, in an operationally effective and computationally tractable fashion, for the appointment scheduling problem. To this end, we first propose a *Tolerance-Aware Delay* (TAD) index that incorporates explicitly the user tolerance information in delay evaluation. We show that the TAD index enjoys decision-theoretical rationale in terms of *Tolerance sensitivity*, *monotonicity*, and *convexity and positive homogeneity*, which enables it to incorporate the frequency and intensity of delays over the tolerance in a coherent manner. Specifically, the convexity of TAD index ensures a tractable modeling of the collective delay dissatisfaction in the appointment scheduling problem. Using the TAD index, we then develop an appointment model with known empirical service time distribution that minimizes the overall tolerance-aware delays of all users. We analyze the impact of delay tolerance on the sequence and schedule decisions and show that the resultant TAD appointment model can be reformulated as a mixed-integer linear program (MILP). Furthermore, we extend the TAD appointment model by considering service time ambiguity. In particular, we encode into the TAD index a moment ambiguity set and a Wasserstein ambiguity set, respectively. The former captures effectively the correlation among service times across positions and user types, whereas the latter captures directly the service time data information. We show that both of the resultant TAD models under ambiguity can be reformulated as polynomial-sized, mixed-integer conic programs (MICPs). Finally, we compare our TAD models with some existing counterpart approaches and the current practice using synthetic data and a case of real hospital data, respectively. Our results demonstrate the effectiveness of the TAD appointment models in capturing the user delay tolerance with heterogeneity and mitigating the worst-case delays.

**Keywords**: joint appointment sequencing and scheduling, delay tolerance, user heterogeneity, linear programming, ambiguity

## 1. Introduction

Managing users' delay or waiting experience is of fundamental importance in many service delivery systems. For waiting experience, users are heterogeneous in terms of not only the characteristics of their service time distributions but also their tolerability to delays. Various research works in the literature suggest the important influence of user delay tolerance on waiting experience. For instance, in a healthcare setting, the level of patients' satisfaction in outpatient care service decreases rapidly once the delay exceeds a given threshold (Huang 1994, McCarthyet al. 2000, Hill and Joonas 2005, Bleustein et al. 2014). The delays may have a bigger impact in hospital emergency

Funding: S. Wang was supported by the National Natural Science Foundation of China [Grants71922020, 72171221, and 71988101, entitled "Econometric Modeling and Economic Policy Studies"], the Fundamental Research Funds for the Central Universities [Grant UCAS-E2ET0808X2], and the Major Program of National Natural Science Foundation of China [Grant 72192843]. S. Wang was also supported by a grant from MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation at UCAS. Supplemental Material: The software that supports the findings of this study is available within the paper and its Supplemental Information (https://pubsonline.informs.org/doi/suppl/10.1287/ijoc.2023.0025) as well as from the IJOC GitHub software repository (https://github.com/INFORMSJoC/2023.0025). The complete IJOC Software and Data Repository is available at https://informsjoc.github.io/.

departments, because exceeding delay tolerances of medical attention can lead to serious consequences (Chan et al. 2005, Robinson and Chen 2011, Nikolova et al. 2016). Some healthcare studies have emphasized the importance of managing the delay tolerance of users with heterogeneity in improving the service outcomes (Cayirli et al. 2008, Sharif et al. 2014). In particular, recent empirical studies have revealed that understanding the patient preference (e.g., delay to care) can actually make the waiting less irritating (Bleustein et al. 2014, Liu et al. 2018). These effects on delay experience are also observable in other industries. For instance, Taylor (1994) showed with an empirical study on airline services that managing customers' perceptions of delay can be more effective in enhancing the service quality than managing only the actual waiting times.

Incorporating the users' delay tolerance into the design of the service system naturally raises two key questions for managers: (1) how to estimate or acquire the users' delay tolerance; and (2) how to design the service system given the acquired delay tolerance information of users? On the first question, there are several methods to do so. A common practice is that companies, hospitals, and/or authorities estimate the users' delay tolerance from the user profile or acquired users' feedback. For instance, in healthcare services, the patient classification system[1] categorizes patients as to how sick they are and alerts the hospital on necessary care needed for them (Malloch and Meisel 2013). The estimation of users' delay tolerances requires different domain-specific skills and is beyond the scope of our current work.

In this paper, we focus on the second question. In particular, we study one aspect of the service system, which is the joint appointment sequencing and scheduling problem. More specifically, we consider a single server and a set of heterogeneous users who seek appointments with the server. The objective is to determine the order (sequence) and the appointment time (schedule) for each user by optimizing the overall user delay experience according to the characteristics of users' delay tolerances and service times.

In the appointment-scheduling literature, most papers optimize the trade-off between the expected (weighted) total waiting time (delay) and some cost-related performance in the appointment system (see, e.g., Gupta and Denton 2008, Kong et al. 2013, Mak et al. 2015, Wang et al. 2019). However, the objective of total expected waiting time may not capture the effects of user delay tolerance with heterogeneity as identified in empirical studies (Hill and Joonas 2005, Bleustein et al. 2014, Liu et al. 2018). Also, the criterion of expected weighted delay, where higher weights are placed on positions whose waiting times are less acceptable, is rather ad hoc and works only when the user sequence is known a priori. Consequently, such criterion does not extend readily to joint sequencing and scheduling problems where the positions assigned to users depend on the sequence (see Remark 3 in Section 4). As for the cost-related performance criterion, it has been witnessed in healthcare studies and practices that financial cost-measurement approaches (e.g., from the perspective of service suppliers) can obscure the value in healthcare and lead to cost containment efforts that are ineffective and sometimes even counterproductive. On the other hand, the challenges of managing the waiting experience are magnified in cases where hospitals attempt to schedule patients' requests for appointments on the day they call. This is referred to as *Advanced Access system* (Murray and Tantau 2000). Implementation of the advanced access system in practice is computationally challenging because the appropriate scheduling of patients to doctors has to be done efficiently and promptly (Kiran and O'Brien 2015). Thus, how to capture the user delay tolerance with heterogeneity in the appointment sequencing and scheduling in an operationally effective and computationally efficient fashion remains a crucial yet challenging research question to be explored.

In this paper, we propose a *Tolerance-Aware Delay* (TAD) index to characterize users' dissatisfaction toward appointment delays by capturing coherently the effects of delay time distribution and delay tolerance with the users' heterogeneity in the appointment-scheduling problem under service time uncertainty. The TAD index incorporates explicitly the delay tolerance information of different users, accounts sensitively for both the frequency and intensity of delays above tolerance, and yields a computationally tractable decision criterion in virtue of its convexity. Using the TAD index to model the tolerance-incorporated delay dissatisfaction of each user, we develop a TAD-based appointment sequencing and scheduling model that minimizes the overall tolerance-aware delay performance of all users in both situations of known service time distribution and ambiguity. We show several operational properties of the TAD model that imply its capability in adapting the appointment decisions to the heterogeneous delay tolerances effectively and derive the convex reformulations of the TAD model that justify its computational tractability. We also demonstrate the effectiveness of the TAD model in our numerical studies and a case study with real outpatient data.

Our work is most related to Qi (2017) in the literature, which proposed a delay unpleasantness measure (DUM) that also considered the users' delay tolerance information for scheduling appointments. Nevertheless, the DUM model in Qi (2017) focused on the unfairness issue and optimized the worst-case tolerance-incorporated delay performance across all users, which is different from our objective of considering the overall tolerance-aware delay performance. Moreover, the criterion of DUM is *nonconvex*, which poses computational

difficulties in optimizing the overall delay performance of all users and, therefore, might not be suitable for situations that require a quick response time, such as the aforementioned advanced access system. A more comprehensive discussion on the research gap is provided in the Literature Review (Section 2). The major contributions of our study can be summarized as follows:

1. We propose a *Tolerance-Aware Delay* (TAD) index for measuring the delay performance in the appointment scheduling problem, which captures the users' delay tolerance with heterogeneity in an operationally effective and computationally tractable fashion. In particular, we show that the TAD index enjoys decision-theoretical rationale in terms of *tolerance sensitivity*, *monotonicity*, *convexity*, and *positive homogeneity*, which ensure its capturing of the frequency and the intensity of the delays over tolerance in a coherent fashion. Computationally, the convexity of the TAD index ensures a tractable modeling[2] for collective delay dissatisfaction (i.e., overall delay performance of all users) in the appointment scheduling problem.
2. We develop a TAD-based appointment sequencing and scheduling model with a known empirical service time distribution that minimizes the total tolerance-aware delay level. We analyze the impacts of user tolerance on the sequence and scheduling decisions with our TAD model and obtain the following insights. (i) For identically distributed service times, it is always optimal to sequence the user with the lowest delay tolerance to the first position. (ii) If the user delay-tolerance at some position is decreased, then the optimal scheduling policy would increase the delay buffering for that position (i.e., increasing the scheduled service time duration of some front users) so as to mitigate the expected delay of the user at that position. Computationally, the TAD model with empirical service time distribution can be transformed into a mixed-integer linear program (MILP).
3. We also extend the proposed TAD appointment model to the situation of service time ambiguity. In particular, we encode into the TAD index a moment ambiguity set and a Wasserstein ambiguity set, respectively. The former captures effectively the correlation among service times across positions and user types, whereas the latter captures directly the service time data information. We show that both of the resultant TAD models under ambiguity can be reformulated as polynomial-sized, mixed-integer conic programs (MICPs).

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature, and research gaps are established. In Section 3, we introduce the Tolerance-Aware Delay (TAD) index and also present its characteristics and decision-theoretical properties. In Section 4, we model the appointment sequencing and scheduling problem using the TAD index with a known empirical service time distribution and discuss the associated operational properties and tractable model reformulation. In Section 5, we extend the TAD model by considering the distributional ambiguity of service times. We present the numerical experiments using synthetic data and a case study with real outpatient data in Section 6, and we conclude our study by including future research directions in Section 7. All proofs are presented in the E-companion.

### Notations

Given $i, N \in \mathbb{N}$ with $i \le N$, we use $[i;N]$ to denote the running indices $\{i, i+1, \ldots, N\}$ and use $[N]$ to denote the complete index set $\{1, 2, \ldots, N\}$. We denote "$\vee$" as the "max" operator and $(r)^+ := r \vee 0$ for any $r \in \mathbb{R}$. We use "$\mathbb{P}$" to represent a specific probability distribution and "$\mathcal{F}$" to represent a set of distributions for modeling the distributional ambiguity, that is, ambiguity set or distributional set. When $\mathcal{F} = \{\mathbb{P}\}$, it reduces to the case of a specific probability distribution $\mathbb{P}$. We use $\mathcal{P}(\Xi)$ to represent the set of all probability distributions $\mathbb{P}$ supported on $\Xi$. We use tilde to denote a random variable, for example, $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^n$, and use $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}$ to denote that the random variable $\tilde{\boldsymbol{\xi}}$ follows probability distribution $\mathbb{P}$. For a random variable $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^n$ with distribution $\mathbb{P}$ and function $\boldsymbol{g} : \mathbb{R}^n \longmapsto \mathbb{R}^m$, we denote $\mathbb{E}_{\mathbb{P}}(\boldsymbol{g}(\tilde{\boldsymbol{\xi}}))$ as the expectation of random variable $\boldsymbol{g}(\tilde{\boldsymbol{\xi}})$ with the probability measure $\mathbb{P}$.

## 2. Literature Review

The literature on appointment scheduling is vast, in which there are generally two approaches to solve the appointment scheduling problem. The first approach solves for the exact appointment time for each user (Chen and Robinson 2014, Samorani and LaGanga 2015, Jiang et al. 2019, Kong et al. 2020, Shehadeh et al. 2021, Homem-de-Mello et al. 2022, Zacharias et al. 2022). The second approach first divides the planning horizon into a number of appointment positions (time slots) and thereafter determines the number of users to be scheduled in each position (Zacharias and Pinedo 2017, Wang et al. 2019, Zacharias and Yunes 2020, Benjaafar et al. 2023). Also, many papers focus on a single server and disregard downstream resources (e.g., downstream recovery units) by noting that sequencing for both upstream (e.g., upstream surgery units) and downstream activities would lead to a class of complex combinatorial and multicriteria stochastic optimization problems (Shehadeh and Padman 2022), which is challenging and usually studied in surgery scheduling with recovery resources (Liu et al. 2019, Bai et al. 2022) that is beyond the scope of our study. Furthermore, most of the studies consider

minimizing the delay, that is, waiting time for users, and idle time and overtime for servers. Only limited papers study the lateness over the random due dates; one of the few examples is Wu and Zhou (2008), in which the authors used a dynamic programming algorithm to solve a single-server scheduling problem with stochastic due dates and processing times for the purpose of minimizing the maximum lateness. In our study, we employ the first approach and consider the case of single-server and delay tolerance heterogeneity. In this section, we review only the streams of studies considering service time uncertainty in appointment scheduling that are closely related to our work, which are summarized in Table 1. We learn from the table that the majority of literature models the (cost for) waiting time, idle time, and overtime as objective while ignoring the delay tolerance heterogeneity. Other streams, such as appointment scheduling with focus on no-shows and/or service interruptions, can be found in reviews from Ahmadi-Javid et al. (2017) and Marynissen and Demeulemeester (2019). For more reviews with multiple servers, see Ahmadi-Javid et al. (2017), Marynissen and Demeulemeester (2019), Kuiper and Lee (2022), and Wu and Zhou (2022) and the references therein.

Uncertainty in service times can have significant impacts on the performance of the appointment scheduling problem. A common approach to model uncertain service times is to describe them using probability distributions. This approach is reasonable if there is sufficient historical data available to calibrate the distributions (see, e.g., Denton et al. 2007, Gupta and Denton 2008, Erdogan and Denton 2013, Qi 2017, Homem-de-Mello et al. 2022). In addition, as we have shown in Table 1, the majority of literature adopting the stochastic programming method also employs the empirical distribution.

In some practical situations, it may be difficult to have precise knowledge of the underlying service time distributions because of the lack of credible historical data. For instance, in healthcare applications, the available data broken down by surgery types and surgeons could be limited (Denton et al. 2007, Macario 2010). Meanwhile, the emergence of new medical technologies could also devalue the assumed probability distribution estimated from the historical healthcare service data. To this end, a stream of appointment scheduling studies has been proposed to employ distributionally robust optimization approaches that harness the limited available distributional information; see the second half of Table 1. The robust appointment scheduling models focus on the worst-case performance outcomes over an *ambiguity set* of all the qualified service time distributions that are partially characterized by, for instance, the moment information (Delage and Ye 2010, Wiesemann et al. 2014) and/or empirical information with statistical metrics, for example, Wasserstein distance (Esfahani and Kuhn 2018, Gao and Kleywegt 2023).

**Table 1.** Related Literature Review in Appointment Scheduling

| References | Objective | Delay tolerance heterogeneity | Uncertainty | Approach | Sevice time ambiguity |
|---|---|---|---|---|---|
| Chen and Robinson (2014) | WT/IT/OT | ✗ | Service time /No-show /No-call | SP | Empirical distribution |
| Samorani and LaGanga (2015) | WT/OT/no. of users | ✗ | No-show | SP | ✗ |
| Shehadeh et al. (2021) | Cost for WT, IT, and OT | ✗ | Service time /Arrival time | SP | Empirical distribution |
| Homem-de-Mello et al. (2022) | Cost for WT, IT, and OT | ✗ | Service time /No-show | SP | Empirical distribution |
| Zacharias et al. (2022) | Cost for delay | ✗ | No. of arrivals | MDP | ✗ |
| Wang et al. (2019) | Cost for WT, IT and OT | ✗ | Walk-in /No-show | SP | ✗ |
| Erdogan and Denton (2013) | Cost for WT and OT | ✗ | Service time /No. of users /No-show | SP | Empirical distribution |
| Mak et al. (2015) | WT/OT | ✗ | Service time | DRO | Moment |
| Zhang et al. (2017) | Cost for WT | ✗ | Service time | DRO | Moment |
| Jiang et al. (2017) | Cost for WT, IT, and OT | ✗ | Service time /No-show | DRO | Moment |
| Kong et al. (2020) | Cost for WT, IT, and OT | ✗ | Service time /No-show | DRO | Moment |
| Kong et al. (2013) | WT/OT | ✗ | Service time | DRO | Moment |
| Qi (2017) | Delay unpleasantness measure | ✓ | Service time | DRO | Moment |
| Jiang et al. (2019) | Cost for WT, IT, and OT | ✗ | Service time /No-show | DRO | Wasserstein |
| **Our work** | Tolerance-aware delay | ✓ | Service time | DRO | Moment/Wasserstein |

*Note.* WT, Waiting time; IT, idle time; OT, overtime; SP, stochastic programming; MDP, Markov decision process; DRO, distributionally robust optimization.

