Addressing Nonlinearity and Uncertainty via Mixed Integer Programming Approaches

by

Minseok Ryu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2020

Doctoral Committee:

Assistant Professor Ruiwei Jiang, Chair
Professor Ian Hiskens
Professor Jon Lee
Dr. Harsha Nagarajan, Los Alamos National Lab
Associate Professor Siqian Shen

Minseok Ryu

msryu@umich.edu

ORCID iD: 0000-0002-1954-7722

To my family

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to Prof. Ruiwei Jiang, who has advised and supported me in many different aspects during the past five years. I have extremely enjoyed working with him and been inspired by his unlimited passion and excellence in exploring and working on new research topics. His brilliant research advice and guidance have motivated me to work hard and led me to the exciting part of doing research. Also, I have been benefited from taking his classes. How he organized course materials, conveyed knowledge, and interacted with students from diverse backgrounds were great examples to me. He helped me to create my own curriculum and provided me an opportunity to teach a course, which was a fantastic experience. I am very happy that I spent my Ph.D. life under the supervision of Prof. Ruiwei Jiang. Thanks for becoming an excellent advisor, teacher, and role model.

I am also deeply thankful to Prof. Siqian Shen for her considerable care, encouragement, and valuable feedback. How she cares about students, manages her research group, and contributes to our research community has inspired me a lot. My deep appreciation goes to Prof. Jon Lee, who provided valuable comments during my preliminary exam and final defense. His feedbacks made me think more than twice, which ended up improving this dissertation. My special thanks go to Prof. Ian Hiskens, who happily served as a doctoral committee member and gave me inspiration on power system modeling and methodologies. I am very grateful to Dr. Harsha Nagarajan, who supported me in numerous ways and helped me to finish Chapter 4 in this dissertation. Collaborating with him has been immensely enjoyable, as he always inspired me. I really appreciate the time I was a research intern at Los Alamos National Lab. Thanks to his help and supports, I was able to have such an enjoyable memory in my life. My special gratitude goes to Prof. Guanglin Xu, who is always eager to open inspiring discussions while collaborating and supported me in various ways. Thanks to his untiring passion, Chapter 2 of this dissertation was able to be completed.

I am very grateful to Prof. Brian Denton, who encouraged and supported me in various ways. I appreciate his excellent comments and feedback on my current and future research directions. Also, I do appreciate how he takes care of each member of our department as a department chair. Thanks to his warm care and passion, I was able to become a member of this family-like department. My Ph.D. life was hugely successful since I became to know many colleagues and friends from the IOE department. My deep gratitudes go to Yiling Zhang, Yuanyuan Guo, Emily Tucker, Karmel Shehadeh, Ece Sanci, Fatemeh Navidi, Xian Yu, Xiangkun Shen, Hao Yuan, Zhiyuan Huang, Qi Luo, Gian-Gabriel Garcia, Wesley Marrero Colon, Esmaeil and Elnaz, Hideaki Nakao, Miao Yu, Haoming Shen, Luze Xu, Junhong Guo. My special thanks to Geunyeong Byeon, Sol Lim, Youngchan Jang, Seokjoo Kwak, and Seokhyun Chung.

Finally, I would like to express my deepest gratitude to my family (Jaemun Ryu, Haejin Kim, Danny Jae-suh Lee, Jeong-ah Ryu, and Adeleine Suh-won Lee) who encourage, support, and love me all the time.

TABLE OF CONTENTS

LIST OF FIGURES

Figures

LIST OF TABLES

Tables

LIST OF APPENDICES

ABSTRACT

The main focus of the dissertation is to develop decision-making support tools that address nonlinearity and uncertainty appearing in real-world applications via mixed integer programming (MIP) approaches. When making decisions under uncertainty, knowing the accurate probability distributions of the uncertain parameters can help us predict their future realization, which in turn helps to make better decisions. In practice, however, it is oftentimes hard to estimate such a distribution precisely. As a consequence, if the estimated distribution is biased, the decisions thus made can end up with disappointing outcomes. To address this issue, we model uncertainty in the decision-making process through a distributionally robust optimization (DRO) approach, which aims to find a solution that hedges against the worst-case distribution within a pre-defined ambiguity set, i.e., a collection of probability distributions that share some distributional and/or statistical characteristics in common. The role of the ambiguity set is crucial as it affects both solution quality and computational tractability of the DRO model. In this dissertation, we tailor the ambiguity sets based on the available historical data in healthcare operations and energy systems, and derive efficient solution approaches for the DRO models via MIP approaches. Through extensive numerical studies, we show that the DRO solutions yield better out-of-sample performance, and the computational performance of the proposed MIP approaches is encouraging. Finally, we provide some managerial insights into the operations of healthcare and energy systems.

CHAPTER 1

Introduction

The increasing availability of data provides a significant opportunity for making better decisions by learning from data. For example, a hospital manager can predict future nurse demands by historical patient census data, which helps make better staffing decisions. Such data-driven decisions can produce a satisfactory future outcome, e.g., satisfying nurse demands with minimum staffing, but this may be challenging due to the data uncertainty (e.g., non-stationary processes, noises, and measurement errors). To address this challenge, several approaches have been developed to inform decision-making under uncertainty.

The first approach is stochastic programming (SP), which aims to find a solution that optimizes the expected outcome given the probability distribution of uncertain parameters. The following is an example of the SP formulation:

$$\text{[SP]:} \quad \min_{x \in \mathcal{X}} \left\{ c^T x + \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x, \tilde{\xi})] \right\},$$

where $\mathcal{X} \subseteq \mathbb{R}^{n_1}$ is a feasible region of the decision variables x, $c \in \mathbb{R}^{n_1}$ is a cost vector, $\mathbb{P}$ is a joint probability distribution of the uncertain parameters $\tilde{\xi}$ supported on the set $\Xi \subseteq \mathbb{R}^{n_2}$, and $\mathcal{Q}(x, \tilde{\xi})$ is a recourse function representing the future outcome, which depends on the decision x and the uncertainty realization $\tilde{\xi}$ (see, e.g., Shapiro et al.[2], Birge and Louveaux[3] for more details about SP). This approach requires a perfect knowledge of the underlying probability distribution $\mathbb{P}$, which can be inferred from the available historical data, prior beliefs, and expert opinions.

Robust optimization (RO) is another approach that informs decision-making under uncertainty. This approach aims to find a solution that hedges against the worst-case outcome of the uncertain parameters, given only a set of all possible outcomes. The following is an example of the RO formulation:

$$\text{[RO]:} \quad \min_{x \in \mathcal{X}} \left\{ c^T x + \max_{\tilde{\xi} \in \Xi} [\mathcal{Q}(x, \tilde{\xi})] \right\},$$

where one minimizes the total cost with regard to the worst-case scenario $\tilde{\xi} \in \Xi$ (see Ben-Tal et al.[4], Ben-Tal et al. [5] for more details). This approach only requires the support set of uncertain parameters (or a confidence region of them), which can be easily estimated from the historical data.

In a wide range of real-world applications, it is often the case that the actual probability distribution of the uncertain parameters cannot be estimated precisely due to the lack of the available data or imprecision of determining distribution using statistical tools (e.g., maximum likelihood estimator). If the chosen distribution $\mathbb{P}$ in [SP] is biased, it may lead to a solution x^{SP} with the disappointing out-of-sample performance. If we do not incorporate the probability distribution and only consider the support set of the uncertain parameters as in [RO], the obtained solution x^{RO} may be too conservative to use in practice. A distributionally robust optimization (DRO) approach has been proposed to address this issue. This approach assumes that the actual probability distribution of the uncertain parameters is ambiguous. Accordingly, it is assumed to belong with a pre-defined *ambiguity set*, which is a set of probability distributions that share some statistical and/or distributional characteristics in common. The following is an example of the DRO formulation:

$$\text{[DRO]:} \quad \min_{x \in \mathcal{X}} \left\{ c^T x + \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x, \tilde{\xi})] \right\},$$

where $\mathcal{D}$ is an ambiguity set, i.e., a collection of probability distributions of the uncertain parameters $\tilde{\xi}$ supported on the set Ξ. This approach aims to find a solution x^{DRO} that optimizes the expected outcome $\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x, \tilde{\xi})]$ over the worst-case probability distribution $\mathbb{P}$ within the ambiguity set $\mathcal{D}$. Depending on the structure of the ambiguity set, the DRO approach has a chance to provide solution x^{DRO} that (i) has better out-of-sample performance than x^{SP} and (ii) is less conservative than x^{RO}. Note that if the ambiguity set $\mathcal{D}$ is a singleton, i.e., a specific distribution, then [DRO] reduces to [SP]. In contrast, if the ambiguity set $\mathcal{D}$ includes all distributions of the uncertain parameters $\tilde{\xi}$ supported on the set Ξ, then [DRO] reduces to [RO].

For the rest of Chapter 1, we review various ambiguity sets and their corresponding DRO models in Section 1.1, review some real-world applications of DRO in Section 1.2, and present overview of the dissertation in Section 1.3.

1.1 Distributionally Robust Optimization

Scarf [6] proposed and studied the first DRO model in a newsvendor problem. The ambiguity set was constructed based on the mean and covariance of the demand, i.e., all distributions within this set have the same mean and covariance. The proposed DRO model aims to find an inventory decision that maximizes the expected profit over the worst-case probability distribution of the demand within the ambiguity set. Since then, there have been numerous literatures (see, e.g., Dupačová [7, 8], Shapiro and Kleywegt [9], Delage and Ye [10], Esfahani and Kuhn [11]) that studied DRO models under various ambiguity sets.

The role of the ambiguity set in DRO is crucial since its structure affects not only the out-of-sample performance of the solution x^{DRO}, but also the computational tractability of the DRO model. Depending on the structure of the ambiguity set, numerous solution methodologies have been developed, and the most common techniques are based on tools in semi-infinite programming and duality. According to Rahimian and Mehrotra [12], the ambiguity sets proposed in the literature can be categorized by (i) discrepancy-based, (ii) moment-based, (iii) shape-preserving-based, and (iv) kernel-based models. In the next two subsections, we review the first two categories of the ambiguity sets, because, in this dissertation, the ambiguity set proposed in Chapter 2 belongs to the discrepancy-based ambiguity sets (Section 1.1.1), and the ambiguity sets proposed in Chapters 3 and 4 belong to the moment-based ambiguity sets (Section 1.1.2).

1.1.1 Discrepancy-based ambiguity set

The discrepancy-based ambiguity set is defined as a collection of probability distributions whose discrepancy to a reference distribution is sufficiently small. The following is a generic form of the discrepancy-based ambiguity set:

$$\mathcal{D}(\mathbb{P}_0, \epsilon) := \left\{ \mathbb{P} \in \mathcal{P}(\Xi) \ : \ d(\mathbb{P}, \mathbb{P}_0) \leq \epsilon \right\},$$

where $\mathcal{P}(\Xi)$ represents the set of all probability distributions supported on the set Ξ, $\mathbb{P}_0$ represents the reference distribution, which, e.g., can be estimated from the available historical data, ϵ is the upper limit of the discrepancy (also called the size of the ambiguity set), and $d(\mathbb{P}, \mathbb{P}_0)$ is a functional that measures the discrepancy between two probability distributions $\mathbb{P}, \mathbb{P}_0 \in \mathcal{P}(\Xi)$. There have been various ways of defining the discrepancy function $d(\mathbb{P}, \mathbb{P}_0)$, for examples, optimal transport discrepancies ([13], [14]), ϕ-divergence ([15], [16], [17]), total variation distance ([18], [19]), Prohorov metric ([20]), and ζ-structure metrics ([21], [22]).

In this dissertation, we consider the p-Wasserstein ambiguity set in Chapter 2, which belongs to the first sub-category, i.e., optimal transport discrepancies. For $p \in [1, \infty)$, the p-Wasserstein distance measures the statistical distance between two distributions $\mathbb{Q}_1$ and

$\mathbb{Q}_2$ as follows:

$$d(\mathbb{Q}_1, \mathbb{Q}_2) = \left(\inf_{\Pi \in \mathcal{P}(\mathbb{Q}_1, \mathbb{Q}_2)} \mathbb{E}_{\Pi} \left[\|u_1 - u_2\|^p \right] \right)^{1/p},$$

where random variables u_1 and u_2 follow marginal distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively, and $\mathcal{P}(\mathbb{Q}_1, \mathbb{Q}_2)$ represents the set of all joint probability distributions Π of (u_1, u_2) with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. The DRO model under the Wasserstein ambiguity set aims to find a solution $x \in \mathcal{X}$ that optimizes the expected outcome $\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x, \tilde{\xi})]$ over the worst-case distribution $\mathbb{P}$ among all the distributions that are located within a radius ϵ of the Wasserstein ball centered at the reference distribution. Multiple Wasserstein-based DRO models and their corresponding solution approaches have been proposed (see, e.g., Gao and Kleywegt [23], Esfahani and Kuhn [11], Zhao and Guan [24], Hanasusanto and Kuhn [25]). Also, Noyan et al. [26], and Luo and Mehrotra [27] studied the decision-dependent Wasserstein-based ambiguity sets, i.e., the ambiguity set is affected by the decision x.

1.1.2 Moment-based ambiguity set

The moment-based ambiguity set is a set of distributions that share the same moment information in common. Under this umbrella, there exist various types of ambiguity sets, and some of them have been given names. For example, a *Markov* ambiguity set contains all distributions with known mean and support, and a *Chebyshev* ambiguity set contains all distributions with bounds on the first and second-order moments (see Hanasusanto et al. [28] for more details).

In this dissertation, we consider the following form of the moment-based ambiguity set:

$$\mathcal{D} = \{\mathbb{P} \in \mathcal{P}(\Xi) \; : \; \mathbb{E}_{\mathbb{P}_j}[\tilde{\xi}_j^q] = \mu_j^q, \; \forall j \in [J], q \in [Q]\},$$

where $\mathcal{P}(\Xi)$ represents the set of all probability distribution on Ξ, $\mathbb{P}_j$ is a marginal distribution of the j-th element $\tilde{\xi}_j$ of the uncertain parameter, and the q-th moment of the uncertain parameter $\mathbb{E}_{\mathbb{P}_j}[\tilde{\xi}_j^q]$ is given as μ_j^q, which can be obtained from the available historical data. Note that we consider (i) a decision-dependent moment-based ambiguity set in Chapter 3, i.e., for some $\hat{j}$, we have $\mu_{\hat{j}} = f(x)$ where $f(x)$ is a function of the decision variables x, and (ii) a *Markov* ambiguity set in Chapter 4, i.e., $Q = 1$. The DRO model under the moment-based ambiguity set aims to find a solution $x \in \mathcal{X}$ that optimizes the expected outcome $\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x, \tilde{\xi})]$ over the worst-case distribution $\mathbb{P}$ among all the probability distributions that have the same moment information defined in the ambiguity set $\mathcal{D}$. Numerous moment-based DRO models and their corresponding solution approaches have been proposed (see, e.g., Scarf [6], Dupačová [29], Delage and Ye [10], Goh and Sim [30], Bertsimas et al. [31], Wiesemann et al. [32]). Also, Royset and Wets [33] studied a DRO model with a decision-dependent moment-based ambiguity set.

1.2 Applications of DRO

We review applications of DRO in two most closely related areas to this dissertation, namely healthcare and energy system operations.