More specifically, Mak et al. (2015) utilized the information of mean and second marginal moments to partially characterize the service time distribution, and the resulting appointment scheduling problem is formulated as a second-order cone program (SOCP). Kong et al. (2013) and Zhang et al. (2017) utilized directly the mean and covariance matrix to characterize the service time correlation in the appointment scheduling. The former includes the nonnegative support information, and the resulting problem is transformed to a copositive program (COP), which, however, is not polynomial time solvable in general and is therefore approximated by a semidefinite program (SDP) relaxation, whereas the latter relaxes the nonnegative support requirement and derives an SDP formulation to approximately solve the (original) problem. Kong et al. (2013) also extended their model to include sequence decisions, which leads to a mixed-integer COP. Qi (2017) studied the unfairness and user tolerance to delays in the appointment sequencing and scheduling under service time uncertainty. In particular, the author proposed a delay unpleasantness measure (DUM) based on conditional value at risk (CVaR) and optimized the worst-case DUM-based delay performance across all users. Qi (2017) also considered the problem with distributional ambiguity and used a *mean absolute deviation* approach to model the service time correlations across positions. Jiang et al. (2017) considered a robust appointment scheduling problem with a mean-support ambiguity set for both random service times and no-shows and derived exact mixed-integer nonlinear programming reformulations for the problem that are solved by a decomposition algorithm. Jiang et al. (2019) considered a data-driven robust appointment scheduling problem that optimized the worst-case expected total operational costs (i.e., waiting, idle, and overtime work) over a Wasserstein ball ambiguity set. They derived a COP reformulation in the general case. In particular, for the cases of one-norm- and two-norm-based Wasserstein balls, they showed that the reformulation reduces tractably to an LP and an SOCP, respectively. More recently, Kong et al. (2020) studied a medical appointment scheduling problem with random service times and time-dependent patient no-show behavior. They deployed a distributionally robust model, which minimizes the worst-case total expected costs of patient waiting and a service provider's idling and overtime. Their model reformulates into a COP or a bilinear COP with time-independent or time-dependent no-shows, respectively, and is approximated by the SDP relaxations.

Among the above related studies, Kong et al. (2013), Qi (2017), and Jiang et al. (2019) are most relevant to our study in the directions of user delay tolerance effect (Qi 2017), joint sequence and schedule optimization (Kong et al. 2013, Qi 2017), and service time ambiguity modeled with Wasserstein ball (Jiang et al. 2019). The research gap between our work and these studies can be established as follows:

(i) We focus on capturing the users' delay tolerance with heterogeneity in the appointment sequencing and scheduling operations, for which we propose a Tolerance-Aware Delay (TAD) index for incorporating the delay tolerance effect in delay performance evaluation. Our resultant TAD appointment models optimize the overall tolerance-aware delay performance under both known service time distribution and ambiguity. These are therefore distinct from Kong et al. (2013) and Jiang et al. (2019) because they aim to optimize the expected total actual delay cost or operational costs. Furthermore, our TAD appointment models under ambiguity are formed by encoding the moment ambiguity set or Wasserstein ambiguity set into the TAD index, which by design have a different modeling structure from their ambiguity models.

(ii) The DUM appointment models by Qi (2017) also considered the user delay tolerance effect in appointment sequencing and scheduling with service time uncertainty. Nevertheless, Qi (2017) focused on the unfairness issue and optimized the worst-case tolerance-incorporated delay performance across all users, whereas we focus on the delay tolerance heterogeneity and optimize the total tolerance-aware delay performance of all users that accounts more sensitively for the tolerance heterogeneity (this is also justified by our experiments in Section 6.2). Moreover, our proposed TAD index by design is a *convex* decision criterion, which, therefore, is more computationally suitable for total delay performance optimization. Finally, we analyze the impact of delay tolerance with heterogeneity on the appointment decisions and also develop the TAD appointment sequencing and scheduling model over Wasserstein ambiguity set. These have not been considered in Qi (2017).

## 3. Tolerance-Aware Delay Index: A Convex Decision Criterion

To capture the delay tolerance effect of the users' dissatisfaction in the appointment scheduling problem, we propose a Tolerance-Aware Delay (TAD) index that enjoys several important appealing features operationally and computationally. In this section, we present the definition of the TAD index and discuss its properties.

**Definition 1 (Tolerance-Aware Delay Index).** Let $\mathcal{W}$ be the space of all the real-valued random variables, $\tau$ the delay tolerance level, and $\tilde{w}$ the delay random variable whose distribution $\mathbb{P}$ can be taken from a set $\mathcal{F}$ of

distributions. The *Tolerance-Aware Delay* (TAD) index $\Pi_\tau(\tilde{w}) : \mathcal{W} \mapsto \mathbb{R}$ is defined as

$$\Pi_\tau(\tilde{w}) := \min_{\alpha\in[0,\tau]} \left\{ \tau - \alpha \;\middle|\; \sup_{\mathbb{P}\in\mathcal{F}} \mathrm{CE}^{\mathbb{P}}(\tilde{w},\alpha) \le \tau \right\}, \tag{1}$$

where $\mathrm{CE}^{\mathbb{P}}(\tilde{w},\alpha)$ is a *certainty equivalent*[3] (CE) function given by

$$\mathrm{CE}^{\mathbb{P}}(\tilde{w},\alpha) := \alpha + \mathbb{E}_{\mathbb{P}}[(\tilde{w}-\alpha)^+], \tag{2}$$

and $\Pi_\tau(\tilde{w}) := \infty$ if no feasible $\alpha$ can be found.

In the above definition, the certainty equivalent $\mathrm{CE}^{\mathbb{P}}(\tilde{w},\alpha)$ can be regarded as the expected delay time being deteriorated or penalized with a parameter $\alpha$; we have $\mathrm{CE}^{\mathbb{P}}(\tilde{w},0) = \mathbb{E}_{\mathbb{P}}[\tilde{w}]$, and in general we have $\mathrm{CE}^{\mathbb{P}}(\tilde{w},\alpha) \ge \mathbb{E}_{\mathbb{P}}[\tilde{w}]$, $\forall\alpha\in[0,\tau]$. Here, the parameter $\alpha$ can be regarded as a penalty level so that the penalized expected delay time of the user is below the tolerance level $\tau$. Because $\mathrm{CE}^{\mathbb{P}}(\tilde{w},\alpha)$ is nondecreasing in $\alpha$, a higher value of $\alpha$ corresponds to a higher penalty level and a worse penalized expected delay time. In this sense, the TAD index essentially searches for the highest penalty level $\alpha^\star\in[0,\tau]$ that is attainable to satisfy the tolerance constraint in (1), and the higher such (maximized) attainable penalty level $\alpha^\star$, the more satisfactory the user is regarding the predescribed tolerance level. Accordingly, the TAD index $\Pi_\tau(\tilde{w}) = \tau - \alpha^\star$ measures the difference between the upper limit of penalty level and the maximized attainable penalty level, which characterizes the *dissatisfaction* level of the user's delay experience with respect to the delay tolerance.

To further understand the TAD index, let us consider a user whose delay tolerance level is $\tau$ and is served with a random delay $\tilde{w}\sim\mathbb{P}$ with $\mathcal{F} = \{\mathbb{P}\}$. If the user is not particularly delay sensitive, then the delay tolerance $\tau$ is relatively large, which would be toward the right tail of $\tilde{w}$. In this case, there is very little chance that $\tilde{w}$ would exceed $\tau$. As such, the penalty level $\alpha$ can be close to $\tau$, and the resultant value of the TAD index is low (low dissatisfaction level). On the other hand, if the user is very delay sensitive, then $\tau$ is relatively small and would be on the left tail of $\tilde{w}$ (i.e., a high chance for $\tilde{w}$ exceeding $\tau$). In this case, $\alpha$ would need to be small so that CE is less than $\tau$, and the resultant value of the TAD index would be high (high dissatisfaction level). It is also possible that there is no such $\alpha$, that is, $\Pi_\tau(\tilde{w}) = \infty$. From this description, we see that the TAD index measures the delay performance via a combination of the likelihood that a user's tolerance is met and the degree to which the expected delay time can be penalized so that the tolerance is still met. In fact, the TAD index is indeed appealing in capturing both the intensity and frequency of delays over tolerance. This will be explicitly reflected by Proposition 2 with Example 1.

In addition, we also point out that the above idea of "the largest penalty imposed to the expected delay time to make the tolerance met" in the TAD index is also analogous to that in the dual representation of a convex risk measure with "acceptable sets"; a convex risk measure can be represented as the "minimal amount of cash" that is added to the position to make it acceptable (Föllmer and Schied 2002). A more detailed discussion of the TAD index in connection to convex risk measure will be provided in Section 3.1.

Next, we present the TAD index under a given distribution of $\tilde{w}$. More specifically, when $\tilde{w}$ follows a normal distribution $\mathbf{N}(\mu,\sigma^2)$ with expected value $\mu < \tau$, by definition we can derive $\Pi_\tau(\tilde{w}) = \min_{\alpha\le\tau}\{\tau-\alpha \mid \mu+\Lambda(\alpha)\sigma \le \tau\}$, where $\Lambda(\alpha)$ is a function of $\alpha$ given by

$$\Lambda(\alpha) := \phi\Big(\frac{\alpha-\mu}{\sigma}\Big) + \Big[\frac{\alpha-\mu}{\sigma}\Big]\Phi\Big(\frac{\alpha-\mu}{\sigma}\Big), \tag{3}$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ being the density function and cumulative distribution function of the standard normal distribution, respectively. Furthermore, $\Lambda(\alpha)$ is an increasing function of $\alpha$ by noting that the derivative $\frac{\mathrm{d}\Lambda(\alpha)}{\mathrm{d}\alpha} = \Phi\big(\frac{\alpha-\mu}{\sigma}\big)/\sigma > 0$, and this implies that the TAD index $\Pi_\tau(\tilde{w})$ can be derived via solving a convex optimization problem and is thus computationally suitable. In particular, it admits a closed-form representation, which is formally stated as follows.

**Proposition 1** (TAD Index of Normal Delay). *If $\tilde{w}\sim\mathbf{N}(\mu,\sigma^2)$ with $\mu<\tau$, then*

$$\Pi_\tau(\tilde{w}) = \tau - \Lambda^{-1}\Big(\frac{\tau-\mu}{\sigma}\Big), \tag{4}$$

*where $\Lambda(\cdot)$ is the Gaussian density-and-distribution function given in (3).*

Proposition 1 reveals that the TAD index of a normal delay distribution is exactly the difference between the delay tolerance and the inverse Gaussian density-and-distribution function of the standardized expected-delay-under-tolerance, that is, $\frac{\tau-\mu}{\sigma}$.

**Remark 1.** As we have emphasized in Literature Review (Section 2), our proposed TAD index shares several properties with the delay unpleasantness measure (DUM) in Qi (2017), including in capturing the delay tolerance effect, which is most relevant to our TAD index in the appointment scheduling literature. Leveraging the closed-form expression presented in Proposition 1, we can illustrate the difference clearly. Consider the random delay $\tilde{w}$ that follows a normal distribution $\mathbf{N}(\mu, \sigma^2)$ with $\mu < \tau$, and it is not difficult to show that the DUM becomes

$$\rho_\tau^{\mathrm{DUM}}(\tilde{w}) = \inf_{\alpha \geq 0} \left\{ \alpha \,\middle|\, \frac{\phi\left(\Phi^{-1}(1-\alpha)\right)}{\alpha} \leq \frac{\tau - \mu}{\sigma} \right\},$$

which is distinct from our TAD index as in (4). Also, it can be seen that both the TAD and DUM utilize the information of standardized expected-delay-under-tolerance ($\frac{\tau-\mu}{\sigma}$), and our closed-form expression (4) implies its computational convenience.

Next, when delay follows an empirical distribution, we can also have the following formula for the TAD index to be readily evaluated.

**Proposition 2** (TAD Index of Empirical Delay). *Let delay $\tilde{w} \sim \hat{\mathbb{P}} := \sum_{k\in[K]} \delta_{w_k} p_k$, where $\delta$ is the Dirac delta function, and assume w.l.o.g. $w_1 < w_2 < \cdots < w_K$. If $0 < \Pi_\tau(\tilde{w}) < \infty$, then*

$$\Pi_\tau(\tilde{w}) = \sum_{k=k^\star+1}^{K} (w_k - \tau) \left[ \frac{p_k}{\sum_{l\in[k^\star]} p_l} \right], \tag{5}$$

*where*

$$k^\star := \max \left\{ l \in [K] \,\middle|\, w_l < \tau, w_l + \sum_{k=l+1}^{K} (w_k - w_l) p_k \leq \tau \right\}. \tag{6}$$

We emphasize that Equation (5) also justifies explicitly the appealing structure of the TAD index that captures both the intensity and frequency of delays over tolerance. This can be further illustrated by comparing the TAD index with the probability measure of delay over tolerance as follows.

**Example 1** (Probability of Intolerable Delays). Consider the following two uncertain service delays, $\tilde{w}_A$ and $\tilde{w}_B$,

$$\tilde{w}_A = \begin{cases} 5 \text{ minutes}, & \text{w.p. } 0.79 \\ 15 \text{ minutes}, & \text{w.p. } 0.21, \end{cases} \quad \tilde{w}_B = \begin{cases} 5 \text{ minutes}, & \text{w.p. } 0.80 \\ 25 \text{ minutes}, & \text{w.p. } 0.20., \end{cases}$$

where w.p. refers to with probability. Suppose that a user has a tolerance level of $\tau = 10$ minutes. The probability $\mathbb{P}[\tilde{w} > \tau]$ is a natural candidate to quantify the delay dissatisfaction experienced. Under this criterion, $\tilde{w}_B$ is preferable because $\mathbb{P}[\tilde{w}_A > 10] = 0.21 > 0.20 = \mathbb{P}[\tilde{w}_B > 10]$. However, the probability criterion focuses only on the delay probability and cannot reflect the intensity of "bad" delays (i.e., the delays that exceed $\tau = 10$). Note that $\tilde{w}_B$, albeit with a slightly better probability, is actually much worse in delay intensity. In contrast, the TAD index is able to capture this effect by noting that $\Pi_{10}(\tilde{w}_A) = \frac{105}{79} < \frac{15}{4} = \Pi_{10}(\tilde{w}_B)$.

### 3.1. Key Properties of TAD Index

In this subsection, we explore several key properties of the TAD index, which shows its potential as a coherent decision criterion for evaluating the tolerance-aware dissatisfaction of users toward uncertain delays.

**Theorem 1** (Decision-Theoretical Properties of the TAD Index). *The TAD index satisfies the properties as follows:*

1. *Tolerance Sensitivity:*
   (a) *Full satisfaction: If $\mathbb{P}\{\tilde{w} \leq \tau\} = 1$, $\forall \mathbb{P} \in \mathcal{F}$, then $\Pi_\tau(\tilde{w}) = 0$.*
   (b) *Abandonment: If $\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[\tilde{w}] > \tau$, then $\Pi_\tau(\tilde{w}) = \infty$.*
   (c) *Delay-tolerance translation: $\Pi_\tau(\tilde{w} + \eta) = \Pi_{\tau-\eta}(\tilde{w})$, $\forall \eta \in [-\underline{\omega}, \tau]$ with $\mathbb{P}[\tilde{w} \geq \underline{\omega}] = 1$.*
2. *Monotonicity:*
   (a) *$\Pi_{\tau_1}(\tilde{w}) \leq \Pi_{\tau_2}(\tilde{w})$ for any delay tolerance thresholds $\tau_1, \tau_2$ with $\tau_1 \geq \tau_2$.*
   (b) *$\Pi_\tau(\tilde{w}_1) \leq \Pi_\tau(\tilde{w}_2)$ for any delays $\tilde{w}_1, \tilde{w}_2$ with $\tilde{w}_1 \leq \tilde{w}_2$.*
3. *Convexity: $\Pi_{\lambda\tau_1+(1-\lambda)\tau_2}(\lambda\tilde{w}_1 + (1-\lambda)\tilde{w}_2) \leq \lambda\Pi_{\tau_1}(\tilde{w}_1) + (1-\lambda)\Pi_{\tau_2}(\tilde{w}_2)$, $\forall \lambda \in [0,1]$.*
4. *Positive homogeneity: $\Pi_{\lambda\tau}(\lambda\tilde{w}) = \lambda\Pi_\tau(\tilde{w})$, $\forall \lambda > 0$.*

In Theorem 1, the properties under *tolerance sensitivity* emphasize the ability of the TAD index in capturing the tolerance awareness. Part 1(a) implies that the delays that are almost surely below the tolerance threshold $\tau$, are always most preferred or satisfactory under the TAD ($\Pi_\tau(\tilde{w}) = 0$). Part 1(b) states that delays $\tilde{w}$ that exceed $\tau$ in expectation are least preferred or intolerable ($\Pi_\tau(\tilde{w}) = \infty$). In other words, if a design cannot even meet the appointment delay tolerance requirement by users in expectation, it will not be considered as a reasonable option for the optimal design. This is important in ensuring good service quality, for example, in the time-critical context of healthcare. We highlight that Part 1(b) also serves to mitigate the unfairness in evaluating total delay dissatisfaction, which helps achieve a good trade-off between mean and worst-case delay performance in collective delay dissatisfaction (Section 3.2). This is also justified by our computational studies (Sections 6.1 and 6.4). Part 1(c) states that the dissatisfaction associated with a delay $\tilde{w}$ reduced (increased) by a constant amount is equivalent to that associated with $\tilde{w}$ by a user with tolerance increased (reduced) by that same constant amount. This property reveals a critical feature of the TAD index on how the user's delay dissatisfaction hinges on the interactions between the actual delay experienced and the delay-tolerance.

Part 2(a) states that users with lower tolerance are always no less dissatisfied than those with higher tolerance for any given delay $\tilde{w}$. Part 2(b) states that if the delay $\tilde{w}_1$ is always no longer than another $\tilde{w}_2$, then the latter is never more satisfactory for any given $\tau$.

Part 3 shows that the TAD index is a *convex function* jointly in $(\tilde{w}, \tau) \in \mathcal{W} \times \mathbb{R}$. The convexity ensures a computationally appealing structure of our TAD-based appointment optimization models that minimizes the total delay dissatisfaction (see Proposition 3 in Section 4 and Proposition 9 in Section 5) where the delay $\tilde{w}$ and tolerance level $\tau$ are functions of decisions.

Part 4 implies that scaling the delay tolerance excess $(\tilde{w} - \tau)$ by a positive multiplier results in the dissatisfaction level being scaled by the same multiplier. Note that $\Pi_\tau(\tilde{w}) = \Pi_0(\tilde{w} - \tau)$ by delay tolerance translation in Part 1(c).

We also point out that our proposed TAD index has a decision-theoretical interpretation, which can be well reflected through its connection with the well-known *coherent risk measures* (Artzner et al. 1999) in the following remark.

**Remark 2** (Connection to Coherent Risk Measures). The properties of the TAD index presented in Theorem 1 share some similarities with the well-known concept of *coherent risk measure* (Artzner et al. 1999) that is popular in risk management. It is known that a (monetary) risk measure is coherent if it has the properties of *translation invariance*, *monotonicity*, *convexity*, and *positive homogeneity* (Föllmer and Schied 2002). Indeed, the TAD index in Theorem 1 also enjoys the properties of monotonicity, convexity, and positive homogeneity in a more general manner by incorporating the effects of the delay tolerance.[4] Moreover, the TAD index strengthens the property of translation invariance into the tolerance sensitivity (Part 1, (a)−(c), of Theorem 1) so as to stress the tolerance awareness in the delay evaluation.

### 3.2. Modeling Collective Delay Dissatisfaction with the TAD Index

Suppose we have multiple users $i \in [I]$, where each user $i$ has a delay tolerance $\tau_i$ and experiences the delay $\tilde{w}_i$. The convexity is critical for measuring the total (collective) delay dissatisfaction in a tractable fashion, which is preserved by the proposed TAD index (Theorem 1, Part 3). In particular, it is natural to consider the total TAD level across all users, that is, $\sum_{i\in[I]}\Pi_{\tau_i}(\tilde{w}_i)$. Alternatively, the convexity also allows us to consider the other variants of the collective delay dissatisfaction with the TAD index, for example, $\max_{i\in[I]}\Pi_{\tau_i}(\tilde{w}_i)$, which is also tractable. In this study, we focus on the total TAD level across all users as the collective delay dissatisfaction criterion for our appointment scheduling problem because this criterion accounts more sensitively for the delay tolerance heterogeneity of different users (as justified by our experiments in Section 6.2). We will show that both the TAD-based appointment models under empirical distribution and distributional ambiguity result in computationally appealing reformulations in Sections 4 and 5, respectively.

## 4. TAD-Based Appointment Sequencing and Scheduling

In this section, we introduce the appointment scheduling problem with multiple user tolerance levels under service time uncertainty. We then model the problem using the TAD index, which minimizes the total TAD level. In particular, we focus on the case of known empirical service time distribution for the computational reformulation of the model in Section 4.2. The computational reformulation with service time ambiguity is discussed in Section 5.

### 4.1. Problem Description

We consider a problem of joint appointment sequencing and scheduling with multiple types of users. Let $I$ denote the number of the available appointment positions and $i$ denote the index of positions, with $i \in [I]$, in order of the sequence to be served. Let $J$ be the total number of the *user types* or *categories*, with $j$ indexing the user category, $j \in [J]$, and $N_j$ the number of category $j$ users, where $\sum_{j\in[J]} N_j = I$. The total session length of the planning horizon is fixed and is denoted by $L$. Denote $r_j$ as the delay tolerance of users in category $j$ and $\tau_i$ the delay tolerance of the user sequenced in position $i$. The latter is now a function of the individual tolerances ($r_j's$) and the sequencing, which will be specified later in (9). We denote the service time of user category $j$ sequenced in position $i$ as a random variable $\tilde{\xi}_{ij}$. Given a list of users known by category (and hence, tolerance levels $r_j$), the manager assigns a position to each user in the appointment sequence and schedules the corresponding appointment time for each position.

Let $x_{ij} \in \{0,1\}$ denote the *sequencing decision*, where $x_{ij} = 1$ if position $i$ is assigned to a user of category $j$, and $x_{ij} = 0$ otherwise, for $i \in [I], j \in [J]$. Let $y_i \in \mathbb{R}_+$ denote the *scheduling decision*, that is, the appointment time for the $i$th position, for $i \in [I]$. The feasible set for $\boldsymbol{x}$ is defined as

$$\mathcal{X} := \left\{ \boldsymbol{x} \in \{0,1\}^{I\times J} \;\middle|\; \sum_{j\in[J]} x_{ij} = 1,\ \forall i \in [I]; \sum_{i\in[I]} x_{ij} = N_j,\ \forall j \in [J] \right\}, \tag{7}$$

where the equality constraints ensure that each position is filled with one user, and all users are allocated. The feasible set for the appointment times $\boldsymbol{y}$ is defined as

$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}_+^I \,|\, y_1 = 0, y_{i-1} \le y_i,\ \forall i \in [2;I], y_I \le L\}, \tag{8}$$

which requires that users are served in chronological order according to the sequence indexed in $i \in [I]$. Note that service staff overtime is permitted, but the last user must be scheduled before time $L$ so that staff working hours are well utilized (at least according to plan).

Because of the user heterogeneity and the sequencing decision $\boldsymbol{x}$, the resulting service time for each position $i$ depends on the type of user assigned to that position and is given by

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\xi}}_i = \sum_{j\in[J]} \tilde{\xi}_{ij} x_{ij},$$

with $\tilde{\boldsymbol{\xi}}_i := (\tilde{\xi}_{i1}, \tilde{\xi}_{i2}, \ldots, \tilde{\xi}_{iJ})^\top$ and $\boldsymbol{x}_i := (x_{i1}, x_{i2}, \ldots, x_{iJ})^\top$. Similarly, denote that $\boldsymbol{r} := (r_1, r_2, \ldots, r_J)^\top$, the delay tolerance threshold $\tau_i(\boldsymbol{x})$ for each position $i$, which also depends on the user assignment, is

$$\tau_i(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}_i = \sum_{j\in[J]} r_j x_{ij}. \tag{9}$$

Let $\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y})$ denote the appointment delay for the user in position $i$. For the first position, it is trivial that the delay $\tilde{w}_1(\boldsymbol{x}, \boldsymbol{y}) \equiv 0$. For the other positions, the waiting times are given by

$$\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y}) = \max\{0, y_{i-1} + \tilde{w}_{i-1}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{x}_{i-1}^\top \tilde{\boldsymbol{\xi}}_{i-1} - y_i\}, \quad \forall i \in [2;I].$$

The above recursion form can be equivalently expressed in the following format:

$$\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y}) = \max_{t\in[i-1]} \left\{ \sum_{l=t}^{i-1} \boldsymbol{x}_l^\top \tilde{\boldsymbol{\xi}}_l - (y_i - y_t) \right\}^+, \quad \forall i \in [2;I]. \tag{10}$$

Given $(\boldsymbol{x}, \boldsymbol{y})$, the TAD index at position $i$, with tolerance level $\tau_i(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}_i$, is denoted as $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y})]$. Hence, the appointment scheduling problem that minimizes the collective delay dissatisfaction can be expressed as

$$\min_{\boldsymbol{x},\boldsymbol{y}} \left\{ \sum_{i\in[2;I]} \Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y})] \;\middle|\; \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y} \right\}. \tag{11}$$

As for the total TAD level minimization in our TAD appointment model (11), we point out that by the full satisfaction property (Part 1(a) of Theorem 1), if the delay of each user is below his or her tolerance almost surely, then our TAD appointment model also achieves its lowest value, that is, $\sum_{i\in[2;I]} \Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y})] = 0$. Furthermore, the abandonment property (Part 1(b) of Theorem 1) ensures that if the delay of even one user exceeds his or her tolerance in expectation (for some $\mathbb{P}$ in $\mathcal{F}$), then $\sum_{i\in[2;I]} \Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x}, \boldsymbol{y})] = \infty$. This enables our TAD appointment

model to reject the extremely bad and/or unfair appointment solutions when evaluating the total delay dissatisfaction across different users, which is also justified in our computational studies (Section 6.1).

Moreover, we also remark that the above TAD model (11) can be readily extended to handle the server overtime and server idle time. In particular, we can introduce an additional delay as $\tilde{w}_{I+1}(\boldsymbol{x},\boldsymbol{y}) = \max\{0, y_I + \tilde{w}_I(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{x}_I^\top \tilde{\boldsymbol{\xi}}_I - L\}$ to model the server overtime (spent by the doctor). Hence, given an overtime tolerance $\tau_{I+1}$, we can incorporate the additional $\Pi_{\tau_{I+1}}[\tilde{w}_{I+1}(\boldsymbol{x},\boldsymbol{y})]$ into (11) to minimize jointly delay and overtime dissatisfaction. Similarly, note that the server idle time is just $L + \tilde{w}_{I+1}(\boldsymbol{x},\boldsymbol{y}) - \sum_{i\in[I]} \boldsymbol{x}_i^\top \tilde{\boldsymbol{\xi}}_i$. Thus, given an idle time tolerance $\tau_{\text{idle}}$, we can also incorporate server idle time in the objective function. The corresponding TAD models incorporated with overtime and idle time are provided in E-companion EC D.

The following proposition shows the convexity of TAD objective function in (11).

**Proposition 3.** *The objective function $\sum_{i\in[2;I]} \Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})]$ of the TAD appointment model (11) is a convex function with respect to $(\boldsymbol{x},\boldsymbol{y})$.*

Proposition 3 implies that the TAD appointment model can yield in computationally attractive formulations. Specifically, we show later that it leads to a mixed-integer linear program for the case under known distribution (Section 4.2) and a mixed-integer second-order conic program for the case under distributional ambiguity (Section 5).

By definition (1) and that $\tau_i(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}_i$, we can rewrite the TAD model (11) explicitly as

$$
\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{y}} \ & \sum_{i\in[2;I]} (\boldsymbol{r}^\top \boldsymbol{x}_i - \alpha_i) \\
\text{s.t.} \ & \alpha_i + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[(\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}) - \alpha_i)^+] \le \boldsymbol{r}^\top \boldsymbol{x}_i, && \forall i \in [2;I] \\
& \tilde{w}_i(\boldsymbol{x},\boldsymbol{y}) = \max_{t\in[i-1]} \left\{ \sum_{l=t}^{i-1} \boldsymbol{x}_l^\top \tilde{\boldsymbol{\xi}}_l - (y_i - y_t) \right\}^+, && \forall i \in [2;I] \\
& 0 \le \alpha_i \le \boldsymbol{r}^\top \boldsymbol{x}_i, && \forall i \in [2;I] \\
& \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}, && (12)
\end{aligned}
$$

where $\boldsymbol{x},\boldsymbol{y}$ are decision variables, and $\mathcal{X}$ and $\mathcal{Y}$ are feasible sets defined by (7) and (8), respectively, and each $\alpha_i$ is an auxiliary variable used to determine the level of TAD index at position $i$.

**Remark 3** (On the Sequence Dependency of Delay Tolerance). Model (12) shows that the TAD model captures, for each position $i$, the interdependency between the sequencing $\boldsymbol{x}$, scheduling $\boldsymbol{y}$, service time distribution $\tilde{\boldsymbol{\xi}}$, and (sequence-dependent) delay tolerance $\tau_i(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}_i$ simultaneously. It is noteworthy to mention that some appointment studies (see, e.g., Kong et al. 2013, Mak et al. 2015, Kong et al. 2020) have employed the expected delay model with "expected weighted waiting time and overtime," where higher weights are placed on positions whose delays are less acceptable. However, such an approach might not be suitable for our joint appointment sequencing and scheduling problem by recalling that the delay tolerance

$$\tau_i(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}_i = \sum_{j\in[J]} r_j x_{ij}$$

in each position $i$ now depends on the user type assigned (i.e., the sequence decision $\boldsymbol{x}$).

**Remark 4** (Delay Tolerance Regularization). In practical applications, if the tolerance levels of users are too low (e.g., when certain users are not comfortable for long waiting) or their delay times are too long (e.g., when the server is over-utilized), it is possible that the problem (12) becomes infeasible. To deal with this issue, we propose a practical regularization approach that is commonly used in statistics/machine learning (Xu et al. 2009, Sugiyama 2015, Bertsimas et al. 2019) to ensure the feasibility of the TAD appointment scheduling model. Specifically, we increase (relax) the original delay tolerance levels $\boldsymbol{r}$ in the TAD model to $\theta\boldsymbol{r}, \theta \ge 1$, where $\theta$ is a relaxation factor. We then solve the following regularized TAD model,

$$\min_{\boldsymbol{x},\boldsymbol{y},\theta} \left\{ \sum_{i\in[2;I]} \Pi_{\tau_i^\theta(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] + \Omega\theta \,\middle|\, \theta \ge 1, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y} \right\}, \tag{13}$$

where $\tau_i^\theta(\boldsymbol{x}) := (\theta\boldsymbol{r})^\top \boldsymbol{x}_i$, and $\Omega\theta$ is the regularization term with $\Omega$ being a given sufficiently large number. The regularized TAD model (13) is always feasible and can be regarded as a reasonable approximation to the

problem (12) in the following sense. Because of the sufficiently large penalty $\Omega$, if the problem (12) is feasible, then the solution $\theta^\star$ is equal (or very close) to 1. As such, we do not lose (much) in optimality. If the problem (12) is infeasible, the model returns an optimal appointment solution $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ with the tightest relaxation $\theta^\star$ for relaxing the delay tolerance levels, and the resulting tolerance levels preserve the original ranking and proportional relation.

It is noted that the regularized TAD model (13) is also computationally attractive. In fact, the bilinear terms $\theta \boldsymbol{x}_i$ in the relaxed tolerance level $\tau_i^\theta(\boldsymbol{x}) = \boldsymbol{r}^\top(\theta \boldsymbol{x}_i)$ can be readily linearized without increasing the number of integers. Specifically, the regularized TAD model (13) is equivalent to the MIP formulation

$$
\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{y}} \quad & \sum_{i\in[2;I]} (\boldsymbol{r}^\top \boldsymbol{z}_i - \alpha_i) + \Omega\theta \\
\text{s.t.} \quad & \alpha_i + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[(\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}) - \alpha_i)^+] \le \boldsymbol{r}^\top \boldsymbol{z}_i, && \forall i \in [2;I] \\
& \tilde{w}_i(\boldsymbol{x},\boldsymbol{y}) = \max_{t\in[i-1]} \left\{ \sum_{l=t}^{i-1} \tilde{\boldsymbol{\xi}}_l^\top \boldsymbol{x}_l - (y_i - y_t) \right\}^+, && \forall i \in [2;I] \\
& z_{ij} \le \theta + (1 - x_{ij})\overline{\theta}, && \forall i \in [2;I], j \in [J] \\
& z_{ij} \le \overline{\theta} x_{ij}, && \forall i \in [2;I], j \in [J] \\
& 0 \le \alpha_i \le \boldsymbol{r}^\top \boldsymbol{z}_i, && \forall i \in [2;I] \\
& \theta \ge 1, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y},
\end{aligned}
\tag{14}
$$

where $\boldsymbol{\alpha}, \boldsymbol{z}$ and $\theta$ are auxiliary variables, and $\overline{\theta}$ can be any upper bound of $\theta$. Clearly, model (14) preserves the structure of the TAD model (12), with the sequence decision $\boldsymbol{x}$ being the only binary variable.