First of all, *healthcare operations* refers to the application area, where a hospital operator aims to make a set of decisions for efficient services in a hospital, which relates to not only operational costs but also the satisfaction of workers and patients. The following are examples of well-known optimization problems in this area; (i) *appointment scheduling problem* aims to decide the arrival time of each appointment for the efficient and smooth flow of patients, (ii) *staffing problem* seeks to determine the size of workforce in a hospital for the safe worker-to-patient ratio. These decisions should be made without knowing the future service duration for each appointment, the exact number of patients incoming to the hospital, and no-show behaviors of patients and/or workers. To provide decisions that aware of these uncertainties, numerous SP models have been proposed under the assumption that the accurate probability distribution of the system uncertainties is known [34, 35, 36, 37]. For example, random service durations in an appointment scheduling problem are assumed to follow Gamma, uniform, and normal distributions [38], exponential distribution [39], and log-normal distribution [40]. This assumption no longer holds as hospital operators consider more types of uncertainties in the system. In other words, obtaining the accurate joint probability distribution of the uncertain parameters becomes even more challenging. Taking this viewpoint, researchers have studied a number of DRO models in healthcare operations. Based on the moment-based ambiguity set described in Section 1.1.2, Kong et al.[41] and Mak et al. [42] studied appointment scheduling problem under uncertain service duration, and these works were extended by Kong et al. [43] and Jiang et al. [44] by incorporating the uncertain patient no-show behaviors into consideration. Additionally, Davis et al. [45] studied the nurse staffing problem under the demand uncertainty. In our study in Chapter 2, we consider the appointment scheduling problem under uncertain service duration and patient no-shows and formulate the problem using a Wasserstein-based DRO model as described in Section 1.1.1. Also, in Chapter 3, we study the nurse staffing problem under uncertain nurse absenteeism, which depends on the staffing decisions, and propose a DRO model under the decision-dependent moment-based ambiguity set as described in Section 1.1.2.

Next, *energy systems* refer to the application area, where an independent system operator (ISO) makes planning, scheduling, and response decisions on the energy generation, transmission, and distribution for effective operations, which relates to not only the operational costs but also the reliability of the system. The followings are examples of optimization problems in this area; (i) *commitment problem* seeks to determine a set of generators and transmission lines to activate so as to satisfy the electricity demand, (ii) *expansion planning problem* aims to decide a set of generating units and transmission lines to expand for enhancing the capacity and reliability of the power system. These decisions should be made without knowing the future electricity demand. Indeed one can predict the future demand and incorporate it into an optimization problem. Alternatively, one can apply the SP approach under the assumption that the probability distribution of the demand is known [46, 47]. As the uncertainty in the system increases with the growing penetration of renewable energy resources, such as wind and solar power, it becomes even more challenging to estimate the probability distribution of the various uncertain parameters accurately. This led to work like [47] that utilized bivariate normal distribution for power load and wind uncertainties. On the other hand, RO approaches were introduced to make robust decisions under uncertainty (see, e.g., [48], [49]). For example, [49] studied robust transmission network expansion planning under uncertain renewable generation and loads. By noting that the solution obtained by the RO approach could be too conservative, there have been several studies that proposed the DRO approach. Based on the discrepancy-based DRO approach as described in Section 1.1.1, Yao et al. [50] studied energy scheduling, Duan et al. [51] studied chance-constrained optimal power flow, and Zhu et al. [52] studied the unit commitment problem. Based on the moment-based DRO approach as described in Section 1.1.2, Zhang et al. [53] studied chance-constrained optimal power flow, Zhao and Jiang [54] studied contingency-constrained unit commitment problem, and Velloso et al. [55] studied transmission expansion planning. In our study in Chapter 4, we study how to operate the transmission grid under uncertain natural disasters based on the moment-based DRO approach.

Not only limited to the applications mentioned earlier, but also the statistical learning community is widely adopting the DRO techniques. For example, the Wasserstein-based DRO models have been widely studied in the context of the maximum likelihood estimation model [56], support vector machines [57], and logistic regression [58].

1.3 Dissertation Overview

In this section, we overview the rest of this dissertation.

In Chapter 2, we study a single-server appointment scheduling problem with a fixed sequence of appointments, for which we must determine the arrival time for each appointment. We specifically examine two stochastic models. In the first model, we assume that all appointees show up at the scheduled arrival times, yet their service durations are random. In the second model, we assume that appointees have random no-show behaviors, and their service durations are random given that they show up at the appointments. In both models, we assume that the probability distribution of the uncertain parameters is unknown but can be partially observed via a set of historical data, which we view as independent samples drawn from the actual unknown distribution. In view of the distributional ambiguity, we propose a data-driven DRO approach based on Wasserstein balls to determine an appointment schedule such that the worst-case (i.e., maximum) expectation of the system's total cost is minimized. A key feature of this approach is that the optimal value and the set of optimal schedules thus obtained provably converge to those of the "true" model, i.e., the stochastic appointment scheduling model with regard to the true probability distribution of the uncertain parameters. While our DRO models are computationally intractable in general, we reformulate them to copositive programs, which are amenable to tractable semidefinite programming problems with high-quality approximations. Furthermore, under some mild conditions, we recast these models as polynomial-sized linear programs or second-order conic programs. Through an extensive numerical study, we demonstrate that our approach yields better out-of-sample performance than two state-of-the-art methods.

In Chapter 3, we study a nurse staffing problem under random nurse demand and absenteeism. While the demand uncertainty is exogenous (stemming from the random patient census), the absenteeism uncertainty is *endogenous*, i.e., the number of nurses who show up for work partially depends on the nurse staffing level. For the quality of care, many hospitals have developed float pools of nurses by cross-training, so that a pool nurse can be assigned to the units short of nurses. In this chapter, we propose a distributionally robust nurse staffing (DRNS) model that considers both exogenous and endogenous uncertainties. We derive a separation algorithm that solves the proposed DRO model under an arbitrary structure of float pools. In addition, we identify several pool structures that often arise in practice and recast the corresponding DRNS model as a monolithic mixed-integer linear program (MILP), which facilitates off-the-shelf commercial solvers. Furthermore, we optimize the float pool design to reduce the cross-training while achieving a specified target staffing costs. The numerical case studies, based on the data of a collaborating hospital, suggest that the units with high absenteeism probability should be pooled together.

In Chapter 4, we study the transmission grid operation under uncertain geomagnetic disturbances (GMD). GMD is one of the natural disasters, mainly caused by a solar storm.

It poses a severe risk to power systems as it drives geomagnetically-induced currents (GIC), which can saturate the cores of transformers, induce hot-spot heating, and increase reactive power losses. As a way to enhance resiliency to GMD, one can establish a damage prevention plan via existing controls in the transmission grid, such as generator dispatch, line switching, and load shedding. These corrective actions should be determined before knowing the exact magnitude and orientation of GMD, which could be challenging due to the insufficient historical data. In this chapter, we propose a DRO approach that models uncertain GMD based on the *Markov* ambiguity set. The proposed DRO model aims to find a set of corrective actions such that the worst-case expectation of the system cost is minimized. For the solution approach, we propose a column-and-constraint generation (CCG) algorithm that solves a sequence of mixed-integer second-order conic programs (MISOCPs) to handle the underlying convex support set of the uncertain GMD. As a special case, when the support set is a polytope with a finite number of selected extreme points, we derive a monolithic MISOCP, which can provide a lower bound to the original problem. Based on this observation, we develop a modified CCG algorithm which utilizes the monolithic MISOCP reformulation to enhance computation. Numerical experiments on 'epri-21' and 'uiuc-150' systems show the efficacy of the proposed algorithms and the exact reformulation of our DRO model.

In Chapter 5, we conclude this dissertation and discuss future research directions.

CHAPTER 2

Data-Driven Distributionally Robust Appointment Scheduling over Wasserstein Balls

2.1 Introductory Remarks

Since the pioneering work of Bailey [59], appointment systems have been extensively studied in many customer service industries with the aim of increasing resource utilization, matching supply and demand, and smoothing customer flows. The core operational activity in many appointment systems is to schedule arrival times for appointments to minimize the system total cost associated with appointment waiting, as well as the server's idleness and overtime. Serving as a central modeling component in a wide range of applications, appointment scheduling (AS) has been applied in outpatient scheduling [60, 61, 39, 41], surgery planning [38], call-center staffing [62], and cloud computing server operations [63]. In this chapter, we study a single-server AS problem where the number and sequence of the appointments are fixed. The main decision in this problem is to schedule the arrival time for each appointment, which helps design the scheduling template of the appointment system.

There are several sources of variability that make AS problems challenging to solve, including random service durations, random appointee no-shows, unpunctuality, and emergency interruptions. In this study, we focus on random service durations and random appointee no-shows. Due to random service durations, an appointment can complete before or after the scheduled starting time of the subsequent appointment and result in the server's idleness or the waiting of the subsequent appointment, respectively. In addition, if the last appointment is completed after the pre-determined time limit, then the server has to work overtime, which is often costly and unpleasant for service providers. Random appointee no-shows arise in many appointment systems, e.g., outpatient clinics. No-shows often results in idleness of resources (e.g., equipment and personnel) and so the loss of opportunities for serving other appointments. In fact, it has been reported that random no-shows have more impact on the performance of an appointment system than random durations (see, e.g., [61]). Hence, it is recommended that we adapt AS models to the anticipated no-show behaviors [64].

In the literature, *stochastic programming* (SP) [65, 66] approaches are often proposed to tackle AS problems given that the true distribution of the service durations and no-shows is fully known [34, 67, 68]. While SP exhibits superior modeling power, it has two inherent shortcomings. First, SP suffers from the notorious curse of dimensionality. Indeed, the computation of expectations necessitates the evaluation of multi-dimensional integrals, which is in general intractable. Second, it is challenging and sometimes impossible to accurately estimate the true distribution. For example, the raw data of the uncertain parameters can typically be explained by many strikingly different distributions. Using a biased estimate of the distribution, the SP approach can yield overfitted decisions, which display an optimistic bias and can lead to post-decision disappointment. For example, if one simply uses the empirical distribution based on the raw data, the obtained solution often results in unpleasant out-of-sample performance.

In view of the distributional ambiguity, one can construct a so-called *ambiguity set* to contain all possible distributions that may govern the generation of the observed data samples. With the ambiguity set, one can formulate a distributionally robust optimization (DRO) problem with the goal of minimizing the worst-case (i.e., maximum) expected system total cost of the appointment system, where the expectations are taken with respect to the distributions from the ambiguity set. The majority of DRO approaches considered in the literature are based on moment ambiguity sets, which consist of all distributions sharing certain moments, e.g., the first and second moments. Although a moment ambiguity set often leads to a tractable optimization problem [69, 44, 41, 42], it typically does not converge to the true distribution even in the situation where more data can be obtained. This gives rise to the central questions of this study: Is there an alternative DRO approach that extracts more information of the underlying true distribution from available (possibly small-size) historical data? If such an approach exists, is the resulting AS model solvable in polynomial time?

In this chapter (see more details in [70]), we endeavor to give affirmative answers to these two research questions. In particular, we propose to construct the ambiguity set using a Wasserstein ball centered at the empirical distribution based on the historical data [11, 13, 71]. Accordingly, we consider two Wasserstein-based distributionally robust appointment scheduling (W-DRAS) models. In the first model, we examine the situation where all appointees show up at their scheduled starting times but their service durations are random. In the second model, we incorporate both random no-shows and random service durations of the appointees (if they show up). Results of the modern measure concentration theory guarantee that the Wasserstein ball has asymptotic consistency, which ensures that the optimal value as well as the optimal solutions of the W-DRAS models converge to their SP counterparts with regard to the true distribution, as the data size tends to infinity. While the resulting optimization problems are, in general, intractable, we reformulate them to copositive programs, which admit tractable semidefinite programs that have high-quality approximations. Furthermore, under some mild conditions, we reformulate the W-DRAS problems to polynomial-sized linear programs or second-order conic programs, which can be efficiently solved by many off-the-shelf optimization solvers.

2.1.1 Literature review

Various methodologies have been applied to formulate and solve the AS problem, including queueing theory (see, e.g., [72]), approximation algorithm (see, e.g., [39]), and optimization. We refer the readers to [73, 74, 64, 68] for comprehensive summaries of these studies. In this section, we conduct a brief literature review on the most relevant literature to the proposed approach.

The SP models of the AS problem assume that the probability distribution of the uncertain parameters is fully known and seek a schedule to minimize the expectation of the system total cost. Begen and Queyranne [35] show for the first time that the AS problem, where the random service durations follow a joint discrete probability distribution, could be solved in polynomial time under some mild conditions. Ge et al. [75] extend the result of [35] to the case where the cost is modeled as piecewise linear convex functions of the waiting time and idleness.

When it comes to general probability distributions, exact calculation of the multi-dimensional integral poses computational challenges. To this end, sample average approximation (SAA) methods are often used to approximately solve AS problems. Denton and Gupta [38] formulate the AS problem as a two-stage stochastic linear program and propose a sequential bounding approach to determine upper bounds on the optimal value. In a more recent work, Begen et. al. [76] propose a sampling-based approach, whereby one can construct an empirical distribution over a set of historical data and quantify the related computation complexity to obtain a near-optimal solution in terms of sample size. However, SAA approaches often lead to optimistic bias, which motivates us to consider a DRO approach in this chapter.

Modeling the no-show behavior in appointment scheduling systems is even more challenging due to the discrete nature of the no-show parameters. For example, the no-show of an appointee is often modeled by a Bernoulli random variable which equals one if the appointee shows up and zero if the appointee does not show up. Ho and Lau [61] take a first step to develop a heuristic approach that suggests to double book the first two appointments and subsequently schedule the remaining appointments. Following the work in [61], a number of more advanced yet more sophisticated approximation and heuristic approaches have been proposed to address the random no-show issue in various settings [40, 38, 67, 77, 78].

In reality, it is often difficult to fit the true distribution of the service durations and no-shows for various reasons, e.g., lack of sufficient data [79] and the existence of correlations [64]. Robust approaches have been proposed to address this challenge based on partial information of the distribution. In particular, the classical robust optimization approach (see, e.g., [80, 81]) models the uncertain parameters based only on an uncertainty set (e.g., the support or a confidence set of the uncertainty). Differently, the DRO approach employs an ambiguity set that incorporates a family of distributions. We refer the readers to the classical works [82, 10, 11, 30, 6] and references therein for general DRO models and solution approaches. Closely related to this research are the following papers. Kong et el. [41] propose a DRO model over a cross-moment ambiguity set consisting of all distributions with common mean and covariance of the random service durations. They recast this model as a copositive program and solve its semidefinite approximation to obtain upper bounds on the optimal value. Differently, Mak et el. [42] consider a marginal-moment ambiguity set consisting of all distributions with common marginal moments up to a finite order. They recast the corresponding DRO model as a semidefinite program for general marginal-moment ambiguity sets and, in particular, a second-order cone program for the mean-variance ambiguity set and a linear program for the mean-support ambiguity set. Recently, Jiang et el. [44] study a mean-support ambiguity set for both random no-shows and service durations. As the no-show parameters are discrete, they propose an integer programming reformulation and develop a decomposition algorithm to solve the resulting mixed integer programs. Kong et el. [43] consider a cross-moment ambiguity set with decision-dependent no-shows, i.e., the first and second moments of the no-shows depend on the appointment arrival times. They propose a copositive programming reformulation for the corresponding DRO model and develop an algorithm to search for an optimal schedule by iteratively solving a series of semidefinite programs. In contrast to the aforementioned work that consider moment ambiguity sets, we propose a Wasserstein-based ambiguity set that enjoys asymptotic consistency (see Section 4.4 for a demonstration based on a *finite* data size).