### 4.2. Empirical TAD Model, Reformulation, and Properties

We focus on the appointment scheduling problem given an empirical service time distribution, that is, $\mathcal{F} = \{\hat{\mathbb{P}}\}$, where we assume that user service times $\tilde{\boldsymbol{\xi}}$ are identified as scenarios $\boldsymbol{\xi}^k = (\xi_{ij}^k)_{I\times J}$ with probabilities $p_k$, $\forall k \in [K]$, which yield the service time distribution as

$$
\hat{\mathbb{P}} := \sum_{k\in[K]} p_k \delta_{\boldsymbol{\xi}^k}, \tag{15}
$$

where $\delta$ is the *Dirac delta* function.

Based on the empirical service time distribution given in (15), given appointment decision $(\boldsymbol{x},\boldsymbol{y})$, we can express the delay scenarios $k$ for each position $i$ as follows:

$$
w_i^k(\boldsymbol{x},\boldsymbol{y}) = \left( \max_{t\in[i-1]} \left( \sum_{l=t}^{i-1} \sum_{j\in[J]} \xi_{lj}^k x_{lj} - y_i + y_t \right) \right)^+, \quad \forall i \in [2;I], k \in [K]. \tag{16}
$$

Accordingly, the TAD index $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})]$ at each position $i$ given $(\boldsymbol{x},\boldsymbol{y})$ can be written as

$$
\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] = \min_{0\le\alpha\le\tau_i(\boldsymbol{x})} \left\{ \boldsymbol{r}^\top \boldsymbol{x}_i - \alpha \,\middle|\, \alpha + \sum_{k\in[K]} p_k \left( w_i^k(\boldsymbol{x},\boldsymbol{y}) - \alpha \right)^+ \le \boldsymbol{r}^\top \boldsymbol{x}_i \right\}.
$$

Furthermore, under the empirical service time distribution, we are able to represent the TAD level $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})]$ of each position $i$ in a closed-form expression. To this end, given decision $(\boldsymbol{x},\boldsymbol{y})$, we let w.l.o.g. the delay realizations in (16) for each position $i \in [2;I]$ be sorted such that

$$
w_i^k(\boldsymbol{x},\boldsymbol{y}) \le w_i^{k'}(\boldsymbol{x},\boldsymbol{y}), \text{ when } k < k'. \tag{17}
$$

Then, by Proposition 2, we have the following representation for the TAD index of each position.

**Corollary 1.** *Given $(\boldsymbol{x},\boldsymbol{y})$ and assuming that the service time probability distribution is given in (15), the ranking in (17) applies. If $0 < \Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] < \infty$, then the TAD index of position i can be calculated by*

$$
\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] = \sum_{k=m_i^*+1}^{K} (w_i^k(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{r}^\top \boldsymbol{x}_i) \left[ \frac{p_k}{\sum_{l\in[m_i^*]} p_l} \right], \tag{18}
$$

*where each* $w_i^k(\boldsymbol{x},\boldsymbol{y})$ *is given by (16) and*

$$m_i^\star := \max\left\{ m \in [K] \,\middle|\, w_i^m(\boldsymbol{x},\boldsymbol{y}) < \boldsymbol{r}^\top \boldsymbol{x}_i, w_i^m(\boldsymbol{x},\boldsymbol{y}) + \sum_{k=m+1}^{K} p_k (w_i^k(\boldsymbol{x},\boldsymbol{y}) - w_i^m(\boldsymbol{x},\boldsymbol{y})) \le \boldsymbol{r}^\top \boldsymbol{x}_i \right\}. \tag{19}$$

Furthermore, with the service time distribution (15), the problem (12) is a stochastic optimization problem that can be formulated as a mixed integer linear program (MILP).

**Proposition 4** (MILP for the Empirical TAD Model). *The TAD appointment model (12) with empirical service time distribution (15) has the MILP reformulation*

$$\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{y}} \quad & \sum_{i\in[2;I]} (\boldsymbol{r}^\top \boldsymbol{x}_i - \alpha_i) \\
\text{s.t.} \quad & \alpha_i + \sum_{k\in[K]} p_k \gamma_{ik} \le \boldsymbol{r}^\top \boldsymbol{x}_i, && \forall i \in [2;I] \\
& -\alpha_i \le \gamma_{ik}, && \forall i \in [2;I], k \in [K] \\
& \sum_{l=t}^{i-1} \sum_{j\in[J]} \xi_{lj}^k x_{lj} + y_t - y_i - \alpha_i \le \gamma_{ik}, && \forall i \in [2;I], t \in [i-1], k \in [K] \\
& \gamma_{ik} \ge 0, && \forall i \in [2;I], k \in [K] \\
& 0 \le \alpha_i \le \boldsymbol{r}^\top \boldsymbol{x}_i, && \forall i \in [2;I] \\
& \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y},
\end{aligned} \tag{20}$$

*where* $\gamma_{ik},\ \forall i \in [2;I], k \in [K]$ *and* $\alpha_i,\ \forall i \in [2;I]$ *are auxiliary variables.*

**Remark 5.** Leveraging on the reformulation (20) of the TAD model with empirical distribution, the regularized TAD model (14) also has an MILP reformulation without inducing additional integers.

Leveraging on the results in Corollary 1 and Proposition 4, in the following we highlight some observations of the TAD appointment model on how the appointment scheduling and sequencing operations are impacted by the tolerance levels $r_j,\ \forall j \in [J]$.

**Proposition 5** (Tolerance Impact on Sequence). *If the service times* $\tilde{\boldsymbol{\xi}}_j, j \in [J]$ *are identically distributed, then the first position is always assigned to the user of the lowest delay tolerance in the sequence solution, that is,*

$$x_{1j^\star}^\star = 1, \quad j^\star \in \operatorname*{argmin}_{j\in[J]} r_j.$$

The above result is actually trivially reasonable, which demonstrates that the first (best) position (which never suffers from delays) should be assigned to the user with lowest delay tolerance if the service times for each category of users have little difference. Furthermore, the following proposition reveals the impact of delay tolerance on the scheduling operations.

**Proposition 6** (Tolerance Impact on Schedule). *Given a fixed sequence decision* $\boldsymbol{x}$*, denote the optimal scheduling solution by* $\boldsymbol{y}^\star(\boldsymbol{\tau})$ *with delay tolerance levels* $\boldsymbol{\tau} = (\boldsymbol{r}^\top \boldsymbol{x}_i)_{i\in[I]}$*. For each position* $i \in [2;I]$*, there exists a threshold* $\Delta_i(\boldsymbol{x},\boldsymbol{\tau}) \ge 0$ *such that if the tolerance level* $\tau_i = \boldsymbol{r}^\top \boldsymbol{x}_i$ *is decreased to* $\tau_i^\flat = \tau_i - \Delta$ *with* $\Delta > \Delta_i(\boldsymbol{x},\boldsymbol{\tau})$*, then*

$$\mathbb{E}_{\hat{\mathbb{P}}}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}^\star(\boldsymbol{\tau}^\flat))] < \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}^\star(\boldsymbol{\tau}))],$$

*and there exists some* $t \in [i-1]$ *such that*

$$y_i^\star(\boldsymbol{\tau}^\flat) - y_t^\star(\boldsymbol{\tau}^\flat) > y_i^\star(\boldsymbol{\tau}) - y_t^\star(\boldsymbol{\tau}).$$

Proposition 6 implies how the optimal scheduling policy $\boldsymbol{y}^\star$ is affected by the delay tolerance changes; when the delay tolerance at some position $i$ is decreased beyond some threshold level, then the expected delay of the user at position $i$ decreases strictly, that is, $\mathbb{E}_{\hat{\mathbb{P}}}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}^\star(\boldsymbol{\tau}^\flat))] < \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}^\star(\boldsymbol{\tau}))]$. In particular, this expected delay reduction is realized by increasing the delay buffering for position $i$, that is, increasing the scheduled service duration(s) for some front position(s): $y_i^\star(\boldsymbol{\tau}^\flat) - y_t^\star(\boldsymbol{\tau}^\flat) > y_i^\star(\boldsymbol{\tau}) - y_t^\star(\boldsymbol{\tau})$, for at least one $t \in [i-1]$.

## 5. TAD Appointment Model Under Service Time Ambiguity

In this section, we extend the TAD model using known service time distribution in Section 4.2 to a model that involves distributional ambiguity, that is, by allowing the distributional set $\mathcal{F}$ to contain partially characterized service time distributions. In particular, it is noted that in the TAD appointment scheduling problem (12) under ambiguity set $\mathcal{F}$, for each $i \in [2;I]$, the TAD index is expressed as

$$\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] = \min_{\alpha\in[0,\tau_i(\boldsymbol{x})]} \left\{ \boldsymbol{r}^\top \boldsymbol{x}_i - \alpha \;\middle|\; \alpha + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[(\tilde{w}_i(\boldsymbol{x},\boldsymbol{y}) - \alpha)^+] \le \boldsymbol{r}^\top \boldsymbol{x}_i \right\}. \tag{21}$$

In the forthcoming subsections, Sections 5.1 and 5.2, we first investigate the tractable reformulations for the TAD index $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})]$ under moment and Wasserstein ambiguity sets, respectively. Then, the resultant MIP formulations of the TAD appointment model are presented in Section 5.3.

### 5.1. Moment TAD Index Model

In this subsection, we use the descriptive statistical information of the support and mean and twofold variance bounds of service times to characterize the distribution of service times $\tilde{\boldsymbol{\xi}}$. The resulting moment ambiguity set $\mathcal{F}_{\mathrm{M}}$ is specified as

$$\mathcal{F}_{\mathrm{M}} := \left\{ \mathbb{P}\in\mathcal{P}(\Xi) \;\middle|\; \begin{array}{ll} \tilde{\boldsymbol{\xi}} \sim \mathbb{P},\ \mathbb{P}[\tilde{\boldsymbol{\xi}}\in\Xi] = 1 & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_{ij}] = \mu_{ij}, & \forall i\in[I], j\in[J] \\ \mathbb{E}_{\mathbb{P}}[(\tilde{\xi}_{ij} - \mathbb{E}_{\mathbb{P}}(\tilde{\xi}_{ij}))^2] \le V_{ij}, & \forall i\in[I], j\in[J] \\ \mathbb{E}_{\mathbb{P}}[(\tilde{\xi}_{ij} + \tilde{\xi}_{\ell\kappa} - \mathbb{E}_{\mathbb{P}}(\tilde{\xi}_{ij} + \tilde{\xi}_{\ell\kappa}))^2] \le D_{ij}^{\ell\kappa}, & \forall i,\ell\in[I], j,\kappa\in[J], (i,j)\neq(\ell,\kappa) \end{array} \right\}, \tag{22}$$

where we assume that $\mu_{ij} > 0, V_{ij} > 0$, and $D_{ij}^{\ell\kappa} > 0$ for technical convenience and $\Xi = \mathbb{R}_+^{I\times J}$. In the ambiguity set $\mathcal{F}$, the first two groups of constraints represent the nonnegative support and the mean values of the service times, respectively. The third and fourth groups of constraints describe the variance bounds of marginal service times and the sum of two service times, respectively. In particular, these variance-bound constraints capture the possible correlations of the service times across user types and positions in a tractable fashion. This can be noted by recalling that

$$\mathbb{E}_{\mathbb{P}}[(\tilde{\xi}_{ij} + \tilde{\xi}_{\ell\kappa} - \mathbb{E}_{\mathbb{P}}(\tilde{\xi}_{ij} + \tilde{\xi}_{\ell\kappa}))^2] = \mathbb{E}_{\mathbb{P}}[(\tilde{\xi}_{ij} - \mathbb{E}_{\mathbb{P}}(\tilde{\xi}_{ij}))^2] + \mathbb{E}_{\mathbb{P}}[(\tilde{\xi}_{\ell\kappa} - \mathbb{E}_{\mathbb{P}}(\tilde{\xi}_{\ell\kappa}))^2] + 2\mathrm{Cov}(\tilde{\xi}_{ij}, \tilde{\xi}_{\ell\kappa}),$$

for all $i,\ell\in[I], j,\kappa\in[J], (i,j)\neq(\ell,\kappa)$ and that the total number of variance-bound constraints in (22) is $M = (I\times J)^2$, which is polynomial with respect to the problem size.

With the above notations, we have the following equivalent reformulation for the TAD index model (21) with $\mathcal{F} = \mathcal{F}_{\mathrm{M}}$ that leverages on the standard conic duality arguments for DRO problems (see Shapiro 2001). In particular, we can arrive at a polynomial-sized second-order cone program (SOCP) formulation of the TAD index at each position $i \in [2;I]$.

**Proposition 7.** *Given the ambiguity set $\mathcal{F} = \mathcal{F}_{\mathrm{M}}$ and a sequence and schedule solution $(\boldsymbol{x},\boldsymbol{y})$, if the TAD index $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] < \infty$, then it solves the SOCP*

$$\min_{\alpha\in[0,\tau_i(\boldsymbol{x})]} \ \boldsymbol{r}^\top \boldsymbol{x}_i - \alpha \tag{23}$$

$$\text{s.t.} \sum_{(l,j)\in\mathcal{A}} \mu_{lj} b_{lj} + \sum_{(l,j)\in\mathcal{A}} V_{lj} d_{ljlj} + \sum_{(l,j)\in\mathcal{A}} \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} D_{lj}^{\ell\kappa} d_{lj\ell\kappa} + \lambda \le \boldsymbol{r}^\top \boldsymbol{x}_i - \alpha \tag{24}$$

$$\sum_{(l,j)\in\mathcal{A}} \left( \sum_{(\ell,\kappa)\in\mathcal{A}} \mu_{lj} v_{lj\ell\kappa} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} \mu_{\ell\kappa} v_{lj\ell\kappa} \right) - \sum_{(l,j)\in\mathcal{A}} \sum_{(\ell,\kappa)\in\mathcal{A}} \frac{\pi_{lj\ell\kappa} + \zeta_{lj\ell\kappa}}{2} + \lambda + \alpha \ge 0 \tag{25}$$

$$\sum_{(l,j)\in\mathcal{A}} \left( \sum_{(\ell,\kappa)\in\mathcal{A}} \mu_{lj} v_{lj\ell\kappa} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} \mu_{\ell\kappa} v_{lj\ell\kappa} \right) - \sum_{(l,j)\in\mathcal{A}} \sum_{(\ell,\kappa)\in\mathcal{A}} \frac{\pi_{lj\ell\kappa} + \zeta_{lj\ell\kappa}}{2} + \lambda \ge 0 \tag{26}$$

$$\sum_{(l,j)\in\mathcal{A}} \left( \sum_{(\ell,\kappa)\in\mathcal{A}} \mu_{lj} v_{tlj\ell\kappa} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} \mu_{\ell\kappa} v_{tlj\ell\kappa} \right) - \sum_{(l,j)\in\mathcal{A}} \sum_{(\ell,\kappa)\in\mathcal{A}} \frac{\varphi_{tlj\ell\kappa} + \psi_{tlj\ell\kappa}}{2} + \lambda + \alpha \ge y_t - y_i,\ \forall t\in[i-1] \tag{27}$$

$$v_{ljlj} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} v_{lj\ell\kappa} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} v_{\ell\kappa lj} \le b_{lj}, \quad \forall (l,j)\in\mathcal{A} \tag{28}$$

$$v_{tljlj} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} v_{tlj\ell\kappa} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} v_{t\ell\kappa lj} \le b_{lj}, \ \forall t\in[i-1], l\in[I]\setminus[t;i-1], j\in[J] \tag{29}$$

$$v_{tljlj} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} v_{tlj\ell\kappa} + \sum_{(\ell,\kappa)\in\mathcal{A}\setminus\{(l,j)\}} v_{t\ell\kappa lj} \le b_{lj} - x_{lj}, \ \forall t\in[i-1], l\in[t;i-1], j\in[J] \tag{30}$$

$$\zeta_{lj\ell\kappa} - \pi_{lj\ell\kappa} \le 2d_{lj\ell\kappa}, \quad \forall (l,j)\in\mathcal{A}, (\ell,\kappa)\in\mathcal{A} \tag{31}$$

$$\psi_{tlj\ell\kappa} - \varphi_{tlj\ell\kappa} \le 2d_{lj\ell\kappa}, \ \forall (l,j)\in\mathcal{A}, (\ell,\kappa)\in\mathcal{A}, t\in[i-1] \tag{32}$$

$$\|(v_{lj\ell\kappa}, \pi_{lj\ell\kappa})^\top\|_2 \le \zeta_{lj\ell\kappa}, \quad \forall (l,j)\in\mathcal{A}, (\ell,\kappa)\in\mathcal{A} \tag{33}$$

$$\|(v_{tlj\ell\kappa}, \varphi_{tlj\ell\kappa})^\top\|_2 \le \psi_{tlj\ell\kappa}, \ \forall (l,j)\in\mathcal{A}, (\ell,\kappa)\in\mathcal{A}, t\in[i-1] \tag{34}$$

$$\boldsymbol{v}, \boldsymbol{\pi}, \boldsymbol{\zeta}\in\mathbb{R}^M, \boldsymbol{\nu}, \boldsymbol{\varphi}, \boldsymbol{\psi}\in\mathbb{R}^{M\times(i-1)}, \boldsymbol{b}\in\mathbb{R}^{I\times J}, \boldsymbol{d}\in\mathbb{R}^M_+, \lambda\in\mathbb{R}. \tag{35}$$

*where* $\boldsymbol{v}, \boldsymbol{\pi}, \boldsymbol{\zeta}, \boldsymbol{\nu}, \boldsymbol{\varphi}, \boldsymbol{\psi}, \boldsymbol{b}, \boldsymbol{d}, \lambda$ are auxiliary variables with the index set $\mathcal{A} := \{(l,j) : l\in[I], j\in[J]\}$.

### 5.2. Wasserstein TAD Index Model

Let $\mathcal{M}(\Xi)$ be the space of all probability distributions supported on $\Xi$ with $\mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\xi}\|_p] < \infty$ for any $\mathbb{Q}\in\mathcal{M}(\Xi)$ where $\|\cdot\|_p$ is the $p$-norm with $p\ge 1$. According to the optimal transportation theory (Kantorovich and Rubinstein 1958, Villani 2008), the type-1 Wasserstein distance (a.k.a. Kantorovich metric)[5] between any two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ supported on $\Xi$, denoted by $\mathrm{W}(\cdot,\cdot) : \mathcal{M}(\Xi)\times\mathcal{M}(\Xi)\longmapsto\mathbb{R}_+$, is the minimum transportation cost of moving from $\mathbb{P}$ to $\mathbb{Q}$ subject to the cost metric $\|\cdot\|_p$:

$$\mathrm{W}(\mathbb{P},\mathbb{Q}) := \inf_{\Pi}\left\{ \mathbb{E}_{\Pi}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|_p] \,\middle|\, \begin{array}{l}\Pi \text{ is joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\zeta}\\ \text{with marginals } \mathbb{P} \text{ and } \mathbb{Q}, \text{ respectively}\end{array} \right\}. \tag{36}$$

Given the empirical distribution $\hat{\mathbb{P}}$ as given in (15), the Wasserstein ambiguity set of service time distributions, parameterized by $\varphi\ge 0$, can be defined as

$$\mathcal{F}_{\mathrm{W}} := \{\mathbb{P}\in\mathcal{M}(\Xi) | \mathrm{W}(\mathbb{P},\hat{\mathbb{P}}) \le \varphi\}. \tag{37}$$

In particular, when $\varphi = 0$, the Wasserstein ambiguity set reduces to the singleton, that is, $\mathcal{F}_{\mathrm{W}} = \{\hat{\mathbb{P}}\}$, which leads to the empirical case as discussed in Section 4.2. Finally, we consider the support to be a nonnegative polyhedron, that is, $\Xi = \{\boldsymbol{\xi}\in\mathbb{R}^{IJ}_+ : \boldsymbol{B}\boldsymbol{\xi}\ge\boldsymbol{q}\}$ with $\boldsymbol{B} = (b^s_{ij})_{S\times IJ}\in\mathbb{R}^{S\times IJ}$. We next show that the TAD index under the Wasserstein ambiguity set $\mathcal{F}_{\mathrm{W}}$ also has a tractable formulation of a norm conic program.

**Proposition 8.** *Given the ambiguity set* $\mathcal{F} = \mathcal{F}_{\mathrm{W}}$ *and a sequence and schedule* $(\boldsymbol{x},\boldsymbol{y})$*, if the TAD index* $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})] < \infty$*, then it solves the norm conic program*

$$\min_{\alpha\in[0,\tau_i(\boldsymbol{x})]} \quad \boldsymbol{r}^\top\boldsymbol{x}_i - \alpha \tag{38}$$

$$\text{s.t. } \eta\varphi + \sum_{k\in[K]} p_k\gamma_k \le \boldsymbol{r}^\top\boldsymbol{x}_i - \alpha \tag{39}$$

$$\gamma_k \ge -\alpha, \qquad \forall k\in[K] \tag{40}$$

$$\sum_{l\in[I]}\sum_{j\in[J]} \xi^k_{lj} z^{kt}_{lj} + \sum_{s\in[S]} \lambda^{kt}_s q_s + y_t - y_i \le \gamma_k + \alpha, \quad \forall k\in[K], t\in[i-1] \tag{41}$$

$$\|\boldsymbol{z}^{kt}\|_{p^*} \le \eta, \qquad \forall k\in[K], t\in[i-1] \tag{42}$$

$$z^{kt}_{lj} + \sum_{s\in[S]} \lambda^{kt}_s b^s_{lj} - x_{lj} \ge 0, \qquad \forall k\in[K], t\in[i-1], l\in[t,i-1], j\in[J] \tag{43}$$

$$z^{kt}_{lj} + \sum_{s\in[S]} \lambda^{kt}_s b^s_{lj} \ge 0, \qquad \forall k\in[K], t\in[i-1], l\in[I]\setminus[t,i-1], j\in[J] \tag{44}$$

$$\boldsymbol{\lambda} \le 0, \gamma_k \ge 0, \qquad \forall k\in[K] \tag{45}$$

*where* $\eta, \boldsymbol{\gamma}, \boldsymbol{\lambda} = (\lambda^{kt}_s)_{S\times K\times(i-1)}$ *and* $\boldsymbol{Z} := (z^{kt}_{lj})_{I\times J\times K\times(i-1)}$ *are auxiliary variables, and* $\|\cdot\|_{p^*}$ *is the dual norm of* $\|\cdot\|_p$*.*

It is noted from Proposition 8 that the evaluation of the Wasserstein TAD index $\Pi_{\tau_i(\boldsymbol{x})}[\tilde{w}_i(\boldsymbol{x},\boldsymbol{y})]$ solves an LP when the order $p\in\{1,\infty\}$ for the cost metric $\|\cdot\|_p$, and it solves an SOCP when $p = 2$.

### 5.3. Mixed-Integer Conic Program Reformulations

Finally, because feasible set $\mathcal{X}$ is a finite set of binaries and $\mathcal{Y}$ is a polytope, Propositions 7 and 8 lead directly to a reformulation of a mixed-integer second-order cone program (MISOCP) or a mixed-integer conic program (MICP) for the TAD appointment model (12) under ambiguity, respectively. These MIP reformulations are formally presented as follows with proof being omitted.

**Proposition 9** (MICPs for TAD Model under Ambiguity). *Given the ambiguity set $\mathcal{F} = \mathcal{F}_{\mathrm{M}}$ or $\mathcal{F} = \mathcal{F}_{\mathrm{W}}$, if the TAD appointment scheduling problem (12) is feasible, then it solves an MISOCP,*

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{y}} \quad & \sum_{i\in[2;I]} (\boldsymbol{r}^\top \boldsymbol{x}_i - \alpha_i) \\ \text{s.t.} \quad & [v^i, \boldsymbol{\pi}^i, \boldsymbol{\zeta}^i, \boldsymbol{\nu}^i, \boldsymbol{\varphi}^i, \boldsymbol{\psi}^i, \boldsymbol{b}^i, \boldsymbol{d}^i, \lambda^i, \alpha_i] \in \{(24)\text{–}(35)\}, \quad \forall i \in [2;I] \\ & \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y},\ 0 \le \alpha_i \le \boldsymbol{r}^\top \boldsymbol{x}_i,\ \forall i \in [2;I] \end{aligned} \tag{46}$$

*where $v^i, \boldsymbol{\pi}^i, \boldsymbol{\zeta}^i, \boldsymbol{\nu}^i, \boldsymbol{\varphi}^i, \boldsymbol{\psi}^i, \boldsymbol{b}^i, \boldsymbol{d}^i, \lambda^i, \alpha_i,\ \forall i \in [2;I]$ are auxiliary variables, or an MICP,*

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{y}} \quad & \sum_{i\in[2;I]} (\boldsymbol{r}^\top \boldsymbol{x}_i - \alpha_i) \\ \text{s.t.} \quad & [\eta^i, \boldsymbol{\gamma}^i, \boldsymbol{\lambda}^i, \boldsymbol{Z}^i, \alpha_i] \in \{(39)\text{–}(45)\}, \qquad \forall i \in [2;I] \\ & \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y},\ 0 \le \alpha_i \le \boldsymbol{r}^\top \boldsymbol{x}_i,\ \forall i \in [2;I] \end{aligned}$$

*where $\eta^i, \boldsymbol{\gamma}^i, \boldsymbol{\lambda}^i, \boldsymbol{Z}^i, \alpha_i,\ \forall i \in [2;I]$ are auxiliary variables*

**Renark 6.** Leveraging on the reformulations of the TAD model under distributional ambiguity in Proposition 9, the regularized TAD model (14) also has the MISOCP and MICP, respectively, with the same number of integers.

## 6. Numerical Experiments

The numerical studies are organized as follows. In Section 6.1, we investigate the impact of the delay tolerance levels on the sequencing and scheduling solutions under a given service time distribution. We compare the TAD model with the expected delay (ED) model that minimizes the expected total delay, which is common in many existing studies (Ahmadi-Javid et al. 2017). We also compare the TAD model with the DUM$_{\mathcal{L}}$ model[6] (Qi 2017), which adopts lexicographic minimization procedure and optimizes the worst-case tolerance-incorporated delay performance across users based on CVaR measure. We perform out-of-sample tests on the three models under different parameter settings and compare their average and worst-case performance. In Section 6.2, we investigate the impact of the number of user categories on the relative performances of the three models. Furthermore, we evaluate in Section 6.3 the performance of the TAD model under service time ambiguity. Finally, in Section 6.4, we demonstrate the performance of the TAD model in a case study with real outpatient data. Moreover, additional numerical experiments are provided in E-companion EC E, which includes (E.1) impact of idle time tolerance and (E.2) impact of sample size. All the models are coded in Python and solved with Gurobi (version 9.5.0). The data and codes for the numerical studies are available in the IJOC GitHub repository (Wang et al. 2023).

### 6.1. Impact of User Delay Tolerance Thresholds

In this section, we study the impact of user delay tolerance (DT) on the scheduling with the above three models. We consider an appointment scheduling problem with 10 users from two categories, named Type 1 and Type 2, with a session length of 20 time units. Type 1 (Type 2) has $N_1$ ($N_2$) number of users with DT denoted by $r_1$ ($r_2$). The service times of Type 1 users follow an uniform distribution $\mathbf{U}[0,\delta]$, where $\delta$ is assumed to be uniformly distributed in $[3,4]$. The service times of Type 2 users follow a normal distribution $\mathbf{N}(2,\sigma^2)$, where $\sigma$ is uniformly distributed in $[0,1/3]$.

We generate 1,000 samples with the above distributions and solve the TAD model (20), the DUM$_{\mathcal{L}}$ model, and the ED model. We assume that $N_1 = N_2$, and we vary $r_2$ from 0.5 to 2 while fixing $r_1 = 1$. For each set $(r_1, r_2)$, we solve the three models to obtain their respective solutions $(\boldsymbol{x},\boldsymbol{y})$. Table 2 shows the results obtained on the scheduled appointment time for each position (scheduling decision $\boldsymbol{y}$) and the user type assigned at each position (sequencing decision $\boldsymbol{x}$).

It is not surprising that the ED model gives the same solution under different sets of $(r_1, r_2)$, because the ED model is indifferent to the value of DT. On the contrary, as $r_2$ increases from 0.5 to 2, the TAD model and the DUM$_{\mathcal{L}}$ model adjust their solutions accordingly, and more Type 2 users (with higher DT) are sequenced to the

**Table 2.** Appointment Schedule and Sequence Solutions (i.e., Time and Type) of the ED, TAD, and DUM$_{\mathcal{L}}$ Models with Different Delay Tolerance Thresholds $(r_1, r_2)$

| | | | TAD model with tolerance $(r_1, r_2)$ | | | | | | | | DUM$_{\mathcal{L}}$ model with tolerance $(r_1, r_2)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ED model | | (1, 0.5) | | (1,1) | | (1,1.5) | | (1,2) | | (1,0.5) | | (1,1) | | (1,1.5) | | (1,2) | |
| Position | Time | Type | Time | Type | Time | Type | Time | Type | Time | Type | Time | Type | Time | Type | Time | Type | Time | Type |
| 1 | 0.00 | 2 | 0.00 | 2 | 0.00 | 2 | 0.00 | 1 | 0.00 | 1 | 0.00 | 2 | 0.00 | 2 | 0.00 | 1 | 0.00 | 1 |
| 2 | 2.06 | 2 | 1.91 | 2 | 1.59 | 2 | 2.13 | 2 | 2.48 | 1 | 1.79 | 2 | 1.51 | 2 | 1.81 | 2 | 2.03 | 2 |
| 3 | 4.20 | 2 | 4.00 | 2 | 3.75 | 2 | 4.41 | 1 | 5.13 | 1 | 3.87 | 2 | 3.55 | 2 | 4.39 | 1 | 4.50 | 1 |
| 4 | 6.30 | 1 | 6.05 | 2 | 5.88 | 2 | 6.77 | 2 | 7.35 | 2 | 5.48 | 1 | 5.51 | 1 | 6.93 | 1 | 7.09 | 1 |
| 5 | 8.50 | 1 | 7.97 | 1 | 7.84 | 1 | 9.02 | 1 | 9.59 | 2 | 8.17 | 1 | 8.26 | 1 | 9.02 | 2 | 8.92 | 2 |
| 6 | 11.47 | 2 | 10.88 | 2 | 10.71 | 2 | 11.68 | 2 | 11.59 | 2 | 11.16 | 2 | 10.78 | 2 | 10.98 | 2 | 11.49 | 2 |
| 7 | 13.55 | 1 | 12.82 | 1 | 12.68 | 1 | 13.70 | 1 | 13.89 | 1 | 12.76 | 1 | 12.81 | 1 | 13.51 | 1 | 13.84 | 1 |
| 8 | 16.33 | 2 | 15.27 | 1 | 15.22 | 1 | 16.31 | 2 | 16.40 | 2 | 15.42 | 1 | 15.40 | 1 | 16.43 | 2 | 16.31 | 1 |
| 9 | 18.42 | 1 | 17.78 | 1 | 17.76 | 1 | 18.46 | 2 | 18.44 | 1 | 18.42 | 2 | 18.00 | 2 | 17.79 | 2 | 19.16 | 2 |
| 10 | 20.00 | 1 | 20.00 | 1 | 20.00 | 1 | 20.00 | 1 | 20.00 | 2 | 20.00 | 1 | 20.00 | 1 | 20.00 | 1 | 20.00 | 2 |

later positions. This suggests that the TAD and DUM$_{\mathcal{L}}$ models are able to utilize the DT information to mitigate the delay accumulation phenomenon of later users in appointment systems (Cayirli and Veral 2003). We also observe that there are more scheduling adjustments in the positions for the TAD model compared with the DUM$_{\mathcal{L}}$ model, which suggests that the TAD model is more sensitive toward changes in DT. One possible explanation is that the DUM$_{\mathcal{L}}$ model focuses on the lexicographic min-max fairness, whereas our TAD model focuses on the overall tolerance-aware delay performance, which can therefore capture more sensitively the DT effects across all positions. Furthermore, in all cases of $(r_1, r_2)$, the TAD model assigns the user with the lower DT to the first position. This observation is consistent with Proposition 5.

We next evaluate the three models using out-of-sample tests. Using their respective solutions $(\boldsymbol{x}, \boldsymbol{y})$, we implement it on 5,000 generated sample instances based on the assumed service time distributions. We evaluate the models based on the following measures. The proportion of delay over tolerance (DOT) ('$P$') shows the proportion of users (i.e., instances) who wait longer than their DT before service starts. The expected DOT ('$E$') shows that the expected time users wait beyond their DT. The standard deviation of the DOT ('$S$') indicates the variability of the waiting time by users beyond their DT. The expected waiting time ('$W$') shows the expected waiting time of users. We summarize the mean and worst case of these measures across all positions in the comparison of the TAD model with the ED model and the DUM$_{\mathcal{L}}$ model in Tables 3 and 4, respectively.