The sequence of the appointments is assumed to be fixed in this chapter. In the literature, determining the best sequence of appointments for a set of heterogeneous appointments is an interesting topic; see [83, 68, 84, 85, 42]. In some situations, appointment systems are allowed to overbook a time slot with multiple appointments. To this end, [86, 87, 88] propose various approaches to provide optimal overbooking policies. Multiple-server AS has also been considered in the literature; see for example [89].

2.1.2 Our contributions

We highlight the main contributions of this chapter as follows.

1. We propose a data-driven DRO approach for the AS problem with random service durations and random no-shows over Wasserstein balls. To the best of our knowledge, this is the first DRO approach applied to the AS problem that enjoys the asymptotic consistency.

2. We make technical contribution to the DRO literature. The W-DRAS model with random service durations is a two-stage DRO problem with random right-hand sides. When we further incorporate the no-shows, randomness arises from both the right-hand sides and the objective coefficients. While two-stage DRO problems are in general NP-hard (see [31]), we recast both W-DRAS models as copositive programs, which are amenable to tractable semidefinite programming approximations. More importantly, under mild conditions, we recast both W-DRAS models as convex programs that are solvable in polynomial time. In particular, if the Wasserstein ball is characterized by an ℓ_p-norm, we show that W-DRAS admits a linear programming reformulation when $p = 1$ and a second-order conic reformulation when $p > 1$ and p is rational.

3. We derive probability distributions that attain the worst-case expected total cost in the W-DRAS models. These distributions can be applied to stress test an appointment schedule generated from any decision-making processes. In addition, when we interpret W-DRAS as a two-person game between the AS scheduler and the nature, these distributions provide interesting insights on how the nature plays against a given appointment schedule.

4. We demonstrate the effectiveness of our approach over diverse test instances. In particular, our approach yields (i) near-optimal appointment schedules even with a small data size and (ii) better out-of-sample performance than two state-of-the-art methods, even when the distribution is misspecified. This demonstrates that the W-DRAS approach is effective in AS systems (i) with scarce data and/or (ii) in a quickly varying environment.

2.1.3 Structure and notation

The remainder of the chpater is organized as follows. We study the W-DRAS problem with random service durations in Section 2.2 and extend the model to incorporate both random no-shows and random service durations in Section 2.3. We test the effectiveness of our approach over diverse test instances in Section 4.4. Finally, we conclude and discuss future research directions in Section 2.5. We relegate all technical proofs to the Appendix.

Notation: For any $m \in \mathbb{N}$, we define $[m]$ as the set of running indices $\{1, \ldots, m\}$. For $i, j \in \mathbb{N}$, we let $[i, j]_{\mathbb{Z}}$ represent the set of running indices $\{i, i+1, \ldots, j\}$. We use boldface notation to denote vectors. In particular, we denote by $\mathbf{e}$ the vector of all ones and by $\mathbf{e}_i$ the i-th standard basis vector. Finally, we denote by $\mathbb{S}^n$ the set of all symmetric matrices in $\mathbb{R}^{n \times n}$.

2.2 Random Service Durations

In this section, we study the W-DRAS model with random service durations. We describe the model, including the Wasserstein ambiguity set, in Section 2.2.1 and reformulate it as a copositive program in Section 2.2.2. In Section 2.2.3, we recast the model as tractable convex programs under a rectangularity condition.

2.2.1 W-DRAS model and Wasserstein ambiguity set

We consider n appointments and determine a time allowance s_i, or equivalently an arrival time, for each appointment $i \in [n]$. We require that all appointments are scheduled to arrive by a fixed end-of-the-day time limit T, which gives rise to the following feasible region for $\boldsymbol{s} := (s_1, \ldots, s_n)^\top$:

$$\mathcal{S} := \left\{ \boldsymbol{s} \in \mathbb{R}^n : \boldsymbol{s} \geq 0,\ \sum_{i=1}^{n} s_i \leq T \right\}.$$

As each appointment i takes up a random service duration u_i, one or multiple of the following three scenarios can happen: (i) an appointment cannot start on time due to a delay of completion of the previous appointment, (ii) the server is idle and waiting for the next appointment due to an early completion of the current appointment, and (iii) the server cannot finish serving all the appointments by T. Specifically, if we let w_i represent the waiting time of appointment i, v_i represent the server's idleness after serving appointment i, and w_{n+1} represent the server's overtime beyond T, then we can recursively express the waiting times, as well as the server's idleness and overtime by $w_1 = 0$ and

$$w_i = \max\Big\{0,\ u_{i-1} + w_{i-1} - s_{i-1}\Big\}, \quad v_{i-1} = \max\Big\{0,\ s_{i-1} - u_{i-1} - w_{i-1}\Big\} \tag{2.1}$$

for all $i \in [2, n+1]_{\mathbb{Z}}$. Note that $w_1 = 0$ because the first appointment always starts on time. To evaluate the performance of the appointment schedule $\boldsymbol{s}$, we denote the unit costs of waiting, idleness, and overtime by $\boldsymbol{c} \in \mathbb{R}^n_+$, $\boldsymbol{d} \in \mathbb{R}^n_+$, and $C \in \mathbb{R}_+$ respectively. In addition, we assume that $d_{i+1} - d_i \leq c_{i+1}$ for all $i \in [n-1]$. This assumption is standard in the literature (see, e.g., [38, 75, 41, 42]). If an appointment system fails to satisfy this assumption, the server would (intentionally) keep idle and make appointee $i+1$ wait after completing all previous appointments, which is unrealistic in practice. Under this assumption, $\boldsymbol{w} := (w_1, \ldots, w_{n+1})^\top$ and $\boldsymbol{v} := (v_1, \ldots, v_n)^\top$ can be obtained from the following linear program:

$$\begin{aligned} f(\boldsymbol{s}, \boldsymbol{u}) := \min_{\boldsymbol{w}, \boldsymbol{v}} \quad & \sum_{i=1}^{n} (c_i w_i + d_i v_i) + C w_{n+1} \\ \text{s.t.} \quad & w_i - v_{i-1} = u_{i-1} + w_{i-1} - s_{i-1} \quad \forall\, i \in [2, n+1]_{\mathbb{Z}} \\ & \boldsymbol{w} \geq 0,\ w_1 = 0,\ \boldsymbol{v} \geq 0, \end{aligned} \tag{2.2}$$

where $\boldsymbol{u} := (u_1, \ldots, u_n)^\top$ and $f(\boldsymbol{s}, \boldsymbol{u})$ represents the system total cost, which is evaluated by a weighted sum of the waiting times, idleness, and overtime under schedule $\boldsymbol{s}$ and service durations $\boldsymbol{u}$. If the probability distribution of $\boldsymbol{u}$, denoted by $\mathbb{P}_{\boldsymbol{u}}$, is fully known, classical stochastic AS approaches seek a schedule that minimizes the expected system total cost, i.e.,

$$Z^\star := \min_{\boldsymbol{s} \in \mathcal{S}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{u}}}\big[f(\boldsymbol{s}, \boldsymbol{u})\big]. \tag{2.3}$$

In this chapter, we assume that $\mathbb{P}_{\boldsymbol{u}}$ is unknown and belongs to a Wasserstein ambiguity set. Specifically, suppose that two probability distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are defined over a common support set $\mathcal{U} \subseteq \mathbb{R}^n$ and, with $p \geq 1$, $\|\cdot\|_p$ represents the p-norm on $\mathbb{R}^n$. Then, the Wasserstein distance $d_p(\mathbb{Q}_1, \mathbb{Q}_2)$ between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is the minimum transportation cost of moving from $\mathbb{Q}_1$ to $\mathbb{Q}_2$, under the premise that the cost of moving from $\boldsymbol{u}_1$ to $\boldsymbol{u}_2$ amounts to $\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_p$. Mathematically,

$$d_p(\mathbb{Q}_1, \mathbb{Q}_2) := \left(\inf_{\Pi \in \mathcal{P}(\mathbb{Q}_1, \mathbb{Q}_2)} \mathbb{E}_\Pi \Big[\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_p^p \Big] \right)^{1/p}$$

where random vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ follow $\mathbb{Q}_1$ and $\mathbb{Q}_2$ respectively, and $\mathcal{P}(\mathbb{Q}_1, \mathbb{Q}_2)$ represents the set of all joint distributions of $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$.

In addition, we assume that $\mathbb{P}_{\boldsymbol{u}}$ is only observed via a finite set of N i.i.d. samples, denoted as $\{\widehat{\boldsymbol{u}}^1, \dots, \widehat{\boldsymbol{u}}^N\}$. For example, these samples can come from the historical data of the service durations $\boldsymbol{u}$. Then, we consider the following p-Wasserstein ambiguity set

$$\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{u}},\, \epsilon) := \left\{ \mathbb{Q}_{\boldsymbol{u}} \in \mathcal{P}(\mathcal{U}) :\ d_p(\mathbb{Q}_{\boldsymbol{u}},\, \widehat{\mathbb{P}}^N_{\boldsymbol{u}}) \leq \epsilon \right\},$$

where $\mathcal{P}(\mathcal{U})$ represents the set of all probability distributions on $\mathcal{U}$, $\widehat{\mathbb{P}}^N_{\boldsymbol{u}}$ represents the empirical distribution of $\boldsymbol{u}$ based on the N i.i.d. samples, i.e., $\widehat{\mathbb{P}}^N_{\boldsymbol{u}} = \frac{1}{N}\sum_{j=1}^N \delta_{\widehat{\boldsymbol{u}}^j}$, and $\epsilon > 0$ represents the radius of the ambiguity set. We seek a schedule that minimizes the expected total cost with regard to the worst-case distribution in the p-Wasserstein ambiguity set, i.e., we solve the following DRO problem:

$$\widehat{Z}(N,\, \epsilon) := \min_{\boldsymbol{s} \in \mathcal{S}} \sup_{\mathbb{Q}_{\boldsymbol{u}} \in \mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{u}},\, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}\big[f(\boldsymbol{s},\, \boldsymbol{u})\big]. \tag{W-DRAS}$$

Many data-driven applications desire asymptotic consistency. In particular, an appointment scheduler may desire that, as the sample size N increases to infinity, $\widehat{Z}(N,\, \epsilon)$ tends to $Z^\star$, which is the optimal value of the "true" model (2.3), i.e., the model with perfect knowledge of $\mathbb{P}_{\boldsymbol{u}}$. Accordingly, an optimal appointment schedule obtained from (W-DRAS) tends to the optimal schedule obtained from (2.3). In addition, if $\widehat{Z}(N,\, \epsilon) \geq Z^\star$ almost surely, then (W-DRAS) provides a safe (upper bound) guarantee on the expected total cost with any *finite* data size N. We close this section by formally establishing the asymptotic consistency and the finite-data guarantee of (W-DRAS). To this end, we make the following assumption on the support set $\mathcal{U}$.

Assumption 1. *The support set $\mathcal{U}$ is nonempty, compact, and convex.*

This assumption is mild and it holds in most practical situations. The following concentration inequality on the p-Wasserstein ambiguity set is adapted from [90].

Lemma 1 (Adapted from Theorem 2 in [90])**.** *Suppose that Assumption 1 holds. Then there exist nonnegative constants c_1 and c_2 such that, for all $N \geq 1$ and $\beta \in (0, \min\{1, c_1\})$,*

$$\mathbb{P}^N_{\boldsymbol{u}} \left\{ d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}^N_{\boldsymbol{u}}) \leq \epsilon_N(\beta) \right\} \geq 1 - \beta,$$

where $\mathbb{P}^N_{\boldsymbol{u}}$ represents the product measure of N copies of $\mathbb{P}_{\boldsymbol{u}}$ and

$$\epsilon_N(\beta) \;=\; \left[\frac{\log(c_1\beta^{-1})}{c_2 N}\right]^{\frac{1}{\max\{n, 3p\}}}.$$

Lemma 1 assures that the p-Wasserstein ambiguity set incorporates the true distribution

$\mathbb{P}_{\boldsymbol{u}}$ with high confidence. This leads to the following two theorems, whose proofs rely on [11] and are provided in the Appendix for completeness.

Theorem 1 (Asymptotic consistency, adapted from Theorem 3.6 in [11])**.** *Suppose that Assumption 1 holds. Consider a sequence of confidence levels* $\{\beta_N\}_{N\in\mathbb{N}}$ *such that* $\sum_{N=1}^{\infty} \beta_N < \infty$ *and* $\lim_{N\to\infty} \epsilon_N(\beta_N) = 0$[1]*, and let* $\widehat{\boldsymbol{s}}(N, \epsilon_N(\beta_N))$ *represent an optimal solution to* (W-DRAS) *with the ambiguity set* $\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon_N(\beta_N))$*. Then,* $\mathbb{P}^\infty_{\boldsymbol{u}}$*-almost surely we have* $\widehat{Z}(N, \epsilon_N(\beta_N)) \to Z^\star$ *as* $N \to \infty$*. In addition, any accumulation point of* $\{\widehat{\boldsymbol{s}}(N, \epsilon_N(\beta_N))\}_{N\in\mathbb{N}}$ *is an optimal solution of* (2.3) $\mathbb{P}^\infty_{\boldsymbol{u}}$*-almost surely.*

Theorem 2 (Finite-data guarantee, adapted from Theorem 3.5 in [11])**.** *For any* $\beta \in (0, 1)$*, let* $\widehat{\boldsymbol{s}}(N, \epsilon_N(\beta))$ *represent an optimal solution of* (W-DRAS) *with the ambiguity set* $\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon_N(\beta))$*. Then,* $\mathbb{P}^N_{\boldsymbol{u}}\{\mathbb{E}_{\mathbb{P}_{\boldsymbol{u}}}[f(\widehat{\boldsymbol{s}}(N, \epsilon_N(\beta)), \boldsymbol{u})] \le \widehat{Z}(N, \epsilon_N(\beta))\} \ge 1 - \beta$*.*

The radius $\epsilon_N(\beta)$ suggested in the above theoretical results is often conservative, i.e., the actual radius ϵ that achieves the $(1 - \beta)$ confidence of $\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)$ may be much smaller than $\epsilon_N(\beta)$. In this chapter, we calibrate the radius of the p-Wasserstein ambiguity set by using the cross validation method. This yields encouraging convergence results and out-of-sample performance of (W-DRAS) (see the detail in Section 4.4).

Problem (W-DRAS) is generically difficult to solve as it involves infinitely many probability distributions. In the following two subsections, we shall recast (W-DRAS) as a (deterministic and finite) copositive program and identify conditions under which we have more tractable reformulations.

2.2.2 Copositive programming reformulations

In this section, we propose a copositive reformulation for (W-DRAS). As the first step, we represent $f(\boldsymbol{s}, \boldsymbol{u})$ by the following dual problem of (2.2):

$$f(\boldsymbol{s}, \boldsymbol{u}) = \max_{y\in\mathcal{Y}} \sum_{i=1}^{n} (u_i - s_i) y_i, \tag{2.4}$$

where dual variables $\boldsymbol{y}$ are associated with the first constraint in (2.2) and the polyhedral feasible set $\mathcal{Y}$ is described as

$$\mathcal{Y} := \left\{ \boldsymbol{y} \in \mathbb{R}^n : \begin{array}{ll} -y_i \le d_i & \forall\, i \in [n] \\ y_{i-1} - y_i \le c_i & \forall\, i \in [2, n]_{\mathbb{Z}} \\ y_n \le C & \end{array} \right\}. \tag{2.5}$$

[1]For example, we can set $\beta_N = \exp\{-\sqrt{N}\}$.