Table 3 shows that the TAD model outperforms the ED model under the mean and worst-case DOT performance criteria in proportion ('$P$'), expectation ('$E$'), and standard deviation ('$S$'). For example, the improvement of the '$E$'-performance is 1,233.3% for $(N_1, N_2) = (3,7), (r_1, r_2) = (1.5, 1)$. The positive values in the statistics '$P$' and

**Table 3.** Out-of-Sample Performance Comparisons of the TAD and ED Solutions, Where a Positive Value Indicates That the TAD Model Outperforms the ED Model Under the Intended Performance Criterion

| | | Mean performance across positions | | | | Worst-case performance across positions | | | |
|---|---|---|---|---|---|---|---|---|---|
| $(N_1, N_2)$ | $(r_1, r_2)$ | $\frac{P^{m}_{ED}}{P^{m}_{TAD}} - 1$ [a] | $\frac{E^{m}_{ED}}{E^{m}_{TAD}} - 1$ [b] | $\frac{S^{m}_{ED}}{S^{m}_{TAD}} - 1$ [c] | $\frac{W^{m}_{ED}}{W^{m}_{TAD}} - 1$ [d] | $\frac{P^{w}_{ED}}{P^{w}_{TAD}} - 1$ | $\frac{E^{w}_{ED}}{E^{w}_{TAD}} - 1$ | $\frac{S^{w}_{ED}}{S^{w}_{TAD}} - 1$ | $\frac{W^{w}_{ED}}{W^{w}_{TAD}} - 1$ |
| (5,5) | (1, 1.5) | 19.5% | 57.2% | 67.9% | −38.3% | 87.7% | 150.0% | 77.5% | −12.1% |
| | (1, 1) | 15.0% | 53.3% | 45.5% | −30.2% | 113.9% | 135.5% | 57.1% | 68.9% |
| | (1.5, 1) | 41.3% | 56.9% | 59.3% | −28.9% | 94.7% | 138.3% | 70.0% | 45.9% |
| (7,3) | (1, 1.5) | 1.8% | 29.9% | 2.3% | −22.8% | 63.8% | 101.9% | 29.5% | 56.2% |
| | (1, 1) | 1.5% | 19.8% | 4.7% | −22.8% | 57.1% | 99.3% | 37.5% | 53.4% |
| | (1.5, 1) | 18.3% | 38.2% | 22.4% | −24.3% | 88.9% | 102.5% | 41.7% | 42.7% |
| (3,7) | (1, 1.5) | 260.5% | 344.4% | 509.9% | −72.4% | 340.9% | 722.0% | 265.9% | −59.7% |
| | (1, 1) | 148.2% | 221.4% | 286.0% | −42.1% | 155.3% | 222.3% | 82.9% | 40.0% |
| | (1.5, 1) | 433.3% | 1,233.3% | 106.5% | −61.6% | 362.0% | 994.4% | 381.8% | −64.2% |

[a] $P^{m}_{A}, P^{w}_{A}$: the mean and worst-case proportion of DOT for model A across all positions.
[b] $E^{m}_{A}, E^{w}_{A}$: the mean and worst-case expected DOT for model A across all positions.
[c] $S^{m}_{A}, S^{w}_{A}$: the mean and worst-case StD of DOT for model A across all positions.
[d] $W^{m}_{A}, W^{w}_{A}$: the mean and worst-case expected waiting time for model A across all positions.

**Table 4.** Out-of-Sample Performance Comparisons of the TAD and DUM$_{\mathcal{L}}$ Solutions, Where a Positive Value Indicates That the TAD Model Outperforms the DUM$_{\mathcal{L}}$ Model Under the Intended Performance Criterion

| | | Mean performance across positions | | | | Worst-case performance across positions | | | |
|---|---|---|---|---|---|---|---|---|---|
| $(N_1, N_2)$ | $(r_1, r_2)$ | $\frac{P^{m}_{DUM_{\mathcal{L}}}}{P^{m}_{TAD}}-1$ | $\frac{E^{m}_{DUM_{\mathcal{L}}}}{E^{m}_{TAD}}-1$ | $\frac{S^{m}_{DUM_{\mathcal{L}}}}{S^{m}_{TAD}}-1$ | $\frac{W^{m}_{DUM_{\mathcal{L}}}}{W^{m}_{TAD}}-1$ | $\frac{P^{w}_{DUM_{\mathcal{L}}}}{P^{w}_{TAD}}-1$ | $\frac{E^{w}_{DUM_{\mathcal{L}}}}{E^{w}_{TAD}}-1$ | $\frac{S^{w}_{DUM_{\mathcal{L}}}}{S^{w}_{TAD}}-1$ | $\frac{W^{w}_{DUM_{\mathcal{L}}}}{W^{w}_{TAD}}-1$ |
| (5,5) | (1, 1.5) | 27.8% | 22.2% | 21.2% | 38.8% | −25.0% | −29.6% | −17.2% | 18.6% |
| | (1, 1) | 55.1% | 51.9% | 51.2% | 27.9% | −29.7% | −44.0% | −32.1% | 25.4% |
| | (1.5, 1) | 2.1% | 9.8% | 7.0% | −5.4% | −29.8% | −16.6% | −2.0% | −4.0% |
| (7,3) | (1, 1.5) | 29.0% | 40.4% | 27.1% | 20.8% | −23.1% | −23.9% | −8.6% | 21.6% |
| | (1, 1) | 17.2% | 30.1% | 20.5% | 13.6% | −33.1% | −33.8% | −8.9% | −25.9% |
| | (1.5, 1) | 13.5% | 16.8% | 12.4% | 20.7% | −38.0% | −36.5% | −10.1% | 16.0% |
| (3,7) | (1, 1.5) | 42.2% | 21.1% | 14.5% | −25.3% | 15.4% | −39.6% | −44.4% | 0.2% |
| | (1, 1) | −17.8% | −7.6% | 24.5% | 19.1% | −43.5% | −47.9% | −24.1% | 0.1% |
| | (1.5, 1) | 3.9% | −29.7% | −42.1% | 59.9% | 81.8% | 10.9% | −31.0% | 28.2% |

'$S$' suggest that the TAD model is also more effective in controlling the variability of the actual DOT levels. This shows the ability of the TAD index in capturing the tolerance effects. The ED model outperforms the TAD model in mean expected waiting time ('$W^{m}$'), which is not surprising because this is the optimization objective of ED model. Interestingly, although the TAD model does not prioritize the worst-case expected waiting time ('$W^{w}$') directly, it does so indirectly by considering the DT effect in individual positions. The TAD model generally outperforms the ED model in the worst-case expected time (see the 10th column of Table 3). Note that the abandonment property (Part 1(b) of Theorem 1) of the TAD index helps mitigate the worst-case delays.

Table 4 shows that the TAD model outperforms the DUM$_{\mathcal{L}}$ model in terms of mean DOT performance in all '$P$', '$E$', and '$S$' statistics under most of the cases. It is also not surprising that the DUM model generally outperforms the TAD model in worst-case DOT performance, because the former optimizes the worst-case performance. Interestingly, Table 4 shows that in most cases the TAD model outperforms the DUM$_{\mathcal{L}}$ model in terms of both mean and worst-case expected waiting time across all positions (the 6th and the 10th columns of Table 4). This "adaptive" performance in worst-case waiting time of the TAD model may also be attributed by the abandonment property of the TAD index, which helps reject the decisions that lead to extremely long waiting times. The DUM$_{\mathcal{L}}$ model, which (by design) prioritizes the worst-case probability level such that the associated CVaR of delay at each position is within tolerance, does not necessarily ensure the optimality in worst-case expected waiting time.

Finally, Table 5 shows the computational time in seconds for the ED, TAD, and DUM$_{\mathcal{L}}$ models. It illustrates that the computational time for the TAD model is slightly more than the ED model, for example, 102.3 seconds versus 98.0 seconds under $(N_1, N_2) = (3,7), (r_1, r_2) = (1,1)$, whereas the TAD model could generally outperform the ED model for the above-mentioned measures as we discussed before. Meanwhile, our TAD model is more efficient than the DUM$_{\mathcal{L}}$ model, which benefits from the convexity of the TAD index that ensures a computationally suitable reformulation.

To summarize, in appointment applications where the users' delay experiences are strongly related to the DT (e.g., healthcare), the TAD model can adapt the scheduling appropriately to users' DT to provide a good DOT performance with moderate computational time. In addition, the TAD model also avoids the extremely bad and unfair waiting times across positions.

## 6.2. Impact of User Heterogeneity

In this section, we study the impact of the number of user categories. We vary the number of user categories (denoted as $J$) from two to five and the tolerance levels from 0.5 to 1.5. Let $\Gamma_J$ be the set containing the DT of each

**Table 5.** Computational Time (in Seconds) of the ED, TAD, and DUM$_{\mathcal{L}}$ Models with Different Numbers of Users $(N_1, N_2)$ and Delay Tolerance Thresholds $(r_1, r_2)$

| $(N_1, N_2)$ | (3, 7) | | | (5, 5) | | | (7, 3) | | |
|---|---|---|---|---|---|---|---|---|---|
| $(r_1, r_2)$ | (1, 1.5) | (1, 1) | (1.5, 1) | (1, 1.5) | (1, 1) | (1.5, 1) | (1, 1.5) | (1, 1) | (1.5, 1) |
| ED | 97.7 | 98.0 | 97.8 | 100.7 | 99.3 | 98.7 | 93.3 | 94.6 | 94.0 |
| TAD | 102.8 | 102.3 | 109.4 | 112.0 | 108.4 | 105.6 | 117.2 | 110.8 | 115.9 |
| DUM$_{\mathcal{L}}$ | 8,362.5 | 7,697.6 | 8,370.5 | 7,144.8 | 6,436.5 | 7,678.1 | 6,691.7 | 6,132.2 | 6,710.4 |

**Table 6.** Out-of-Sample Performance Comparisons of Three Models for Different Numbers of User Types ($J$)

| | Mean performance across positions | | | | Worst-case performance across positions | | | |
|---|---|---|---|---|---|---|---|---|
| $J$ | $\frac{P^{m}_{ED}}{P^{m}_{TAD}}-1$ | $\frac{E^{m}_{ED}}{E^{m}_{TAD}}-1$ | $\frac{S^{m}_{ED}}{S^{m}_{TAD}}-1$ | $\frac{W^{m}_{ED}}{W^{m}_{TAD}}-1$ | $\frac{P^{w}_{ED}}{P^{w}_{TAD}}-1$ | $\frac{E^{w}_{ED}}{E^{w}_{TAD}}-1$ | $\frac{S^{w}_{ED}}{S^{w}_{TAD}}-1$ | $\frac{W^{w}_{ED}}{W^{w}_{TAD}}-1$ |
| 2 | 27.8% | 68.6% | 8.0% | −49.3% | 87.7% | 149.6% | 75.3% | −12.2% |
| 3 | 1.4% | 44.8% | 17.3% | −42.2% | 74.7% | 176.9% | 71.6% | 17.3% |
| 4 | 14.2% | 52.0% | 27.4% | −67.4% | 140.4% | 277.5% | 101.1% | 157.2% |
| 5 | 121.7% | 156.8% | 44.6% | −30.1% | 136.7% | 190.3% | 35.6% | 101.3% |
| | Mean performance across positions | | | | Worst-case performance across positions | | | |
| $J$ | $\frac{P^{m}_{DUM_{\mathcal{L}}}}{P^{m}_{TAD}}-1$ | $\frac{E^{m}_{DUM_{\mathcal{L}}}}{E^{m}_{TAD}}-1$ | $\frac{S^{m}_{DUM_{\mathcal{L}}}}{S^{m}_{TAD}}-1$ | $\frac{W^{m}_{DUM_{\mathcal{L}}}}{W^{m}_{TAD}}-1$ | $\frac{P^{w}_{DUM_{\mathcal{L}}}}{P^{w}_{TAD}}-1$ | $\frac{E^{w}_{DUM_{\mathcal{L}}}}{E^{w}_{TAD}}-1$ | $\frac{S^{w}_{DUM_{\mathcal{L}}}}{S^{w}_{TAD}}-1$ | $\frac{W^{w}_{DUM_{\mathcal{L}}}}{W^{w}_{TAD}}-1$ |
| 2 | 30.8% | 36.5% | 28.1% | 7.7% | 2.8% | 0.1% | −0.6% | −32.7% |
| 3 | 37.2% | 41.0% | 30.9% | 21.8% | −13.9% | −21.4% | −12.7% | −12.2% |
| 4 | 44.8% | 68.7% | 39.5% | −1.4% | −7.7% | 26.7% | 27.3% | −10.0% |
| 5 | 54.7% | 67.3% | 33.5% | 28.5% | −8.0% | −16.9% | −9.4% | 1.4% |

category and $\Psi_J$ the set containing the respective number (of users) in each category. For example, for $J = 2$, we consider $\tau_1 = 0.5, \tau_2 = 1.5, N_1 = 5, N_2 = 5$. Thus, we have $\Gamma_2 = \{0.5, 1.5\}$ and $\Psi_2 = \{5, 5\}$. We also consider the following cases: $\Gamma_3 = \{0.5, 1, 1.5\}$ with $\Psi_3 = \{3, 4, 3\}$ for $J = 3$, $\Gamma_4 = \{0.5, 0.8, 1.2, 1.5\}$ with $\Psi_4 = \{3, 2, 3, 2\}$ for $J = 4$, and $\Gamma_5 = \{0.5, 0.8, 1, 1.2, 1.5\}$ with $\Psi_4 = \{2, 2, 2, 2, 2\}$ for $J = 5$. Table 6 illustrates a similar pattern as we found in Section 6.1. More specifically, the TAD model outperforms the ED model in both mean and worst-case DOT performance measures, and meanwhile, it also outperforms the $DUM_{\mathcal{L}}$ model in mean DOT performance criteria across different user categories. It is worth noting that the mean DOT performance improvement of the TAD model over the ED and $DUM_{\mathcal{L}}$ models generally increases as heterogeneity increases ($J$ increases). One possible explanation is that the TAD model captures this by considering the sum of user's TAD index. In contrast, the $DUM_{\mathcal{L}}$ model focuses more on the worst-case situation across all positions, whereas the ED model is indifferent to users' DT.

Finally, we record the computational time of ED, TAD, and $DUM_{\mathcal{L}}$ models with different numbers of user categories as well as sample size in Table 7; it shows that as either heterogeneity across user categories or sample size increases, the computational time for each model increases, which is not surprising, although it is worth noting that it implies the same pattern as what we found in Section 6.1, that is, that the computational time for the TAD model is comparable to the ED model but more computationally suitable than the $DUM_{\mathcal{L}}$ model, which again justifies the benefit in computation of our proposed TAD model.

### 6.3. Impact of Service Time Distribution Ambiguity

Finally, we study the performance of the TAD model under distributional ambiguity with the MISOCP reformulation (46), which we refer to as the r-TAD model. To investigate the impact of the service time ambiguity, we compare the r-TAD model's performance with that of the TAD model under different out-of-sample variability situations. Specifically, we use the 1,000 samples generated in Section 6.1 to estimate the mean values and variance bounds and use these as inputs to solve the r-TAD model. The computational times for the r-TAD model are between 4,500 and 5,800 seconds for the instances of Table 8. We then compare the solution of the r-TAD model with the solution of the TAD model (20). We increase the variability of the 5,000 out-of-sample scenarios by varying $\delta$ from $\delta \sim \mathbf{U}[3,4]$ to $\delta \sim \mathbf{U}[1.5,5.5]$ and $\sigma$ from $\sigma \sim \mathbf{U}[0,1/3]$ to $\sigma \sim \mathbf{U}[0.3,1/3]$ of the service time distributions.