The strong duality between (2.2) and (2.4) holds because $\mathcal{Y}$ is nonempty and compact, as summarized in the following lemma.

Lemma 2. *$\mathcal{Y}$ is nonempty and compact.*

Second, to obtain a copositive reformulation, we consider the inner maximization problem of (W-DRAS)

$$\sup_{\mathbb{Q}_{\boldsymbol{u}} \in \mathcal{D}_p(\widehat{\mathbb{P}}_{\boldsymbol{u}}^N, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}\big[f(\boldsymbol{s}, \boldsymbol{u})\big] \tag{2.6}$$

for fixed $\boldsymbol{s} \in \mathcal{S}$. In the following proposition, we present an equivalent dual formulation of (2.6) via duality theory.

Proposition 1. *The optimal value of formulation* (2.6) *equals that of the following formulation:*

$$\begin{aligned} \inf_{\rho, \boldsymbol{\theta}} \ & \epsilon^p \rho + \frac{1}{N} \sum_{j=1}^{N} \theta_j \\ \textit{s.t.} \ & \sup_{\boldsymbol{u} \in \mathcal{U}} \big\{ f(\boldsymbol{s}, \, \boldsymbol{u}) - \rho \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p \big\} \leq \theta_j \quad \forall\, j \in [N] \\ & \rho \geq 0, \ \boldsymbol{\theta} \in \mathbb{R}^N. \end{aligned} \tag{2.7}$$

While this dual formulation is in the format of a classical robust optimization problem, it is inherently difficult. Indeed, reformulating the semi-infinite constraint (2.7) entails solving N non-convex optimization problems (note that both $f(\boldsymbol{s}, \boldsymbol{u})$ and $\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p$ are convex in $\boldsymbol{u}$), which are in general computationally intractable. The presence of the norm $\|\cdot\|_p^p$ with a general $p \geq 1$ introduces additional computational difficulties.

In this section, we propose to use copositive programming reformulation techniques to tackle the intractable constraint for the cases of $p = 1, 2$. For ease of exposition, we define sets $\mathcal{F}_1^j$ for all $j \in [N]$ and $\mathcal{F}_2$ by using $\mathcal{U}$ and $\mathcal{Y}$:

$$\mathcal{F}_1^j := \Big\{ \ (\boldsymbol{u}^+, \boldsymbol{u}^-, \boldsymbol{y}) \in \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{R}^n : \boldsymbol{u}^+ - \boldsymbol{u}^- + \widehat{\boldsymbol{u}}^j \in \mathcal{U}, \, \boldsymbol{y} \in \mathcal{Y} \ \Big\},$$

$$\mathcal{F}_2 := \Big\{ (\boldsymbol{u}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n : \ \boldsymbol{u} \in \mathcal{U}, \ \boldsymbol{y} \in \mathcal{Y} \ \Big\},$$

and then construct their perspective sets $\mathcal{K}_1^j$ and $\mathcal{K}_2$ respectively:

$$\mathcal{K}_1^j := \text{closure}\Big(\Big\{ (t, \boldsymbol{u}^+, \boldsymbol{u}^-, \boldsymbol{y}) : (\boldsymbol{u}^+/t, \boldsymbol{u}^-/t, \boldsymbol{y}/t) \in \mathcal{F}_1^j, \ t > 0 \Big\} \Big), \tag{2.8}$$

$$\mathcal{K}_2 := \text{closure}\Big(\Big\{ (t, \boldsymbol{u}, \boldsymbol{y}) : (\boldsymbol{u}/t, \boldsymbol{y}/t) \in \mathcal{F}_2, \ t > 0 \Big\} \Big), \tag{2.9}$$

where closure($\mathcal{K}$) denotes the closure of the set $\mathcal{K}$. In fact, $\mathcal{K}_1^j$ and $\mathcal{K}_2$ are closed and convex cones by Assumption 1. Furthermore, for a closed and convex cone $\mathcal{K}$, we define $\mathcal{COP}(\mathcal{K})$ as the set of all copositive matrices with respect to $\mathcal{K}$, i.e., $\mathcal{COP}(\mathcal{K}) := \left\{\boldsymbol{M} \in \mathbb{S}^n : \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} \geq 0 \, \forall \, \boldsymbol{x} \in \mathcal{K}\right\}$. The dual cone of $\mathcal{COP}(\mathcal{K})$ is denoted by $\mathcal{CP}(\mathcal{K})$, which is the set of all completely positive matrices with respect to $\mathcal{K}$. We refer the reader to [91, 92, 93, 94] for a thorough discussion about copositive programming.

Now, we are ready to present the copositive programming reformulation of (W-DRAS) for the cases of $p = 1, 2$ in the definition of $\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)$. For fixed ρ and $\boldsymbol{s}$, define

$$\boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) := \begin{bmatrix} 0 & -\frac{1}{2}\rho\boldsymbol{e}^\top & -\frac{1}{2}\rho\boldsymbol{e}^\top & \frac{1}{2}(\widehat{\boldsymbol{u}}^j - \boldsymbol{s})^\top \\ -\frac{1}{2}\rho\boldsymbol{e} & 0 & 0 & \frac{1}{2}\boldsymbol{I} \\ -\frac{1}{2}\rho\boldsymbol{e} & 0 & 0 & -\frac{1}{2}\boldsymbol{I} \\ \frac{1}{2}(\widehat{\boldsymbol{u}}^j - \boldsymbol{s}) & \frac{1}{2}\boldsymbol{I} & -\frac{1}{2}\boldsymbol{I} & 0 \end{bmatrix} \quad \forall\, j \in [N],$$

$$\boldsymbol{H}_j^2(\rho,\, \boldsymbol{s}) := \begin{bmatrix} -\rho\|\widehat{\boldsymbol{u}}^j\|^2 & \rho(\widehat{\boldsymbol{u}}^j)^\top & -\frac{1}{2}\boldsymbol{s}^\top \\ \rho\widehat{\boldsymbol{u}}^j & -\rho\boldsymbol{I} & \frac{1}{2}\boldsymbol{I} \\ -\frac{1}{2}\boldsymbol{s} & \frac{1}{2}\boldsymbol{I} & 0 \end{bmatrix} \quad \forall\, j \in [N],$$

where $\boldsymbol{I} \in \mathbb{S}^n$ and $\boldsymbol{e} \in \mathbb{R}^n$ denote by the identity matrix and all-ones vector respectively.

Theorem 3. *Suppose that Assumption 1 holds. Then, when $p = 1$,* (W-DRAS) *yields the same optimal value and the same set of optimal solutions as the following linear conic program:*

$$\begin{aligned} \widehat{Z}(N,\, \epsilon) = \inf \quad & \epsilon\rho + \frac{1}{N}\sum_{j=1}^N \beta_j \\ \text{s.t.} \quad & \rho \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^N \\ & \beta_j\, \boldsymbol{e}_1\boldsymbol{e}_1^\top - \boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}_1^j) \quad \forall\, j \in [N] \\ & \rho \geq 0,\ \boldsymbol{s} \in \mathcal{S}, \end{aligned} \tag{2.10}$$

where $\boldsymbol{e}_1$ represents the first standard basis vector in $\mathbb{R}^{3n+1}$.

In addition, when $p = 2$, (W-DRAS) *yields the same optimal value and the same set of optimal solutions as the following linear conic program:*

$$\begin{aligned} \widehat{Z}(N,\, \epsilon) = \inf \quad & \epsilon^2\rho + \frac{1}{N}\sum_{j=1}^N \beta_j \\ \text{s.t.} \quad & \rho \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^N \\ & \beta_j\, \boldsymbol{e}_1\boldsymbol{e}_1^\top - \boldsymbol{H}_j^2(\rho,\, \boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}_2) \quad \forall\, j \in [N] \\ & \rho \geq 0,\ \boldsymbol{s} \in \mathcal{S}, \end{aligned} \tag{2.11}$$

where $\boldsymbol{e}_1$ represents the first standard basis in $\mathbb{R}^{2n+1}$.

Remark 1. *The copositive programming reformulations in* (2.10) *and* (2.11) *are amenable to semidefinite programming solution schemes. Specifically, there exists a hierarchy of increasingly tight semidefinite-based inner approximations that converge to $\mathcal{COP}(\mathcal{K})$ for a general closed and convex cone $\mathcal{K}$ [95, 96, 97, 98, 99]. Replacing the cone $\mathcal{COP}(\mathcal{K})$ with these inner approximations gives rise to conservative semidefinite programs that can be solved using standard off-the-shelf solvers (such as MOSEK [100] and SDPT3 [101]).*

2.2.3 Tractable reformulations

In this subsection, we consider a setting that admits tractable reformulations of (W-DRAS) for general $p \geq 1$. In particular, we recast problem (W-DRAS) as a linear program when $p = 1$ and as a second-order cone program when $p > 1$ and p is rational. To this end, we make the following assumption on the support set $\mathcal{U}$.

Assumption 2 (Rectangular support of service durations)**.** *Assume that $\mathcal{U}$ is defined as*

$$\mathcal{U} := \Big\{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{u}^{L} \leq \boldsymbol{u} \leq \boldsymbol{u}^{U}\Big\},$$

for $0 \leq u_i^L < u_i^U < \infty$ for all $i \in [n]$.

Assumption 2 is mild because any compact $\mathcal{U} \subset \mathbb{R}^n_+$ can be relaxed to be rectangular. In addition, the asymptotic consistency and the finite-data guarantee of (W-DRAS) (i.e., Theorems 1–2) hold valid under Assumption 2. We note that rectangular $\mathcal{U}$ does not imply that $\{u_i\}_{i\in[n]}$ are probabilistically independent. First, following Proposition 1, we rewrite (W-DRAS) as

$$\begin{aligned} \widehat{Z}(N, \epsilon) = \quad & \inf_{\rho, \boldsymbol{s}} \quad \epsilon^p \rho + \frac{1}{N}\sum_{j=1}^{N} \sup_{\boldsymbol{u}\in\mathcal{U}} \left\{ f(\boldsymbol{s}, \boldsymbol{u}) - \rho\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p \right\} \\ & \text{s.t.} \quad \rho \geq 0,\ \boldsymbol{s} \in \mathcal{S}. \end{aligned} \tag{2.12}$$

Recall that formulation (2.12) is potentially prohibitive to compute because it entails solving N non-convex optimization problems, in which $f(\boldsymbol{s}, \boldsymbol{u}) - \rho\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p$ is neither convex nor concave in $\boldsymbol{u}$. Fortunately, Assumption 2 enables us to recast these problems as linear programs for fixed ρ and $\boldsymbol{s}$, as summarized in the following proposition.

Proposition 2. *Suppose that Assumption 2 holds. Given $p \geq 1$, $\rho \geq 0$, and $\boldsymbol{s} \in \mathcal{S}$, we denote*

$$\omega_j(\rho, \boldsymbol{s}) := \sup_{\boldsymbol{u}\in\mathcal{U}} \Big\{ f(\boldsymbol{s}, \boldsymbol{u}) - \rho\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p \Big\}.$$

Then, we have

$$\omega_j(\rho, \boldsymbol{s}) \;=\; \max_t \; \sum_{k=1}^{n}\sum_{\ell=k}^{n+1}\left(\sum_{i=k}^{\min\{\ell,n\}} z_{i\ell j}\right) t_{k\ell} \tag{2.13a}$$

$$\text{s.t.} \;\; \sum_{k=1}^{i}\sum_{\ell=i}^{n+1} t_{k\ell} = 1 \quad \forall\, i \in [n] \tag{2.13b}$$

$$t_{k\ell} \geq 0 \quad \forall\, k \in [n],\; \forall\, \ell \in [k, n+1]_{\mathbb{Z}}. \tag{2.13c}$$

In formulation (2.13a)–(2.13c)*, $z_{i\ell j}$ is defined as $z_{i\ell j} = -s_i\pi_{i\ell} + \sup_{u_i^L \leq u_i \leq u_i^U}\{\pi_{i\ell}u_i - \rho|u_i - \widehat{u}_i^j|^p\}$ for all $j \in [N]$, $i \in [n]$, and $\ell \in [i, n+1]_{\mathbb{Z}}$, where*

$$\pi_{i\ell} := \begin{cases} -d_\ell + \sum_{q=i+1}^{\ell} c_q & \text{if } i \in [n] \text{ and } \ell \in [i, n]_{\mathbb{Z}} \\ C + \sum_{q=i+1}^{n} c_q & \text{if } i \in [n] \text{ and } \ell = n+1. \end{cases} \tag{2.13d}$$

Proposition 2 contrasts with the general computational tractability of two-stage DRO with right-hand side uncertainty (note that random variables $\boldsymbol{u}$ appear in the right-hand side of the linear program (2.2) that defines (W-DRAS)). In general, evaluating the objective function value of (W-DRAS) with a *fixed* appointment schedule $\boldsymbol{s}$ entails solving an exponential-size convex program (see Remark 5.5 in [11]). In contrast, Proposition 2 indicates that (W-DRAS) can be solved in polynomial time to find an *optimal* $\boldsymbol{s}$. Indeed, (W-DRAS) is a convex program because the function $\omega_j(\rho, \boldsymbol{s})$ is jointly convex in ρ and $\boldsymbol{s}$. Additionally, Proposition 2 implies that the epigraph of $\omega_j(\rho, \boldsymbol{s})$, denoted as $E := \{(\theta, \rho, \boldsymbol{s}) : \theta \geq \omega_j(\rho, \boldsymbol{s})\}$, can be separated in polynomial time. That is, given a point $(\bar{\theta}, \bar{\rho}, \bar{\boldsymbol{s}}) \in \mathbb{R}^{n+2}$, one can either verify that $(\bar{\theta}, \bar{\rho}, \bar{\boldsymbol{s}}) \in E$ or generate a hyperplane that separates $(\bar{\theta}, \bar{\rho}, \bar{\boldsymbol{s}})$ from E in polynomial time. Then, following the equivalence between separation and convex optimization established in the seminal work [102], we conclude that (W-DRAS) can be solved in polynomial time. This contrast in computational tractability arises because we study a particular DRO model in appointment scheduling and we take advantage of the structure of (W-DRAS).

We proceed to consider special p values. In particular, the following theorem recasts (W-DRAS) as a linear program and a second-order cone program for $p = 1$ and $p = 2$, respectively. In both cases, the solution of (W-DRAS) does not reply on specialized algorithms (such as separation) and can be obtained via off-the-shelf solvers (such as Gurobi and CPLEX).

Theorem 4. *Suppose that Assumption 2 holds. Then, when $p = 1$,* (W-DRAS) *yields the* *same optimal value and the same set of optimal solutions as the following linear program:*

$$\begin{aligned}
\min_{\rho,\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{z}} \quad & \epsilon\rho + \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\gamma_{ij} \\
\text{s.t.} \quad & \sum_{k=i}^{\min\{\ell,n\}} \gamma_{kj} \geq \sum_{k=i}^{\min\{\ell,n\}} z_{k\ell j} && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& z_{i\ell j} + \pi_{i\ell}s_i + |u_i^{L} - \widehat{u}_i^{j}|\rho \geq \pi_{i\ell}u_i^{L} && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& z_{i\ell j} + \pi_{i\ell}s_i \geq \pi_{i\ell}\widehat{u}_i^{j} && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& z_{i\ell j} + \pi_{i\ell}s_i + |u_i^{U} - \widehat{u}_i^{j}|\rho \geq \pi_{i\ell}u_i^{U} && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& \rho \geq 0,\ \boldsymbol{s}\in\mathcal{S}.
\end{aligned} \tag{2.14}$$

In addition, when $p = 2$, (W-DRAS) *yields the same optimal value and the same set of optimal solutions as the following second-order cone program:*

$$\begin{aligned}
\min_{\rho,\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{r}} \quad & \epsilon^2\rho + \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\gamma_{ij} \\
\text{s.t.} \quad & \sum_{k=i}^{\min\{\ell,n\}} \gamma_{kj} \geq \sum_{k=i}^{\min\{\ell,n\}} z_{k\ell j} && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& z_{i\ell j} + \pi_{i\ell}s_i - (u_i^{U} - u_i^{L})\beta_{i\ell j}^{U} \\
& -(\widehat{u}_i^{j} - u_i^{L})\beta_{i\ell j}^{L} - r_{i\ell j} \geq \pi_{i\ell}\widehat{u}_i^{j} && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& \left\| \begin{bmatrix} \pi_{i\ell} + \beta_{i\ell j}^{L} - \beta_{i\ell j}^{U} \\ r_{i\ell j} - \rho \end{bmatrix} \right\|_2 \leq r_{i\ell j} + \rho && \forall\, i\in[n],\ \forall\,\ell\in[i,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& \rho \geq 0,\ \boldsymbol{s}\in\mathcal{S}.
\end{aligned} \tag{2.15}$$

We note that both reformulations (2.14) and (2.15) involve $\mathcal{O}(Nn^2)$ decision variables and $\mathcal{O}(Nn^2)$ constraints. More generally, we show that (W-DRAS) admits a second-order conic reformulation for all $p > 1$, as long as p is rational. But as p typically takes integer values (e.g., $p = 1, 2$) in real-world applications, we relegate this more general result to Theorem 15 in the Appendix.

In addition to tractable computation of an optimal appointment schedule, Theorem 4 suggests an approach to stress testing *any* appointment schedule $\bar{s} \in \mathcal{S}$. Specifically, the following theorem derives a worst-case probability distribution $\mathbb{Q}_{\boldsymbol{u}}^{\star}$ of the random service durations $\boldsymbol{u}$ that attains $\sup_{\mathbb{Q}_{\boldsymbol{u}}\in\mathcal{D}_1(\widehat{\mathbb{P}}_{\boldsymbol{u}}^N,\epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}},\boldsymbol{u})]$. This distribution can be applied, for example, to assess the quality of an appointment schedule generated by any decision-making processes.[2]

[2] Although the derivation of worst-case distributions in Theorem 5 is based on the formulation when $p = 1$, similar conclusions can be obtained for the case when $p > 1$ and p is rational.

Theorem 5. *For fixed $\bar{\boldsymbol{s}} \in \mathcal{S}$ and $\epsilon \geq 0$, $\sup_{\mathbb{Q}_{\boldsymbol{u}} \in \mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}}, \boldsymbol{u})]$ equals the optimal value of the following linear program:*

$$
\begin{aligned}
\max_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{r}} \quad & \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\pi_{i\ell}\left[(u_i^L - \widehat{u}_i^j)q_{i\ell j} + (u_i^U - \widehat{u}_i^j)r_{i\ell j} + \left(\sum_{k=1}^{i} p_{k\ell j}\right)(\widehat{u}_i^j - \bar{s}_i)\right] \\
\text{s.t.} \quad & \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\left[(\widehat{u}_i^j - u_i^L)q_{i\ell j} + (u_i^U - \widehat{u}_i^j)r_{i\ell j}\right] \leq \epsilon \\
& \sum_{\ell=i}^{n+1}\sum_{k=1}^{i} p_{k\ell j} = 1 \quad \forall\, i \in [n],\ \forall\, j \in [N] \\
& \sum_{k=1}^{i} p_{k\ell j} - q_{i\ell j} - r_{i\ell j} \geq 0 \quad \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N] \\
& p_{i\ell j} \geq 0,\ q_{i\ell j} \geq 0,\ r_{i\ell j} \geq 0 \quad \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N].
\end{aligned}
\tag{2.16}
$$

Let $\{p^\star_{k\ell j}, q^\star_{i\ell j}, r^\star_{i\ell j}\}$ be an optimal solution of (2.16) *and define*

$$
\mathcal{T} = \left\{ \boldsymbol{t} \in \{0,1\}^{(n+1)(n+2)/2} :\ \sum_{k=1}^{i}\sum_{\ell=i}^{n+1} t_{k\ell} = 1,\ \forall i \in [n+1] \right\}.
$$

Then, there exists a distribution $\mathbb{P}^j_{\boldsymbol{t}}$ on $\mathcal{T}$ such that $\mathbb{P}^j_{\boldsymbol{t}}\{t_{k\ell} = 1\} = p^\star_{k\ell j}$ for all $j \in [N]$, $k \in [n+1]$, and $\ell \in [k, n+1]_{\mathbb{Z}}$. Furthermore, define the probability distribution

$$
\mathbb{Q}^\star_{\boldsymbol{u}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{\boldsymbol{\tau} \in \mathcal{T}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\}\delta_{\boldsymbol{u}^j(\boldsymbol{\tau})},
$$

where, for all $i \in [n]$ and $j \in [N]$,

$$
u_i^j(\boldsymbol{\tau}) = \sum_{\ell=i}^{n+1}\left(\sum_{k=1}^{i}\tau_{k\ell}\right)u_{i\ell j} \ \text{ and } \ u_{i\ell j} = \widehat{u}_i^j + \frac{q^\star_{i\ell j}(u_i^L - \widehat{u}_i^j)}{\sum_{k=1}^{i} p^\star_{k\ell j}} + \frac{r^\star_{i\ell j}(u_i^U - \widehat{u}_i^j)}{\sum_{k=1}^{i} p^\star_{k\ell j}},
$$

where we adopt an extended arithmetic given by $0/0 = 0$. Then, $\mathbb{Q}^\star_{\boldsymbol{u}}$ belongs to the Wasserstein ambiguity set $\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)$ and $\mathbb{E}_{\mathbb{Q}^\star_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}}, \boldsymbol{u})] = \sup_{\mathbb{Q}_{\boldsymbol{u}} \in \mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}}, \boldsymbol{u})]$.

Remark 2. *Intuitively, the* (W-DRAS) *model can be viewed as a two-person game between the AS scheduler and the nature, who picks a $\mathbb{Q}^\star_{\boldsymbol{u}}$ that maximizes the expected total cost after the schedule $\boldsymbol{s}$ is determined. Theorem 5 gives us a clear picture on how the nature makes its pick. Specifically, $\mathbb{Q}^\star_{\boldsymbol{u}}$ is a mixture of N distributions, with each pertaining to a data sample. That is, $\mathbb{Q}^\star_{\boldsymbol{u}} = (1/N)\sum_{j=1}^{N}\mathbb{Q}^j_{\boldsymbol{u}}$, where $\mathbb{Q}^j_{\boldsymbol{u}} := \sum_{\boldsymbol{\tau} \in \mathcal{T}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\}\delta_{\boldsymbol{u}^j(\boldsymbol{\tau})}$ for all $j \in [N]$. Note that $\mathcal{T}$ is the collection of all possible partitions of the set $[n+1]$ into sub-intervals. Hence, what the nature does under $\mathbb{Q}^j_{\boldsymbol{u}}$ is to: (i) randomly pick a partition*

$\boldsymbol{\tau} \in \mathcal{T}$ following distribution $\mathbb{P}_{\boldsymbol{t}}^j$ and (ii) for each $i \in [n]$, if i belongs to the sub-interval $[k,\ell]_{\mathbb{Z}}$ (i.e., if $\tau_{k\ell} = 1$ with $k \le i \le \ell$) then the i^{th} appointment has a service duration $u_{i\ell j}$.

Remark 3. *In the situation where the radius of the Wasserstein ball is set to zero, the worst-case distribution $\mathbb{Q}_{\boldsymbol{u}}^\star$ reduces to the empirical distribution. Indeed, setting $\epsilon = 0$ enforces all $q_{i\ell j}$ and $r_{i\ell j}$ to be zero in (2.16). It follows that $u_{i\ell j} = \widehat{u}_i^j$ and so $u_i^j(\boldsymbol{\tau}) = \widehat{u}_i^j$ for all $\boldsymbol{\tau} \in \mathcal{T}$. Therefore, we have $\mathbb{Q}_{\boldsymbol{u}}^\star = \frac{1}{N}\sum_{j=1}^N \delta_{\widehat{\boldsymbol{u}}^j}$.*

2.3 Random No-Shows and Service Durations

In many AS systems, appointments have random no-shows, i.e., the appointee cancels her appointment too late such that the scheduler cannot make a substitute. In such a situation, the approach studied in Section 2.2 is no longer applicable. In Section 2.3.1, we extend (W-DRAS) and the Wasserstein ambiguity set to incorporate both random no-shows and service durations. We reformulate this extended model as copositive programs in Section 2.3.2 and as tractable convex programs under a mild condition in Section 2.3.3.

2.3.1 Extended model for random no-shows

We model the random no-show of appointment i by using a Bernoulli random variable λ_i such that $\lambda_i = 1$ if appointee i shows up and $\lambda_i = 0$ otherwise. We denote the support of $\boldsymbol{\lambda}$ by $\Lambda \subseteq \{0,1\}^n$. If $\Lambda = \{0,1\}^n$ then it includes all possible scenarios of no-shows. Unfortunately, it has been observed (e.g., in [44]) that such a Λ often results in poor out-of-sample performance. Intuitively, this is because the set $\{0,1\}^n$ includes many unlikely scenarios (e.g., a majority of appointees do not show up), rendering the resulting schedule over-conservative. In this chapter, we propose to consider a less conservative, budget-constrained support set

$$\Lambda := \left\{ \boldsymbol{\lambda} \in \{0,1\}^n : \sum_{i=1}^n (1-\lambda_i) \le K \right\},$$

where the integer parameter $K \in [n]$ denotes the "budget" of no-shows and controls the conservativeness of Λ. Intuitively, Λ contains only scenarios with no more than K no-shows out of the n appointments. For example, if $K = 0$ then all n appointees show up for their appointments, yielding the least conservative support set; and if $K = n$ then $\Lambda = \{0,1\}^n$, yielding the most conservative case. The parameter K can be determined based on the scheduler's knowledge, risk attitude, and/or the historical data of no-shows (see (2.18) below).

Observing that an AS system spends no time on a no-show appointment, we let $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n)^\top$ represent the *actual* service durations of appointments and, for ease of exposition, let $\boldsymbol{\xi} := (\boldsymbol{\mu}^\top, \boldsymbol{\lambda}^\top)^\top$. Then, under a rectangularity condition, we define the support set of $\boldsymbol{\xi}$ as

$$\Xi := \Big\{(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathbb{R}^n \times \Lambda : \ u_i^{\mathrm{L}}\lambda_i \leq \mu_i \leq u_i^{\mathrm{U}}\lambda_i \ \ \forall\, i \in [n]\Big\}.$$

We note that the constraint $u_i^{\mathrm{L}}\lambda_i \leq \mu_i \leq u_i^{\mathrm{U}}\lambda_i$ ensures that (i) if $\lambda_i = 1$ (i.e., if appointee i shows up) then $\mu_i \in [u_i^{\mathrm{L}}, u_i^{\mathrm{U}}]$ and (ii) if $\lambda_i = 0$ (i.e., if appointee i does not show up) then $\mu_i = 0$ (i.e., the actual service duration is zero). Under the standard assumption that $d_{i+1} - d_i \leq c_{i+1}$ for all $i \in [n-1]$ (see Section 2.2.1 for elaboration of this assumption), the total cost of the appointment system for given $\boldsymbol{s}$ and $\boldsymbol{\xi}$ can be obtained from the following linear program:

$$\begin{aligned} g(\boldsymbol{s}, \boldsymbol{\xi}) := \min_{\boldsymbol{w}, \boldsymbol{v}} \quad & \sum_{i=1}^{n} (c_i \lambda_i w_i + d_i v_i) + C w_{n+1} \\ \text{s.t.} \quad & w_i - v_{i-1} = \mu_{i-1} + w_{i-1} - s_{i-1} \quad \forall\, i \in [2, n+1]_{\mathbb{Z}} \\ & \boldsymbol{w} \geq 0, \ w_1 = 0, \ \boldsymbol{v} \geq 0. \end{aligned} \tag{2.17}$$

Here, the cost of waiting time $c_i \lambda_i w_i$ is modeled from the perspective of appointments, i.e., this cost is waived if appointee i does not show up. In addition, we note that the variables $\boldsymbol{u}$ do not explicitly appear in the above definitions of Ξ and $g(\boldsymbol{s}, \boldsymbol{\xi})$. In this section, we interpret u_i as the service duration *if appointee i shows up*, i.e., $\boldsymbol{u}$ are conditional random variables depending on $\boldsymbol{\lambda}$. As a result, $\boldsymbol{u}$ may not even be observable when no-shows take place, while $\boldsymbol{\xi}$ are always observable, for example, from the historical data of service durations and no-shows. Similar to Section 2.2, we assume that we observe a finite set of N i.i.d. samples of $\boldsymbol{\xi}$, denoted as $\{\widehat{\boldsymbol{\xi}}^1, \ldots, \widehat{\boldsymbol{\xi}}^N\}$. Then, we consider the following p-Wasserstein ambiguity set

$$\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}, \epsilon) := \Big\{\mathbb{Q}_{\boldsymbol{\xi}} \in \mathcal{P}(\Xi) : \ d_p(\mathbb{Q}_{\boldsymbol{\xi}}, \widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}) \leq \epsilon\Big\},$$

where $\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}$ represents the empirical distribution of $\boldsymbol{\xi}$ based on the N i.i.d. samples, i.e., $\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}} = \frac{1}{N}\sum_{j=1}^{N} \delta_{\widehat{\boldsymbol{\xi}}^j}$. Additionally, we can determine K, the budget of no-shows in Λ, based on these samples. For example, we can set

$$K := \max_{j \in [N]} \left\{ \sum_{i=1}^{n} (1 - \widehat{\lambda}_i^j) \right\}. \tag{2.18}$$

With the above Wasserstein ambiguity set, we formulate the following DRO model to seek a schedule that minimizes the expected total cost with regard to the worst-case distribution in $\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\, \epsilon)$:

$$\widehat{Z}_{\text{NS}}(N,\, \epsilon) := \min_{\boldsymbol{s}\in\mathcal{S}} \sup_{\mathbb{Q}_{\boldsymbol{\xi}}\in\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\boldsymbol{s},\, \boldsymbol{\xi})]. \tag{W-NS}$$

We close this section by noting that similar asymptotic consistency and finite-data guarantee as in Section 2.2 (see Theorems 1–2) also hold for (W-NS). In particular, as the data size N increases to infinity, $\widehat{Z}_{\text{NS}}(N,\, \epsilon)$ converges to the optimal value of the stochastic AS model, where perfect information of the probability distribution of $\boldsymbol{\xi}$ is known. Accordingly, a (W-NS) optimal appointment schedule converges to the optimal schedule obtained from this stochastic model. In addition, with high confidence, (W-NS) provides an upper bound on the optimal value of the stochastic model with any finite data size N. We skip the formal statement of these two results to avoid repetition.