Table 8 shows that the relative performance of the r-TAD model over the TAD model improves with increasing variability in the out-of-sample data. In the low-variability case, that is, when the out-of-sample data are

**Table 7.** Computational Time (in Seconds) of the ED, TAD, and $DUM_{\mathcal{L}}$ Models with Different Numbers of User Types ($J$) and Sample Size ($K$), Where OOT Represents Out of Time (more than 20,000)

| | $K = 100$ | | | | $K = 500$ | | | | $K = 1{,}000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | $J = 2$ | $J = 3$ | $J = 4$ | $J = 5$ | $J = 2$ | $J = 3$ | $J = 4$ | $J = 5$ | $J = 2$ | $J = 3$ | $J = 4$ | $J = 5$ |
| ED | 2.2 | 7.6 | 12.2 | 16.6 | 32.1 | 84.6 | 147.5 | 273.9 | 99.3 | 395.5 | 687.0 | 1,282.9 |
| TAD | 2.7 | 10.5 | 13.4 | 34.0 | 36.4 | 141.5 | 174.7 | 283.6 | 124.5 | 633.7 | 784.7 | 1,727.7 |
| $DUM_{\mathcal{L}}$ | 129.1 | 291.3 | 681.1 | 762.5 | 2,005.4 | 5,224.6 | 15,390.3 | OOT | 7,225.8 | OOT | OOT | OOT |

**Table 8.** Out-of-Sample Performance Comparisons Between r-TAD and TAD $(N_1, N_2) = (5,5)$

| $(r_1, r_2)$ | Variability Level in Out-of-Sample Scenarios | $\frac{P^{m}_{TAD}}{P^{m}_{rTAD}} - 1$ | $\frac{E^{m}_{TAD}}{E^{m}_{rTAD}} - 1$ | $\frac{S^{m}_{TAD}}{S^{m}_{rTAD}} - 1$ |
|---|---|---|---|---|
| (1, 1.5) | Low variability[a] | −19.5% | −31.2% | 17.1% |
| | Medium variability[b] | 24.5% | 1.8% | 28.9% |
| | High variability[c] | 40.7% | 13.3% | 33.0% |
| (1, 1) | Low variability | −13.8% | −26.8% | −5.0% |
| | Medium variability | 13.9% | −6.8% | −0.6% |
| | High variability | 40.0% | 22.4% | 0.5% |
| (1.5, 1) | Low variability | −20.9% | −33.3% | −8.3% |
| | Medium variability | 7.3% | −9.1% | −1.6% |
| | High variability | 20.2% | 0.7% | −0.4% |

[a]Low variability: out-of-sample scenarios generated with $\delta \sim \mathbf{U}[3.0, 4.0], \sigma \sim \mathbf{U}[0.0, 1/3]$.
[b]Medium variability: out-of-sample scenarios generated with $\delta \sim \mathbf{U}[2.0, 5.0], \sigma \sim \mathbf{U}[0.2, 1/3]$.
[c]High variability: out-of-sample scenarios generated with $\delta \sim \mathbf{U}[1.5, 5.5], \sigma \sim \mathbf{U}[0.3, 1/3]$.

drawn from the original given service time distributions with $\delta \sim \mathbf{U}[3.0, 4.0], \sigma \sim \mathbf{U}[0.0, 1/3]$, the TAD model generally achieves better DOT performances. This is not surprising, because the TAD solution is obtained using the sample drawn from the distributions with the same $(\delta, \sigma)$. This performance diminishes when the variability increases. In the case of high out-of-sample variability with $\delta \sim \mathbf{U}[1.5, 5.5], \sigma \sim \mathbf{U}[0.3, 1/3]$, the r-TAD model outperforms the TAD model in DOT performance criteria across most instances. This demonstrates the robustness of the r-TAD model in maintaining the performance under high variability in the service times.

### 6.4. A Case Using Real Patient Data

In this section, we use three months of outpatient data collected from a comprehensive hospital in Singapore to evaluate the performance of the TAD model. Appointments for morning sessions start at 0830 and end at 1130. The afternoon sessions start at 1400 and end at 1700, and the hospital schedules 10 patients in each session (for each doctor). Figure 1, (a) and (b), illustrates the distributions of the waiting times (minutes) and consulting times (minutes) of the outpatient data, respectively. We also summarize the statistics for waiting time and consulting time in Table 9, which shows that the average waiting time and consulting time per patient are 32.9 and 15.1 minutes, respectively. It implies that the consulting time under different types, that is, first visit and revisit, are significantly different, which motivates us to divide the patients via their types for further analysis.

We depict the performance of the TAD model under two studies. In Section 6.4.1, we consider two categories of patients: first-visit patients and revisit patients. The follow-up consultation for revisiting patients usually has different service time characteristics from first-visit patients. In Section 6.4.2, we assume that the patients are categorized according to the length of the consulting time. This covers more general situations in practice. For instance, the consulting times for young patients are usually shorter than that for the elderly. The consultation for patients with some chronic diseases could take a longer time than others.

**6.4.1. Categorizing Patients as First Visits and Revisits.** In the first study, we categorize the patients in terms of first visits (FV) and revisits (RV), with delay tolerance (DT) as $r_{FV}$ and $r_{RV}$, respectively. Given a parameter $\beta$, we assume that a $(1-\beta)$ proportion of patients in that category are not satisfied with the waiting experience. We estimate the DT of each category by approximating it as the $\beta-$percentile of the historical waiting times, from the

**Figure 1.** (Color online) Distributions of Waiting Time (a) and Consultation Time (b) Identified from Data

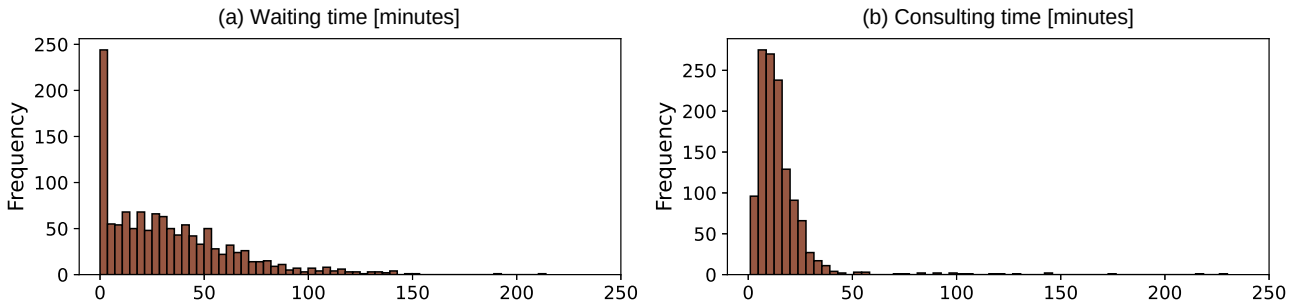


**Table 9.** Statistics of Waiting Time and Consulting Time

| | Waiting time (minutes) | | | Consulting time (minutes) | | |
|---|---|---|---|---|---|---|
| Statistics | First-visit | Revisit | Total | First-visit | Revisit | Total |
| Mean | 35.7 | 32.1 | 32.9 | 21.3 | 13.4 | 15.1 |
| Median | 31.0 | 25.0 | 27.0 | 16.0 | 11.0 | 12.0 |
| Maximum | 148.0 | 214.0 | 214.0 | 230.0 | 176.0 | 230.0 |
| Minimum | 0.0 | 0.0 | 0.0 | 2.0 | 1.0 | 1.0 |
| Std | 29.3 | 30.8 | 30.5 | 25.1 | 11.9 | 16.1 |

first two months' data. As a comparison, we vary $\beta$. Specifically, we set $\beta$ to be 20%, 30%, and 40% and refer to these cases as Cases A1, A2, and A3, respectively (see Table 10).

For each case, we solve for the appointment decisions $(\boldsymbol{x}, \boldsymbol{y})$ from the respective ED, TAD, and DUM$_{\mathcal{L}}$ models and record computational times. To evaluate, we implement out-of-sample tests using the third month's data. For each position, we record the proportion of patients with positive delay over tolerance (DOT), the expected DOT, the standard deviation of the DOTs, and the waiting time. We compute the mean of the above measures across the positions and refer to them as '$P_{\mathrm{m}}$', '$E_{\mathrm{m}}$','$S_{\mathrm{m}}$', and '$W_{\mathrm{m}}$', respectively. We also compute the worst case of the above measures across all positions and refer to them as '$P_{\mathrm{w}}$', '$E_{\mathrm{w}}$','$S_{\mathrm{w}}$', and '$W_{\mathrm{w}}$', respectively. We summarize the results in Table 11.

The TAD model has the best mean DOT performance in all cases (see column '$P_{\mathrm{m}}$' to column '$S_{\mathrm{m}}$' of Table 11). Although the ED model optimizes expected time, it has the worst mean DOT performance because it ignores the tolerance effects. For the worst-case performance, the DUM$_{\mathcal{L}}$ model outperforms the TAD model in '$P_{\mathrm{w}}$', whereas TAD is better in '$E_{\mathrm{w}}$' and '$S_{\mathrm{w}}$'. This could be due to (i) the out-of-sample effects and/or (ii) the design of the DUM$_{\mathcal{L}}$ model, whose optimization criterion based on CVaR is the worst-case probability (confidence) level across positions, which does not ensure optimality in '$E_{\mathrm{m}}$' and '$S_{\mathrm{m}}$' criteria.

It is not surprising that the ED model has the best mean performance of expected waiting time. However, the ED model results in an extremely long waiting for some positions (specifically the last one) by noting that mean and worst-case waiting time ('$W_{\mathrm{m}}$' and '$W_{\mathrm{w}}$') are 3.2 and 40.904 minutes, respectively. In contrast, both the TAD model and DUM$_{\mathcal{L}}$ model can mitigate this unfairness issue, and the TAD model has the best performance in worst-case waiting time in most cases. These superior performances of the TAD model show again its ability in mitigating the unfairness and extremely bad delay worst-case performance, which are consistent with our numerical results in Section 6.1.

**6.4.2. Categorizing Patients with Consulting Time.** In the second study, we categorize the patients, by the length of their consulting times, into three categories: Type I (0 to 15 min), Type II (15 to 30 min), and Type III (above 30 min). We let $r_I, r_{II}$, and $r_{III}$ denote the DT of patients for each category, respectively. Similarly, we consider three cases and set $r_I, r_{II}$, and $r_{III}$ as 20%, 30%, and 40% quantiles of the waiting times of respective types of patients. We refer to them as cases B1, B2, and B3, respectively (Table 12). We then compare the DOT and waiting time performance of three models of TAD, ED, and DUM$_{\mathcal{L}}$ and present the mean and worst-case out-of-sample performance statistics as well as the computational times in Table 13.

Similarly, the TAD model outperforms the ED and DUM$_{\mathcal{L}}$ models in '$E_{\mathrm{m}}$', '$P_{\mathrm{m}}$', and '$W_{\mathrm{w}}$' statistics. The TAD model performs very closely to the DUM$_{\mathcal{L}}$ model in '$E_{\mathrm{w}}$', '$P_{\mathrm{w}}$', and '$S_{\mathrm{w}}$' statistics in Cases B1 and B2 and even outperforms the DUM$_{\mathcal{L}}$ model in case B3. Clearly, the TAD model performs relatively better here than in the first study (in Section 6.4.1). This could be due to the higher heterogeneity because we have three categories here as compared with two in the first study. This is consistent with observations in Section 6.2 that the TAD model performs relatively better when adapting to higher heterogeneity effects.

Finally, the worst-case performance of the ED model is bad in all cases. For instance, when the DT is tight (cases B1 and B2), its worst-case proportion of DOT is 100%, and this always happens to the last position. Thus,

**Table 10.** Delay Tolerance Levels of Each Type of Patient Taken as Quantiles of Historical Waiting Times

| Delay tolerance level | Case A1 | Case A2 | case A3 |
|---|---|---|---|
| $r_{\mathrm{FV}}$ | 15 minutes | 22 minutes | 25 minutes |
| $r_{\mathrm{RV}}$ | 8 minutes | 12 minutes | 18 minutes |

**Table 11.** Comparison of TAD, $DUM_{\mathcal{L}}$, and ED Models in Mean and Worst-Case Performance and Computational Performance for Cases A1, A2, and A3

| | | Mean performance | | | | Worst-case performance | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(r_{FV}, r_{RV})$ | Model | $P_m$[a] | $E_m$[b] | $S_m$[c] | $W_m$[d] | $P_w$[a] | $E_w$[b] | $S_w$[c] | $W_w$[d] | Time |
| Case A1 | TAD | 12.81% | 1.560 | 5.380 | 4.900 | 28.60% | 2.560 | 6.980 | 11.921 | 12.6 s |
| | $DUM_{\mathcal{L}}$ | 13.75% | 1.844 | 6.149 | 5.140 | 17.70% | 2.896 | 8.284 | 15.810 | 1,105.4 s |
| | ED | 15.95% | 4.730 | 13.600 | 3.200 | 40.60% | 8.152 | 17.142 | 42.904 | 11.0 s |
| Case A2 | TAD | 11.50% | 1.310 | 4.863 | 4.948 | 26.60% | 2.146 | 6.415 | 12.93 | 12.3 s |
| | $DUM_{\mathcal{L}}$ | 12.03% | 1.579 | 5.639 | 8.014 | 14.70% | 2.539 | 7.587 | 15.93 | 1,190.4 s |
| | ED | 15.28% | 4.419 | 13.117 | 3.200 | 38.30% | 7.350 | 16.640 | 42.904 | 11.1 s |
| Case A3 | TAD | 7.40% | 0.860 | 4.110 | 4.850 | 17.30% | 1.311 | 5.390 | 14.700 | 12.4 s |
| | $DUM_{\mathcal{L}}$ | 8.91% | 1.098 | 4.820 | 7.910 | 12.70% | 1.596 | 5.886 | 14.670 | 1,254.1 s |
| | ED | 14.30% | 3.930 | 11.046 | 3.200 | 27.70% | 5.180 | 13.919 | 42.904 | 11.2 s |

[a]$P_m$ and $P_w$: the mean of and worst-case proportion of DOT across all positions.
[b]$E_m$ and $E_w$: the mean of and worst-case expected DOT across all positions.
[c]$S_m$ and $S_w$: the mean of and worst-case standard deviation of DOT across all positions.
[d]$W_m$ and $W_w$: the mean of and worst-case expected waiting time across all positions.

the ED model that ignores the DT effect could impose a long waiting time for the patient at one individual position, as we already observed in Section 6.4.1.

We summarize the results from both studies. The TAD model, which considers the sum of the patients' TAD index, performs well in terms of mean performance of DOT and worst-case waiting time against the ED model and $DUM_{\mathcal{L}}$ model. In contrast, the ED model performs badly in mean DOT, worst-case DOT, and waiting time statistics, whereas the $DUM_{\mathcal{L}}$ performs badly in mean waiting time statistics. In worst-case DOT performance, the TAD model can also be close to the $DUM_{\mathcal{L}}$ model. These suggest that the TAD model gives steadily good mean DOT performance and also protects the worst-case DOT and waiting time performances in practical settings.

**6.4.3. Performance Improvement with the TAD Model Over Current Practice.** Finally, we compare the results from the TAD model with the results based on the hospital's schedules made in the third month. We observe that the TAD model generally improves the waiting time beyond the patient's DT level significantly in all cases. Because the results are similar for Cases A1, A2, and A3, we present only the out-of-sample results for Case A1 in Figure 2(a). Similarly, we present only the out-of-sample results for Case B1 in Figure 2(b).

The spikes in Figure 2, (a) and (b), indicate that the majority of the patients from the TAD model experience very little waiting beyond their DT level. In contrast, the red bins over [0, 150] in the horizontal axis in both parts of the figure indicate that there is a higher proportion of patients with long waiting times beyond their DT resulting from the hospital's schedule. Specifically, the TAD model improves the average proportion of the patients with zero DOT across different categories of patients from 16% and 15% to 52% and 45% in Case A1 and Case B1, respectively.

The higher spike in Figure 2(b) indicates that the TAD model has better improvements in Case B1 than in Case A1. Recall that we have two categories in Case A1 and three categories in case B1 based on two different categorizations. This leads to an important managerial insight that an appropriate categorization of patients can have a big impact on the performance of the model. Thus, to improve the waiting experience of users, operations managers should not downplay the importance of a good categorization of users.