2.3.2 Copositive programming reformulations

In this section, we propose a copositive reformulation for (W-NS). As the first step, we represent $g(\boldsymbol{s}, \boldsymbol{\xi})$ by the following dual linear program of (2.17):

$$g(\boldsymbol{s}, \boldsymbol{\xi}) \;=\; \max_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \;\sum_{i=1}^{n}(\mu_i - s_i)y_i \tag{2.19}$$

where dual variables $\boldsymbol{y}$ are associated with the first constraint in (2.17) and the polyhedral feasible set $\mathcal{Y}(\boldsymbol{\lambda})$ is described as

$$\mathcal{Y}(\boldsymbol{\lambda}) := \left\{ \boldsymbol{y}\in\mathbb{R}^n : \begin{array}{l} -y_i \le d_i \quad \forall\, i\in[n] \\ y_{i-1} - y_i \le c_i\lambda_i \quad \forall\, i\in[2,n]_{\mathbb{Z}} \\ y_n \le C \end{array} \right\}.$$

The strong duality between (2.17) and (2.19) holds because $\mathcal{Y}(\boldsymbol{\lambda})$ is nonempty and compact for any $\boldsymbol{\lambda}\in\Lambda$. We formally state this fact in the following lemma and omit its proof due to its similarity to that of Lemma 2.

Lemma 3. *For any $\boldsymbol{\lambda}\in\Lambda$, $\mathcal{Y}(\boldsymbol{\lambda})$ is nonempty, compact, and convex.*

Second, we derive a deterministic formulation to compute the worst-case expectation in (W-NS), $\sup_{\mathbb{Q}_{\boldsymbol{\xi}}\in\mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\boldsymbol{s},\, \boldsymbol{\xi})]$, for fixed $\boldsymbol{s}\in\mathcal{S}$. We state this formulation in the following proposition and omit the proof due to the similarity to that of Proposition 1.

Proposition 3. *For fixed $\boldsymbol{s} \in \mathcal{S}$, $\sup\limits_{\mathbb{Q}_{\boldsymbol{\xi}} \in \mathcal{D}_p(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\boldsymbol{s}, \boldsymbol{\xi})]$ equals the optimal value of the following formulation:*

$$
\begin{aligned}
&\inf \;\; \epsilon^p \rho + \frac{1}{N} \sum_{j=1}^{N} \sup_{\boldsymbol{\xi} \in \Xi} \Big\{ g(\boldsymbol{s}, \boldsymbol{\xi}) - \rho \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}^j\|_p^p \Big\} \qquad (2.20)\\
&\textit{s.t.} \;\; \rho \geq 0.
\end{aligned}
$$

The deterministic formulation in Proposition 3 is computationally intractable particularly due to the maximization problem in the objective function (2.20). As compared to its counterpart in Section 2.2 (see (2.7)), the problem in (2.20) involves both binary decision variables $\boldsymbol{\lambda}$ and continuous decision variables $\boldsymbol{\mu}$, which significantly increases the computational difficulty. Nevertheless, in the following theorem we derive a copositive reformulation of (W-NS) for $p \in \{1, 2\}$. Similarly, we define the sets $\mathcal{F}^j_{\text{NS},1}$ for all $j \in [N]$ and $\mathcal{F}_{\text{NS},2}$:

$$
\mathcal{F}^j_{\text{NS},1} := \left\{ (\boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-, \boldsymbol{y}) \in \mathbb{R}^{5n} : \begin{array}{l}
\boldsymbol{\mu}^+ \in \mathbb{R}^n_+,\ \boldsymbol{\mu}^- \in \mathbb{R}^n_+,\ \boldsymbol{\lambda}^+ \in \mathbb{R}^n_+,\ \boldsymbol{\lambda}^- \in \mathbb{R}^n_+,\ \boldsymbol{y} \in \mathbb{R}^n \\
0 \leq \boldsymbol{\lambda}^+ - \boldsymbol{\lambda}^- + \widehat{\boldsymbol{\lambda}}^j \leq \boldsymbol{e},\ \boldsymbol{\lambda}^+ \leq \boldsymbol{e},\ \boldsymbol{\lambda}^- \leq \boldsymbol{e} \\
\sum\limits_{i=1}^{n} \Big[1 - (\lambda^+_i - \lambda^-_i + \widehat{\lambda}^j_i) \Big] \leq K \\
u^{\text{L}}_i (\lambda^+_i - \lambda^-_i + \widehat{\lambda}^j_i) \leq \mu^+_i - \mu^-_i + \widehat{\mu}^j_i \ \ \forall\, i \in [n] \\
\mu^+_i - \mu^-_i + \widehat{\mu}^j_i \leq u^{\text{U}}_i (\lambda^+_i - \lambda^-_i + \widehat{\lambda}^j_i) \ \ \forall\, i \in [n] \\
y_n \leq C,\ -y_i \leq d_i \ \ \forall\, i \in [n] \\
y_{i-1} - y_i \leq c_i (\lambda^+_i - \lambda^-_i + \widehat{\lambda}^j_i) \ \ \forall\, i \in [2, n]_{\mathbb{Z}}
\end{array} \right\},
$$

$$
\mathcal{F}_{\text{NS},2} := \left\{ (\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{y}) \in \mathbb{R}^{3n} : \begin{array}{l}
\boldsymbol{\mu} \in \mathbb{R}^n,\ \boldsymbol{\lambda} \in \mathbb{R}^n,\ \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{\lambda}) \\
\sum\limits_{i=1}^{n} (1 - \lambda_i) \leq K,\ 0 \leq \boldsymbol{\lambda} \leq \boldsymbol{e} \\
u^{\text{L}}_i \lambda_i \leq \mu_i \leq u^{\text{U}}_i \lambda_i \ \ \forall i \in [n]
\end{array} \right\},
$$

and then construct their perspective sets $\mathcal{K}^j_{\text{NS},1}$ and $\mathcal{K}_{\text{NS},2}$ respectively:

$$
\begin{aligned}
\mathcal{K}^j_{\text{NS},1} &:= \text{closure}\Big(\Big\{ (t, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-, \boldsymbol{y}) : (\boldsymbol{\mu}^+/t, \boldsymbol{\mu}^-/t, \boldsymbol{\lambda}^+/t, \boldsymbol{\lambda}^-/t, \boldsymbol{y}/t) \in \mathcal{F}^j_{\text{NS},1},\ t > 0 \Big\} \Big), \\
\mathcal{K}_{\text{NS},2} &:= \text{closure}\Big(\Big\{ (t, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{y}) : (\boldsymbol{\mu}/t, \boldsymbol{\lambda}/t, \boldsymbol{y}/t) \in \mathcal{F}_{\text{NS},2},\ t > 0 \Big\} \Big).
\end{aligned}
$$

Now, we are ready to present copositive programming reformulations of (W-NS) for the cases of $p = 1, 2$. For ease of exposition, we define

$$
\boldsymbol{G}_j^1(\rho,\boldsymbol{s}) := \begin{bmatrix} 0 & -\frac{1}{2}\rho\boldsymbol{e}^\top & -\frac{1}{2}\rho\boldsymbol{e}^\top & -\frac{1}{2}\rho\boldsymbol{e}^\top & -\frac{1}{2}\rho\boldsymbol{e}^\top & \frac{1}{2}(\widehat{\boldsymbol{\mu}}^j-\boldsymbol{s})^\top \\ -\frac{1}{2}\rho\boldsymbol{e} & 0 & 0 & 0 & 0 & \frac{1}{2}\boldsymbol{I} \\ -\frac{1}{2}\rho\boldsymbol{e} & 0 & 0 & 0 & 0 & -\frac{1}{2}\boldsymbol{I} \\ -\frac{1}{2}\rho\boldsymbol{e} & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{2}\rho\boldsymbol{e} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2}(\widehat{\boldsymbol{\mu}}^j-\boldsymbol{s}) & \frac{1}{2}\boldsymbol{I} & -\frac{1}{2}\boldsymbol{I} & 0 & 0 & 0 \end{bmatrix},
$$

$$
\boldsymbol{J}_i := \begin{bmatrix} 0\\0\\0\\\boldsymbol{e}_i\\0\\0 \end{bmatrix}\begin{bmatrix} 0\\0\\0\\\boldsymbol{e}_i\\0\\0 \end{bmatrix}^\top - \frac{1}{2}\begin{bmatrix} 0\\0\\0\\\boldsymbol{e}_i\\0\\0 \end{bmatrix}\begin{bmatrix} 1\\0\\0\\0\\0\\0 \end{bmatrix}^\top - \frac{1}{2}\begin{bmatrix} 1\\0\\0\\0\\0\\0 \end{bmatrix}\begin{bmatrix} 0\\0\\0\\\boldsymbol{e}_i\\0\\0 \end{bmatrix}^\top,
$$

$$
\boldsymbol{M}_i := \begin{bmatrix} 0\\0\\0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}\begin{bmatrix} 0\\0\\0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}^\top - \frac{1}{2}\begin{bmatrix} 0\\0\\0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}\begin{bmatrix} 1\\0\\0\\0\\0\\0 \end{bmatrix}^\top - \frac{1}{2}\begin{bmatrix} 1\\0\\0\\0\\0\\0 \end{bmatrix}\begin{bmatrix} 0\\0\\0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}^\top,
$$

$$
\boldsymbol{G}_j^2(\rho,\boldsymbol{s}) := \begin{bmatrix} -\rho(\|\widehat{\boldsymbol{\mu}}_j\|^2+\|\widehat{\boldsymbol{\lambda}}_j\|^2) & \rho\widehat{\boldsymbol{\mu}}_j^\top & \rho\widehat{\boldsymbol{\lambda}}_j^\top & -\frac{1}{2}\boldsymbol{s}^\top \\ \rho\widehat{\boldsymbol{\mu}}_j & -\rho\boldsymbol{I} & 0 & \frac{1}{2}\boldsymbol{I} \\ \rho\widehat{\boldsymbol{\lambda}}_j & 0 & -\rho\boldsymbol{I} & 0 \\ -\frac{1}{2}\boldsymbol{s} & \frac{1}{2}\boldsymbol{I} & 0 & 0 \end{bmatrix},
$$

$$
\boldsymbol{N}_i := \begin{bmatrix} 0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}\begin{bmatrix} 0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}^\top - \frac{1}{2}\begin{bmatrix} 0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}\begin{bmatrix} 1\\0\\0\\0 \end{bmatrix}^\top - \frac{1}{2}\begin{bmatrix} 1\\0\\0\\0 \end{bmatrix}\begin{bmatrix} 0\\0\\\boldsymbol{e}_i\\0 \end{bmatrix}^\top.
$$

Theorem 6. *When $p = 1$,* (W-NS) *yields the same optimal value and the same optimal* *solutions for the following copositive program:*

$$
\begin{aligned}
\widehat{Z}_{\text{NS}}(N,\,\epsilon) = \quad \inf \quad & \epsilon\rho + \frac{1}{N}\sum_{j=1}^{N}\beta_j \\
\text{s.t.} \quad & \rho\in\mathbb{R},\ \boldsymbol{\psi}\in\mathbb{R}^{n\times N}, \boldsymbol{\phi}\in\mathbb{R}^{n\times N} \\
& \beta_j\boldsymbol{e}_1\boldsymbol{e}_1^\top + \sum_{i=1}^{n}\psi_{ij}\boldsymbol{M}_i + \sum_{i=1}^{n}\phi_{ij}\boldsymbol{J}_i - \boldsymbol{G}_j^1(\rho,\,\boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}^j_{\text{NS},1}) \quad \forall\, j\in[N] \\
& \rho\geq 0,\ \boldsymbol{s}\in\mathcal{S}.
\end{aligned}
\tag{2.21}
$$

In addition, when $p = 2$*,* (W-NS) *yields the same optimal value and the same optimal solutions for the following copositive program:*

$$
\begin{aligned}
\widehat{Z}_{\text{NS}}(N,\,\epsilon) = \quad \inf \quad & \epsilon^2\rho + \frac{1}{N}\sum_{j=1}^{N}\beta_j \\
\text{s.t.} \quad & \rho\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^{N},\ \boldsymbol{\psi}\in\mathbb{R}^{n\times N} \\
& \beta_j\boldsymbol{e}_1\boldsymbol{e}_1^\top + \sum_{i=1}^{n}\psi_{ij}\boldsymbol{N}_i - \boldsymbol{G}_j^2(\rho,\,\boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}_{\text{NS},2}) \quad \forall\, j\in[N] \\
& \rho\geq 0,\ \boldsymbol{s}\in\mathcal{S}.
\end{aligned}
\tag{2.22}
$$

2.3.3 Tractable reformulations

In this subsection, we consider a setting that admits a tractable reformulation of (W-NS) for general $p \geq 1$. Different from Section 2.2 that considers random service durations only, (W-NS) incorporates no-shows and the accompanying Bernoulli random variables. To address the amplified computational challenge, we adopt a different approach based on dynamic programming and network flow techniques. This leads to a (polynomial-size) linear programming reformulation of (W-NS) when $p = 1$ and a (polynomial-size) second-order cone programming reformulation when $p > 1$ and p is rational. To this end, we make the following assumption on the costs of waiting and idleness of the AS system.

Assumption3 (Homogeneous Costs)**.** *The costs of appointment waiting and server idleness are homogeneous, i.e.,* $c_0 := c_1 = c_2 = \cdots c_n$ *and* $d_0 := d_1 = d_2 = \cdots = d_n$*.*

Assumption 3 is non-stringent because (1) the server idleness is always associated with the same server, and (2) although the waiting times are associated with different appointments, the scheduler should consider a homogeneous cost to ensure fairness among all appointments. Under this assumption, we can assume $c_0 = 1$ without loss of generality.

We first recast and identify an optimality condition (OC) of the maximization problem $\sup_{\boldsymbol{\xi}\in\Xi}\{g(\boldsymbol{s},\,\boldsymbol{\xi}) - \rho\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}^j\|_p^p\}$ in formulation (2.20).

Proposition 4. *Denote* $\omega'_j(\rho, \boldsymbol{s}) = \sup_{\boldsymbol{\xi}\in\Xi} \{g(\boldsymbol{s}, \boldsymbol{\xi}) - \rho\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}^j\|_p^p\}$ *and suppose that Assumptions 2–3 hold. Then, for all* $p \geq 1$, $j \in [N]$, $\rho \geq 0$, *and* $\boldsymbol{s} \in \mathcal{S}$,

$$\omega'_j(\rho, \boldsymbol{s}) \;=\; \sup_{\boldsymbol{\lambda}\in\Lambda,\ \boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \left\{\sum_{i=1}^{n} f_{ij}(\lambda_i, y_i)\right\}, \tag{2.23}$$

where

$$f_{ij}(\lambda_i, y_i) := \sup_{u_i^L\lambda_i \leq \mu_i \leq u_i^U\lambda_i} \left\{y_i(\mu_i - s_i) - \rho|\mu_i - \widehat{\mu}_i^j|^p - \rho|\lambda_i - \widehat{\lambda}_i^j|^p\right\} \tag{2.24}$$

for all $i \in [n]$. *In addition, without loss of optimality, variables* $\boldsymbol{\lambda}$ *and* $\boldsymbol{y}$ *satisfy the following optimality condition:*

$$\begin{cases} y_n = C \text{ or } y_n = -d_0 & \\ y_i = -d_0 \text{ or } y_i = y_{i+1} + \lambda_{i+1} & \forall i \in [n-1] \\ y_i \in \mathcal{Y}_i & \forall i \in [n], \end{cases} \tag{OC}$$

where $\mathcal{Y}_i := [-d_0, -d_0 + n - i]_{\mathbb{Z}} \cup [C, C + n - i]_{\mathbb{Z}}$.

(OC) shrinks the search space of problem (2.23) from a union of polytope to a finite set of points. Specifically, for each $i \in [n]$, variable y_i has $2(n-i+1)$ possible choices because $\mathcal{Y}_i$ consists of $2(n-i+1)$ elements. More importantly, given the value of y_{i+1}, y_i can take only two values: $-d_0$ or $y_{i+1}+\lambda_{i+1}$. This allows us to recast $\omega'_j(\rho, \boldsymbol{s})$ as a dynamic program (DP), that is, we solve (2.23) by sequentially determining (λ_1, y_1), (λ_2, y_2), and so on. To this end, for each $i \in [n]$, we define the state of this DP as $(\bar{\lambda}_i, y_i) \in [0, K]_{\mathbb{Z}} \times \mathcal{Y}_i$, where $\bar{\lambda}_i := \sum_{k=1}^{i}(1 - \lambda_k)$ records the total number of no-shows among the first i appointments[3], and define the value function of this DP through

$$\begin{aligned} V_{nj}(\bar{\lambda}_n, y_n) \;&=\; 0 \;\; \forall\, (\bar{\lambda}_n, y_n) \in [0, K]_{\mathbb{Z}} \times \mathcal{Y}_n, \\ V_{(i-1)j}(\bar{\lambda}_{i-1}, y_{i-1}) \;&=\; \sup_{\bar{\lambda}_i, y_i} \;\; f_{ij}(\bar{\lambda}_{i-1} - \bar{\lambda}_i + 1, y_i) + V_{ij}(\bar{\lambda}_i, y_i) \\ &\qquad\;\; \text{s.t. } \; \bar{\lambda}_i \in \{\bar{\lambda}_{i-1}, \bar{\lambda}_{i-1} + 1\}, \;\; y_i \in \{-d_0,\ y_{i-1} - (\bar{\lambda}_{i-1} - \bar{\lambda}_i + 1)\} \\ \forall\, i \in [2, n]_{\mathbb{Z}}, \;\; \forall\, &(\bar{\lambda}_{i-1}, y_{i-1}) \in [0, K]_{\mathbb{Z}} \times \mathcal{Y}_{i-1}. \end{aligned}$$

It follows that $\omega'_j(\rho, \boldsymbol{s}) = \sup_{(\bar{\lambda}_1, y_1)\in\{0,1\}\times\mathcal{Y}_1}\{f_{1j}(1 - \bar{\lambda}_1, y_1) + V_{1j}(\bar{\lambda}_1, y_1)\}$.

To provide more intuition, we map the states and state transitions of this DP onto

[3]Note that $\lambda_1 = 1 - \bar{\lambda}_1$ and $\lambda_i = \bar{\lambda}_{i-1} - \bar{\lambda}_i + 1$ for all $i \in [2, n]_{\mathbb{Z}}$. In addition, although $\bar{\lambda}_i \in [0, \min\{i, K\}]_{\mathbb{Z}}$, we consider $\bar{\lambda}_i \in [0, K]_{\mathbb{Z}}$ in this DP for notational brevity.

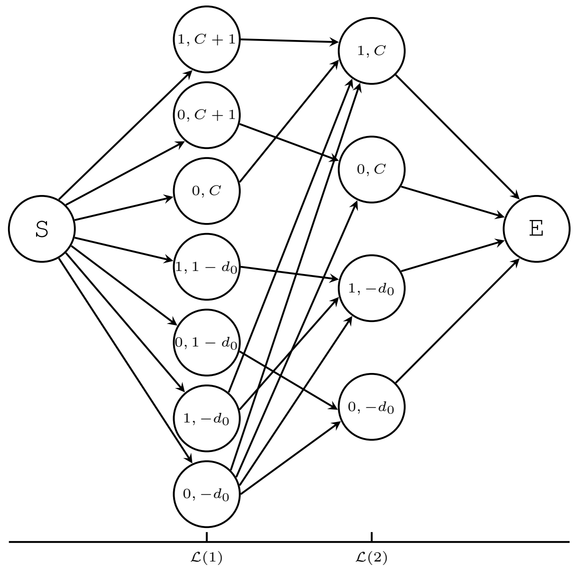


Figure 2.1: An Example of the Network $(\mathcal{G}, \mathcal{E})$ with $n = 2$ and $K = 1$. The nodes in layers 1 and 2 are arranged along the coordinates $\mathcal{L}(1)$ and $\mathcal{L}(2)$, respectively.

an n-layer network. In layer i, $\forall\, i \in [n]$, we construct a set $\mathcal{L}(i)$ of nodes with $\mathcal{L}(i) := \big\{(\bar{\lambda}_i, y_i) : \bar{\lambda}_i \in [0, K]_{\mathbb{Z}},\ y_i \in \mathcal{Y}_i\big\}$. In addition, we construct directed arcs between two consecutive layers in the network. For $i \in [2, n]_{\mathbb{Z}}$, there is an arc from node $(\bar{\lambda}_{i-1}, y_{i-1}) \in \mathcal{L}(i-1)$ to node $(\bar{\lambda}_i, y_i) \in \mathcal{L}(i)$ if $\bar{\lambda}_i \in \{\bar{\lambda}_{i-1}, \bar{\lambda}_{i-1} + 1\}$ and $y_i \in \{-d_0,\ y_{i-1} - (\bar{\lambda}_{i-1} - \bar{\lambda}_i + 1)\}$. Finally, we construct a (dummy) starting node S and a (dummy) ending node E, as well as arcs from S to all nodes $(\bar{\lambda}_1, y_1) \in \mathcal{L}(1)$ with $\bar{\lambda}_1 \in \{0, 1\}$ and from all nodes in $\mathcal{L}(n)$ to E. We depict an example of this network in Figure 2.1. It can be observed that each feasible solution of the DP corresponds to an S-E path in the network, and vice versa. This indicates that, by associating appropriate length to each arc, solving the longest-path problem in the network produces an optimal solution to the DP. We summarize this observation in the following proposition.

Proposition 5. *Suppose that Assumptions 2–3 hold and consider a directed network* $(\mathcal{G}, \mathcal{E})$ *such that* $\mathcal{G} = \bigcup_{i=1}^{n} \mathcal{L}(i) \cup \{S, E\}$ *and* $\mathcal{E} = \{((\bar{\lambda}_{i-1}, y_{i-1}),\ (\bar{\lambda}_i, y_i)) \in \mathcal{L}(i-1) \times \mathcal{L}(i) : \bar{\lambda}_i \in \{\bar{\lambda}_{i-1}, \bar{\lambda}_{i-1} + 1\},\ y_i \in \{-d_0,\ y_{i-1} - (\bar{\lambda}_{i-1} - \bar{\lambda}_i + 1)\},\ i \in [2, n]_{\mathbb{Z}}\} \cup \{(S,\ (\bar{\lambda}_1, y_1)) : \bar{\lambda}_1 \in \{0, 1\},\ (\bar{\lambda}_1, y_1) \in \mathcal{L}(1)\} \cup \{((\bar{\lambda}_n, y_n),\ E) :\ (\bar{\lambda}_n, y_n) \in \mathcal{L}(n)\}$*. For each arc* $(k, \ell) \in \mathcal{E}$*, let* $g_{k\ell j}$ *represent its length such that*

$$g_{k\ell j} = \begin{cases} f_{1j}(1 - \bar{\lambda}_1,\ y_1) & \text{if } k = S \text{ and } \ell = (\bar{\lambda}_1, y_1) \\ f_{ij}(\bar{\lambda}_{i-1} - \bar{\lambda}_i + 1,\ y_i) & \text{if } k = (\bar{\lambda}_{i-1}, y_{i-1}) \text{ and } \ell = (\bar{\lambda}_i, y_i) \text{ for } i \in [2, n]_{\mathbb{Z}} \\ 0 & \text{if } \ell = E, \end{cases}$$

where $f_{ij}(\cdot,\cdot)$ *is defined in* (2.24). *Then, we have*

$$\omega_j'(\rho, \boldsymbol{s}) \;=\; \max_{\boldsymbol{z}} \sum_{(k,\ell)\in\mathcal{E}} g_{k\ell j} z_{k\ell} \tag{2.25a}$$

$$\text{s.t.} \quad \sum_{\ell:(k,\ell)\in\mathcal{E}} z_{k\ell} - \sum_{\ell:(\ell,k)\in\mathcal{E}} z_{\ell k} = \begin{cases} 1 & \text{if } k = S \\ 0 & \text{if } k \neq S, E \\ -1 & \text{if } k = E \end{cases} \quad \forall\, k \in \mathcal{G} \tag{2.25b}$$

$$z_{k\ell} \geq 0 \quad \forall\,(k,\ell) \in \mathcal{E}. \tag{2.25c}$$

We note that $|\mathcal{G}| = \mathcal{O}(Kn^2)$ and $|\mathcal{E}| = \mathcal{O}(Kn^2)$. This indicates that $\omega_j'(\rho, \boldsymbol{s})$ can be computed in polynomial time by solving the linear program (2.25a)–(2.25c). Following the equivalence between separation and convex optimization (see [102]), Proposition 5 implies that (W-NS) can be solved in polynomial time to find an optimal appointment schedule under both random no-shows and service durations.

As a generalization of Theorem 4, the following theorem recasts (W-NS) as a linear program and a second-order cone program for $p = 1$ and $p = 2$, respectively. Similar second-order conic reformulations can be obtained for any rational $p \geq 1$.

Theorem 7. *Suppose that Assumptions 2–3 hold and define*

$$\begin{aligned}
&\mathcal{E}_1^0 := \{(S,\ell) \in \mathcal{E} :\ \ell = (1, y_i)\}\,,\ \ \mathcal{E}_1^1 := \{(S,\ell) \in \mathcal{E} :\ \ell = (0, y_i)\}\,,\\
&\mathcal{E}_i^0 := \{(k,\ell) \in \mathcal{E} :\ k = (\bar\lambda_{i-1}, y_{i-1}), \ell = (\bar\lambda_i, y_i),\ \textit{and}\ \bar\lambda_i = \bar\lambda_{i-1} + 1\}\ \ \forall\, i \in [2,n]_{\mathbb{Z}},\\
&\mathcal{E}_i^1 := \{(k,\ell) \in \mathcal{E} :\ k = (\bar\lambda_{i-1}, y_{i-1}), \ell = (\bar\lambda_i, y_i),\ \textit{and}\ \bar\lambda_i = \bar\lambda_{i-1}\}\ \ \forall\, i \in [2,n]_{\mathbb{Z}},\\
&\textit{and}\ \ \mathcal{E}_E := \{(k,\ell) \in \mathcal{E} :\ \ell = E\}.
\end{aligned}$$

Then, when $p = 1$, (W-NS) *yields the same optimal value and the same set of optimal* *appointment schedules as the following linear program:*

$$\begin{aligned}
\min_{\rho,\boldsymbol{s},\boldsymbol{\alpha}} \ & \epsilon\rho + \frac{1}{N}\sum_{j=1}^{N}\left(\alpha_{S}^{j} - \alpha_{E}^{j}\right) \\
\text{s.t.} \ & \alpha_k^j - \alpha_\ell^j \geq 0 \ \ \forall (k,\ell) \in \mathcal{E}_E \ \ \forall j \in [N] \\
& \alpha_k^j - \alpha_\ell^j + y_i s_i + (\widehat{\mu}_i^j + \widehat{\lambda}_i^j)\rho \geq 0 \ \ \forall i \in [n], \ \forall (k,\ell) \in \mathcal{E}_i^0, \ \forall j \in [N] \\
& \left.\begin{aligned}
& \alpha_k^j - \alpha_\ell^j + y_i s_i + (1 - \widehat{\lambda}_i^j + \widehat{\mu}_i^j - u_i^L)\rho \geq u_i^L y_i \\
& \alpha_k^j - \alpha_\ell^j + y_i s_i + (1 - \widehat{\lambda}_i^j)\rho \geq \widehat{\mu}_i^j y_i \\
& \alpha_k^j - \alpha_\ell^j + y_i s_i + (1 - \widehat{\lambda}_i^j + u_i^U - \widehat{\mu}_i^j)\rho \geq u_i^U y_i
\end{aligned}\right\} \begin{aligned} & \forall i \in [n], \ \forall (k,\ell) \in \mathcal{E}_i^1 \\ & \forall j \in [N] \end{aligned} \\
& \rho \geq 0, \ \boldsymbol{s} \in \mathcal{S}.
\end{aligned} \tag{2.26}$$

In addition, when $p = 2$, (W-NS) *yields the same optimal value and the same set of optimal appointment schedules as the following second-order cone program:*

$$\begin{aligned}
\min_{\substack{\rho,\boldsymbol{s},\boldsymbol{\alpha} \\ \boldsymbol{\beta},\boldsymbol{\varphi}}} \ & \epsilon^2\rho + \frac{1}{N}\sum_{j=1}^{N}\left(\alpha_{S}^{j} - \alpha_{E}^{j}\right) \\
\text{s.t.} \ & \alpha_k^j - \alpha_\ell^j \geq 0 \ \ \forall (k,\ell) \in \mathcal{E}_E \ \ \forall j \in [N] \\
& \alpha_k^j - \alpha_\ell^j + y_i s_i + \left((\widehat{\mu}_i^j)^2 + \widehat{\lambda}_i^j\right)\rho \geq 0 \ \ \forall i \in [n], \ \forall (k,\ell) \in \mathcal{E}_i^0, \ \forall j \in [N] \\
& \alpha_k^j - \alpha_\ell^j + y_i s_i + (1 - \widehat{\lambda}_i^j)\rho - \varphi_{k\ell j} - (\widehat{\mu}_i^j - u_i^L)\beta_{k\ell j}^L - (u_i^U - \widehat{\mu}_i^j)\beta_{k\ell j}^U \geq \widehat{\mu}_i^j y_i \\
& \forall i \in [n], \ \forall (k,\ell) \in \mathcal{E}_i^1, \ \forall j \in [N] \\
& \left\| \begin{bmatrix} \beta_{k\ell j}^L - \beta_{k\ell j}^U + y_i \\ \varphi_{k\ell j} - \rho \end{bmatrix} \right\|_2 \leq \varphi_{k\ell j} + \rho \ \ \forall i \in [n], \ \forall (k,\ell) \in \mathcal{E}_i^1, \ \forall j \in [N] \\
& \rho \geq 0, \ \boldsymbol{s} \in \mathcal{S}.
\end{aligned}$$

We note that both reformulations in Theorem 7 involve $\mathcal{O}(KNn^2)$ decision variables and $\mathcal{O}(KNn^2)$ constraints. We further derive a worst-case probability distribution $\mathbb{Q}_{\boldsymbol{\xi}}^{\star}$ of the random no-shows and service durations that attains $\sup_{\mathbb{Q}_{\boldsymbol{\xi}} \in \mathcal{D}_1(\widehat{\mathbb{P}}_{\boldsymbol{\xi}}^N, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}}, \boldsymbol{\xi})]$. This distribution can be applied to stress test an appointment schedule generated from any decision-making processes.