## 7. Conclusion

In this paper, we study a joint appointment sequencing and scheduling problem with service time uncertainty. We define a new measure, the TAD index, to quantify the users' dissatisfaction to delay experience. The TAD

**Table 12.** Delay Tolerance Levels of Each Type of Patient Taken as Quantiles of Historical Waiting Times

| Delay tolerance level | Case B1 | Case B2 | Case B3 |
|---|---|---|---|
| $r_I$ | 7 minutes | 12 minutes | 18 minutes |
| $r_{II}$ | 11 minutes | 15 minutes | 21 minutes |
| $r_{III}$ | 10 minutes | 13 minutes | 20 minutes |

**Table 13.** Comparison of TAD, $DUM_{\mathcal{L}}$, and ED Models in Mean and Worst-Case Performance and Computational Performance for Cases B1, B2, and B3

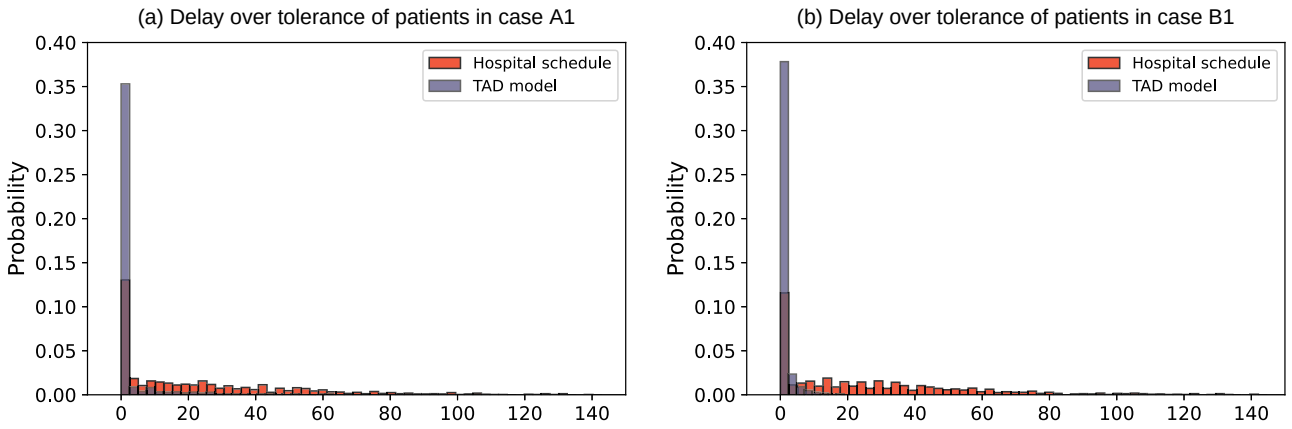
| | | Mean performance | | | | Worst-case performance | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(r_I, r_{II}, r_{III})$ | Model | $P_m$ | $E_m$ | $S_m$ | $W_m$ | $P_w$ | $E_w$ | $S_w$ | $W_w$ | Time |
| Case B1 | TAD | 17.60% | 0.645 | 1.568 | 5.381 | 37.70% | 2.008 | 3.566 | 11.850 | 42.3 s |
| | $DUM_{\mathcal{L}}$ | 36.20% | 1.410 | 2.530 | 7.272 | 37.50% | 1.870 | 3.530 | 12.751 | 2,779.1 s |
| | ED | 11.25% | 1.170 | 0.260 | 3.891 | 100.00% | 10.50 | 2.130 | 20.520 | 41.6 s |
| Case B2 | TAD | 3.98% | 0.149 | 0.517 | 5.312 | 22.90% | 1.081 | 2.553 | 10.680 | 43.1 s |
| | $DUM_{\mathcal{L}}$ | 16.30% | 0.546 | 1.450 | 7.580 | 21.50% | 1.020 | 2.520 | 12.503 | 2,945.2 s |
| | ED | 11.10% | 0.830 | 0.230 | 3.891 | 100.00% | 7.500 | 2.120 | 20.520 | 42.2 s |
| Case B3 | TAD | 0.70% | 0.019 | 0.140 | 6.304 | 4.70% | 0.140 | 0.260 | 11.420 | 42.1 s |
| | $DUM_{\mathcal{L}}$ | 0.71% | 0.020 | 0.143 | 7.912 | 4.90% | 0.140 | 0.730 | 12.684 | 3,233.1 s |
| | ED | 4.41% | 0.110 | 0.210 | 3.891 | 37.30% | 1.070 | 1.760 | 20.520 | 41.8 s |

index is a convex decision criterion that captures the frequency and intensity of delays above tolerance. We use the TAD index to minimize the collective delay dissatisfaction level across heterogeneous users in the appointment scheduling problem. The TAD model effectively coordinates the users' heterogeneous tolerances with the service time distribution and adapts the appointment sequence-and-schedule decision to optimize the users' total tolerance-aware delay level. The convexity of the TAD index endows it with an appealing computational advantage in modeling the appointment scheduling problem, which enjoys computationally attractive reformulations of mixed-integer linear program and mixed-integer conic programs under known empirical distribution and distributional ambiguity, respectively. Also, our TAD model is extended to incorporate overtime and idle time.

We identify several insights of the TAD model. (i) It is always optimal to assign the user of the lowest delay tolerance to the first position when service times are identically distributed. (ii) When the delay tolerance at some position is decreased beyond some threshold level, then the model increases the scheduled service time duration of some front users so as to mitigate the expected delay of the user at that position. (iii) Our computational results show that the proposed TAD model performs consistently well in mean delay-over-tolerance performance and also mitigate effectively the worst-case delay-over-tolerance and waiting time. The relative performance of the TAD model over existing approaches improves when the number of user categories (heterogeneity) increases. We also evaluate the TAD model using a case study with real outpatient data, which justifies the effectiveness of the TAD model in a practical setting. We highlight an important managerial insight, that managers should not neglect the importance of categorization of users because it may lead to potential improvement in the system performance.

Finally, in the current paper, we focus only on how to develop the scheduling model for the service system given the acquired delay tolerance information of users, and an implicit assumption for our focal setting is that the delay tolerance is estimated using the group-level information (e.g., the patient classification system in the hospitals). Nevertheless, estimating the delay tolerance using only the group-level information may conflict with the user's individual intention and characteristics in different circumstances that are valuable for the scheduling.

**Figure 2.** (Color online) DOT Distributions of Hospital Schedules and TAD Schedules in Case A1 (a) and Case B1(b)



This issue could be mitigated by utilizing more side information from the users' features to characterize the users' delay tolerance. In order to address the above issue, as for our future research, we can make the users' delay tolerance and the associated scheduling in a data-driven fashion that adapts to the side information. This also leads to several modeling challenges. For instance, how to model the data-driven scheduling problem with feature-based delay tolerance in a tractable and incentive-compatible fashion is practice relevant—that is also tough—by noting that some of the side information on the users' feedback of their expected delay time could contain misreporting. On the one hand, our proposed framework could be extended to incorporate prescriptive approaches with statistical and machine-learning models (Bertsimas and Kallus 2020, Bertsimas and Koduri 2022, Hu et al. 2022) for more accurately characterizing the service time uncertainty. These interesting and important topics will be discussed in our forthcoming studies.

## Endnotes

[1] The characteristics of patients' delay tolerance can be studied based on the information acquired according to patient classifications, for example, the medical department, symptoms, chronic disease history, gender, and age (Moschis et al. 2003, Hill and Joonas 2005). Also, in emergency medicine in Canada, the Canadian Triage and Acuity Scale (Bullard et al. 2008) has classified patients into five priority classes, and each class associates with an access time threshold standard.

[2] The tractability here is in the sense that all the binary variables in the TAD models are merely the sequencing decisions that are present in the original (deterministic) appointment scheduling problem, and when binary sequence variables are fixed, the resulting TAD appointment models are polynomial-time solvable.

[3] The certainty equivalent—as it will imply in the forthcoming discussions—has several appealing properties, which is also closely related to the coherent risk measures (Ben-Tal and Teboulle 2007, Drapeau and Kupper 2013, Vinel and Krokhmal 2017).

[4] This can be seen if we normalize the uncertain delays $\tilde{w}$ to $\tilde{v} := \tilde{w} - \tau$ (i.e., delay-tolerance excesses) and tolerance $\tau$ to zero accordingly; the monotonicity, convexity, and positive homogeneity of the TAD index then coincide with those of the coherent risk measures, with respect to $\tilde{v}$.

[5] The more general $r$-Wasserstein distance ($r \geq 1$) can also be defined by setting the cost function as $\|\cdot\|_p^r$. In this paper, we follow Esfahani and Kuhn (2018) to focus on the 1-Wasserstein distance to construct the ambiguity set.

[6] The $\mathrm{DUM}_{\mathcal{L}}$ model with the lexicographic minimization procedure solves a sequence of optimization problems and takes care every user's delay. It should be noted that because the DUM is quasi-convex, each optimization problem is therefore solved with binary search in the $\mathrm{DUM}_{\mathcal{L}}$ model.

## References


Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *European J. Oper. Res.* 258(1):3–4.

Artzner P, Delbaen F, Eber J, Heath D (1999) Coherent measures of risk. *Math. Finance* 9(3):203–228.

Bai M, Storer RH, Tonkay GL (2022) Surgery sequencing coordination with recovery resource constraints. *INFORMS J. Comput.* 34(2):1207–1223.

Ben-Tal A, Teboulle M (2007) An old–new concept of convex risk measures: The optimized certainty equivalent. *Math. Finance* 17(3):449–476.

Benjaafar S, Chen D, Wang R, Yan Z (2023) Appointment scheduling under a service-level constraint. *Manufacturing Service Oper. Management* 25(1):70–87.

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Sci.* 66(3):1025–1044.

Bertsimas D, Koduri N (2022) Data-driven optimization: A reproducing kernel Hilbert space approach. *Oper. Res.* 70(1):454–471.

Bertsimas D, Dunn J, Pawlowski C, Zhuo YD (2019) Robust classification. *INFORMS J. Optimization* 1(1):2–34.

Bleustein C, Rothschild DB, Valen A, Valaitis E, Schweitzer L, Jones R (2014) Wait times, patient satisfaction scores, and the perception of care. *Amer. J. Management Care* 20(5):393–400.

Bullard MJ, Unger B, Spence J, Grafstein E, CTAS National Working Group (2008) Revisions to the Canadian emergency department triage and acuity scale (CTAS) adult guidelines. *CJEM* 10(2):136–142.

Cayirli T, Veral E (2003) Outpatient scheduling in healthcare: A review of literature. *Production Oper. Management* 12(4):519–549.

Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. *Production Oper. Management* 17(3):338–353.

Chan TC, Killeen JP, Kelly D, Guss DA (2005) Impact of rapid entry and accelerated care at triage on reducing emergency department patient wait times, lengths of stay, and rate of left without being seen. *Ann. Emerg. Med.* 46(6):491–497.

Chen RR, Robinson LW (2014) Sequencing and scheduling appointments with potential call-in patients. *Production Oper. Management* 23(9):1522–1538.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Sci.* 10(1):13–24.

Drapeau S, Kupper M (2013) Risk preferences and their robust representation. *Math. Oper. Res.* 38(1):28–62.

Erdogan SA, Denton B (2013) Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS J. Comput.* 25(1):116–132.

Esfahani PM, Kuhn D (2018) Data-driven distributional robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program* 171(1–2):115–166.

Föllmer H, Schied A (2002) Convex measures of risk and trading constraints. *Finance Stoch.* 6:429–447.

Gao R, Kleywegt AJ (2023) Distributional robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* 48(2):603–655.
Gupta D, Denton B (2008) Appointment scheduling in healthcare: Challenges and opportunities. *IIE Trans.* 40(9):800–819.
Hill CJ, Joonas K (2005) The impact of unacceptable wait time on healthcare patients' attitudes and actions. *Health Marketing Q.* 23:69–87.
Hu J, Chen Z, Wang S (2022) Budget-driven multi-period hub location: A robust time series approach. Preprint, submitted September 22, https://dx.doi.org/10.2139/ssrn.4221971.
Huang XM (1994) Patient attitude toward waiting in an outpatient clinic and its applications. *Health Serv. Management Res.* 7(1):2–8.
Homem-de-Mello T, Kong Q, Godoy-Barba R (2022) A simulation optimization approach for the appointment scheduling problem with decision-dependent uncertainties. *INFORMS J. Comput.* 34(5):2845–2865.
Jiang R, Ryu M, Xu G (2019) Data-driven distributionally robust appointment scheduling over Wasserstein balls. Preprint, submitted July 7, https://arxiv.org/abs/1907.03219.
Jiang R, Shen S, Zhang Y (2017) Integer programming approaches for appointment scheduling with random no-shows and service durations. *Oper. Res.* 65(6):1638–1656.
Kantorovich LV, Rubinstein SG (1958) On a space of totally additive functions. *Vestnik St. Petersburg Univ. Math.* 13(7):52–59.
Kiran T, O'Brien P (2015) Challenge of same-day access in primary care. *Can. Fam. Physician.* 61(5):399–400.
Kong QX, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):711–726.
Kong QX, Li S, Liu N, Teo CP, Yan Z (2020) Appointment scheduling under time-dependent patient no-show behavior. *Management Sci.* 66(8):3480–3500.
Kuiper A, Lee RH (2022) Appointment scheduling for multiple servers. *Management Sci.* 68(10):7422–7440.
Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Sci.* 64(5):1975–1996.
Liu N, Truong VA, Wang X, Anderson BR (2019) Integrated scheduling and capacity planning with considerations for patients' length-of-stays. *Production Oper. Management* 28(7):1735–1756.
Macario A (2010) Is it possible to predict how long a surgery will last? *Medscape Anesthesiology* (July 14), http://medscape.com/viewarticle/724756.
Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Sci.* 61(2):316–334.
Malloch K, Meisel M (2013) Patient classification systems: State of the science. *Nurse Lead.* 11(6):35–40.
Marynissen J, Demeulemeester E (2019) Literature review on multi-appointment scheduling problems in hospitals. *European J. Oper. Res.* 272(2):407–419.
McCarthy K, McGee HM, O'Boyle CA (2000) Outpatient clinic delays and non-attendance as indicators of quality. *Psychol. Health Med.* 5(3):287–293.
Moschis GP, Bellinger DN, Curasi CF (2003) What influences the mature customer? *Marketing Health Serv.* 23:16–21.
Murray M, Tantau C (2000) Same-day appointments: Exploding the access paradigm. *Fam. Pract. Management* 7(8):45–50.
Nikolova S, Harrison M, Sutton M (2016) The impact of waiting time on health gains from surgery: Evidence from a national patient-reported outcome data set. *Health Econom.* 25(8):955–968.
Qi J (2017) Mitigating delays and unfairness in appointment systems. *Management Sci.* 63(2):566–583.
Robinson LW, Chen RR (2011) Estimating the implied value of the customer's waiting time. *Manufacturing Service Oper. Management* 13(1):53–57.
Samorani M, LaGanga LR (2015) Outpatient appointment scheduling given individual day-dependent no-show predictions. *European J. Oper. Res.* 240(1):245–257.
Shapiro A (2001) On duality theory of conic linear problems. Goberna MA, López MA, eds. *Semi-Infinite Programming: Recent Advances* (Springer, Berlin, Heidelberg), 135–165.
Sharif AB, Stanford DA, Taylor P, Ziehins H (2014) A multi-class multi-server accumulating priority queue with application to healthcare. *Oper. Res. Health Care* 3:73–79.
Shehadeh KS, Padman R (2022) Stochastic optimization approaches for elective surgery scheduling with downstream capacity constraints: Models, challenges, and opportunities. *Comput. Oper. Res.* 137:105523.
Shehadeh KS, Cohn AE, Jiang R (2021) Using stochastic programming to solve an outpatient appointment scheduling problem with random service and arrival times. *Naval Res. Logist.* 68(1):89–111.
Sugiyama M (2015) *Introduction to Statistical Machine Learning* (Morgan Kaufmann, San Francisco).
Taylor S (1994) Waiting for service: The relationship between delays and evaluations of service. *J. Marketing* 58(2):56–69.
Villani C (2008) *Optimal Transport: Old and New*, vol. 338 (Springer Science & Business Media, Berlin).
Vinel A, Krokhmal PA (2017) Certainty equivalent measures of risk. *Ann. Oper. Res.* 249(1–2):75–95.
Wang S, Liu N, Wan G (2019) Managing appointment-based services in the presence of walk-in customers. *Management Sci.* 66(2):667–686.
Wang S, Li J, Ang M, Ng TS (2023) Data for appointment scheduling with delay-tolerance heterogeneity. https://dx.doi.org/10.1287/ijoc.2023.0025.cd, https://github.com/INFORMSJoC/2023.0025.
Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.
Wu X, Zhou X (2008) Stochastic scheduling to minimize expected maximum lateness. *European J. Oper. Res.* 190(1):103–115.
Wu X, Zhou S (2022) Sequencing and scheduling appointments on multiple servers with stochastic service durations and customer arrivals. *Omega* 106:102523.
Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *J. Mach. Learn. Res.* 10:1485–1510.
Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. *Manufacturing Service Oper. Management* 19(4):639–656.
Zacharias C, Yunes T (2020) Multimoduarity in the stochastic appointment scheduling with discrete arrival epochs. *Management Sci.* 22(6):744–763.
Zacharias C, Liu N, Begen MA (2022) Dynamic interday and intraday scheduling. *Oper. Res.*, ePub ahead of print August 24, https://doi.org/10.1287/opre.2022.2342.
Zhang Y, Shen S, Erdogan SA (2017) Distributionally robust appointment scheduling with moment-based ambiguity set. *Oper. Res. Lett.* 45:139–144.