Theorem 8. *For fixed $\bar{\boldsymbol{s}} \in \mathcal{S}$ and $\epsilon \geq 0$, $\sup\limits_{\mathbb{Q}_{\boldsymbol{\xi}} \in \mathcal{D}_1(\widehat{\mathbb{P}}_{\boldsymbol{\xi}}^N, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}}, \boldsymbol{\xi})]$ equals the optimal value* *of the following linear program:*

$$\max_{\boldsymbol{o},\boldsymbol{p},\boldsymbol{q},\boldsymbol{w},\boldsymbol{r}} \ \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n} y_i \left[\sum_{(k,\ell)\in\mathcal{E}_i^1} \left((u_i^L - \bar{s}_i)q_{k\ell j} + (\widehat{\mu}_i^j - \bar{s}_i)w_{k\ell j} + (u_i^U - \bar{s}_i)r_{i\ell j}\right) - \sum_{(k,\ell)\in\mathcal{E}_i^0} \bar{s}_i p_{k\ell j}\right] \tag{2.27a}$$

$$\text{s.t.} \ \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0}(\widehat{\mu}_i^j + \widehat{\lambda}_i^j)p_{k\ell j} + \sum_{(k,\ell)\in\mathcal{E}_i^1}\bigg(\big(1-\widehat{\lambda}_i^j + \widehat{\mu}_i^j - u_i^L\big)q_{k\ell j}$$

$$+ \big(1-\widehat{\lambda}_i^j\big)w_{k\ell j} + \big(1-\widehat{\lambda}_i^j + u_i^U - \widehat{\mu}_i^j\big)r_{k\ell j}\bigg)\Bigg] \le \epsilon \tag{2.27b}$$

$$\sum_{\ell:(k,\ell)\in\mathcal{E}^0} p_{k\ell j} + \sum_{\ell:(k,\ell)\in\mathcal{E}^1}(q_{k\ell j} + w_{k\ell j} + r_{k\ell j}) - \sum_{\ell:(\ell,k)\in\mathcal{E}^0} p_{\ell k j}$$

$$- \sum_{\ell:(\ell,k)\in\mathcal{E}^1}(q_{\ell k j} + w_{\ell k j} + r_{\ell k j}) = \begin{cases} 1 & \textit{if } k = S \\ 0 & \textit{if } k \neq S \end{cases} \quad \forall\, k \in \mathcal{E}\setminus\big(\mathcal{L}(n)\cup E\big),\ \forall\, j\in[N] \tag{2.27c}$$

$$o_{kEj} - \sum_{\ell:(\ell,k)\in\mathcal{E}^0} p_{\ell k j} - \sum_{\ell:(\ell,k)\in\mathcal{E}^1}(q_{\ell k j} + w_{\ell k j} + r_{\ell k j}) = 0 \ \ \forall\, k\in\mathcal{L}(n),\ \forall\, j\in[N] \tag{2.27d}$$

$$\sum_{k:(k,E)\in\mathcal{E}_E} o_{kEj} = 1 \ \ \forall\, j\in[N] \tag{2.27e}$$

$$o_{k\ell j},\ p_{k\ell j},\ q_{k\ell j},\ w_{k\ell j},\ r_{k\ell j} \ge 0 \ \ \forall\,(k,\ell)\in\mathcal{E},\ \forall\, j\in[N], \tag{2.27f}$$

where $\mathcal{E}^0 := \cup_{i=1}^n \mathcal{E}_i^0$ *and* $\mathcal{E}^1 := \cup_{i=1}^n \mathcal{E}_i^1$. *Let* $\{o_{k\ell j}^\star, p_{k\ell j}^\star, q_{k\ell j}^\star, w_{k\ell j}^\star, r_{k\ell j}^\star\}$ *be an optimal solution of the above linear program and define* $\mathcal{P} = \{\boldsymbol{z}\in\{0,1\}^{|\mathcal{E}|} : \text{(2.25b)}\}$. *Then, for all* $j\in[N]$*, there exists a distribution* $\mathbb{P}_{\boldsymbol{z}}^j$ *on* $\mathcal{P}$ *such that (i)* $\mathbb{P}_{\boldsymbol{z}}^j\{z_{k\ell}=1\} = p_{k\ell j}^\star$ *for all* $(k,\ell)\in\mathcal{E}^0$ *and (ii)* $\mathbb{P}_{\boldsymbol{z}}^j\{z_{k\ell}=1\} = q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star$ *for all* $(k,\ell)\in\mathcal{E}^1$. *Furthermore, define the probability distribution*

$$\mathbb{Q}_{\boldsymbol{\xi}}^\star = \frac{1}{N}\sum_{j=1}^{N}\sum_{\boldsymbol{\zeta}\in\mathcal{P}} \mathbb{P}_{\boldsymbol{z}}^j\{\boldsymbol{z}=\boldsymbol{\zeta}\}\delta_{\boldsymbol{\xi}^j(\boldsymbol{\zeta})},$$

where, for all $i\in[n]$ *and* $j\in[N]$,

$$\xi_i^j(\boldsymbol{\zeta}) = \sum_{(k,\ell)\in\mathcal{E}_i^0}\zeta_{k\ell}(0,0) + \sum_{(k,\ell)\in\mathcal{E}_i^1}\zeta_{k\ell}(\mu_{k\ell j},1) \ \textit{and} \ \mu_{k\ell j} = \widehat{\mu}_i^j + \frac{q_{k\ell j}^\star(u_i^L - \widehat{\mu}_i^j)}{q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star} + \frac{r_{k\ell j}^\star(u_i^U - \widehat{\mu}_i^j)}{q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star}.$$

Here we adopt an extended arithmetic given by $0/0 = 0$. *Then,* $\mathbb{Q}_{\boldsymbol{\xi}}^\star$ *belongs to the Wasser-*

stein ambiguity set $\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}, \epsilon)$ *and* $\mathbb{E}_{\mathbb{Q}^\star_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}}, \boldsymbol{\xi})] = \sup_{\mathbb{Q}_{\boldsymbol{\xi}} \in \mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}, \epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}}, \boldsymbol{\xi})]$.

Remark 4. *Once again, we can view the* (W-NS) *model as a two-person game between the AS scheduler and the nature. Theorem 8 provides an intuitive interpretation of how the nature picks the worst-case distribution* $\mathbb{Q}^\star_{\boldsymbol{\xi}}$. *Specifically,* $\mathbb{Q}^\star_{\boldsymbol{\xi}}$ *is a mixture of* N *distributions, i.e.,* $\mathbb{Q}^\star_{\boldsymbol{\xi}} = (1/N)\sum_{j=1}^N \mathbb{Q}^j_{\boldsymbol{\xi}}$, *where each* $\mathbb{Q}^j_{\boldsymbol{\xi}} := \sum_{\boldsymbol{\zeta} \in \mathcal{P}} \mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z} = \boldsymbol{\zeta}\}\delta_{\boldsymbol{\xi}^j(\boldsymbol{\zeta})}$ *pertains to the* j^{th} *data sample. Note that* $\mathcal{P}$ *consists of all the* S*–*E *paths in the network* $(\mathcal{G}, \mathcal{E})$. *Hence, what the nature does under* $\mathbb{Q}^j_{\boldsymbol{\xi}}$ *is to: (i) randomly pick an* S*–*E *path* $\boldsymbol{\zeta} \in \mathcal{P}$ *following distribution* $\mathbb{P}^j_{\boldsymbol{z}}$ *and (ii) for each* $i \in [n]$, *if the* i^{th} *arc of the path belongs to* $\mathcal{E}^0_i$ *(i.e., if* $(k, \ell) \in \mathcal{E}^0_i$*) then appointment* i *does not show up and accordingly the service duration equals zero; and if this arc belongs to* $\mathcal{E}^1_i$ *then appointment* i *shows up and lasts for* $\mu_{k\ell j}$ *long.*

2.4 Numerical Experiments

In this section, we demonstrate the effectiveness of the proposed approach via numerical experiments. Throughout these experiments, we adopt 1-Wasserstein ambiguity sets, i.e., $p = 1$. All computations are conducted on an 8-core 2.3 GHz Linux PC with 16 GB RAM. We implement the experiments in Python 2.7.12 and solve the linear programming problems by using Gurobi 8.0.1.

2.4.1 Random service durations

We begin with the (W-DRAS) model discussed in Section 2.2, where all appointments show up and the service durations are random. We consider $n = 10$ appointments and the unit waiting, idleness, and overtime costs are set to be $c_1 = \cdots = c_{10} = 2$, $d_1 = \cdots = d_{10} = 1$, and $C = 20$, respectively. Additionally, we consider each of the following three distributions to generate the data in our experiments:

- LN: The service duration u_i of each appointment i independently follows a lognormal distribution with mean μ_i and standard deviation σ_i, which are uniformly sampled from the intervals $[0.9, 1.1]$ and $[0.1, 0.9]$, respectively.
- UB: Each appointment i has a random service duration $u_i = 2\beta_i$, where β_i independently follows the (U-shaped) beta distribution $B(0.5, 0.5)$.
- NG: Each appointment i has a random service duration $u_i = \phi + \gamma_i$, where $\phi \geq 0$ follows a truncated normal distribution $N(1, 0.5^2)$ and γ_i independently follows the gamma distribution $\Gamma(\alpha, 1/\alpha)$ with α uniformly sampled from the interval $[0.5, 1]$. This is

to mimic a situation in which the service duration is influenced by both the service provider (modeled by the shared random variable ϕ) and the appointment characteristics (modeled by each individual γ_i).

It can be observed that the mean value of the service duration equals 1 in LN and UB, and is larger than 2 in NG. Accordingly, we set the time limit $T = 15$ in the LN and UB instances, and set $T = 30$ in the NG instances. Finally, we set $\boldsymbol{u}^{\text{L}}$ and $\boldsymbol{u}^{\text{U}}$ such that for all $i \in [n]$:

$$u_i^{\text{L}} := \min_{j \in [N]} \{\widehat{u}_i^j\} \quad \text{and} \quad u_i^{\text{U}} := \max_{j \in [N]} \{\widehat{u}_i^j\},$$

where $\{\widehat{\boldsymbol{u}}^j\}_{j=1}^N$ are the generated data sets.

2.4.1.1 Calibration of the Wasserstein ball radius

Our first experiment is to investigate the impact of the Wasserstein ball radius ϵ on the out-of-sample performance of the (W-DRAS) optimal solution, denoted by $\widehat{\boldsymbol{s}}(\epsilon, N)$, with respect to the data size N. Specifically, we evaluate the out-of-sample performance by computing

$$\mathbb{E}_{\mathbb{P}_{\text{approx}}}\Big[f(\widehat{\boldsymbol{s}}(\epsilon, N), \boldsymbol{u})\Big],$$

where $\mathbb{P}_{\text{approx}}$ represents an empirical distribution over a set of $100,000$ samples independently drawn from the true distribution $\mathbb{P}_{\boldsymbol{u}}$. In addition, we consider a discrete set of values $\Omega := \big\{0.01, 0.02, \cdots, 0.1, \cdots, 1, \cdots, 10\big\}$ for the selection of ϵ. For each $\epsilon \in \Omega$, we randomly sample 30 data sets of size $N \in \{5, 50, 500\}$ from each of the three distributions LN, UB, and NG. We solve an instance of the model (W-DRAS) via its LP reformulation (2.14) for each of the generated data sets and each of the candidate Wasserstein radius ϵ.

Figure 2.2 visualizes the impact of ϵ on the out-of-sample performance of $\widehat{\boldsymbol{s}}(\epsilon, N)$, which are derived over the data sets generated from distribution LN. Specifically, Figure 2.2 illustrates the tubes between the 20th and 80th percentiles (shaded areas) and the mean values (solid lines) of the out-of-sample performance $Z(\widehat{\boldsymbol{s}}(\epsilon, N))$ as a function of ϵ. The percentiles and mean values are estimated over the 30 independent simulation runs. We observe that the out-of-sample performance improves up to a critical ϵ value and then deteriorates. Hence, there exists a Wasserstein radius ϵ_{best} such that the corresponding optimal distributionally robust solutions have the lowest (i.e., best) out-of-sample performance. We note that same trends are observed from the data sets generated by distributions UB and NG.

In practice, however, a large data set is usually unavailable to construct $\mathbb{P}_{\text{approx}}$ and thus seeking ϵ_{best} by computing the out-of-sample performance is not viable. In this chapter, we

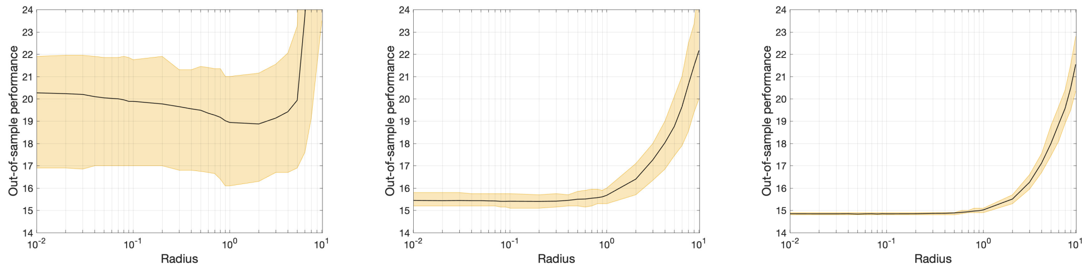


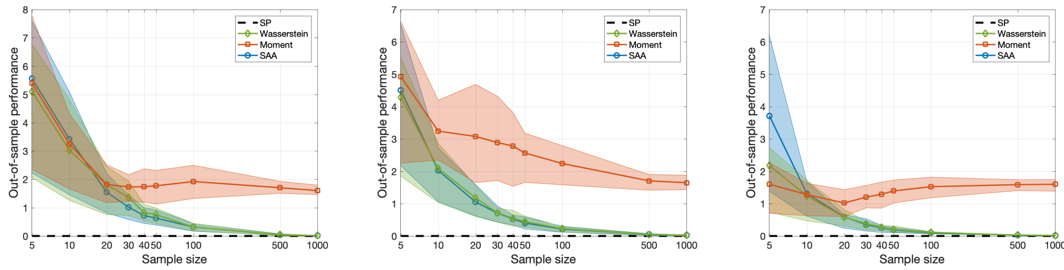
Figure 2.2: Out-of-sample performance as a function of the Wasserstein radius ϵ. The data sets are generated from distribution LN. Sample size: $N = 5$ (left), $N = 50$ (middle), and $N = 500$ (right).

implement a cross validation method that mimics the above out-of-sample evaluation procedure to approximate ϵ_{best} based on the N in-sample data. More specifically, we randomly partition the data $\{\widehat{\boldsymbol{u}}^j\}_{j=1}^N$ into two parts: a training data set consisting of $0.8N$ data and a validation data set consisting of the remaining $0.2N$ data. Using only the training data, we solve (W-DRAS) to obtain optimal solutions $\widehat{\boldsymbol{s}}(\epsilon, 0.8N)$ for each $\epsilon \in \Omega$. Then, we evaluate these solutions by computing $\mathbb{E}_{\mathbb{P}_{0.2N}}[f(\widehat{\boldsymbol{s}}(\epsilon, 0.8N), \boldsymbol{u})]$, where $\mathbb{P}_{0.2N}$ is the empirical distribution based only on the validation data, and we set $\widehat{\epsilon}_{\text{best}}^N$ to any ϵ that minimizes this quantity, i.e., $\widehat{\epsilon}_{\text{best}}^N \in \arg\min_{\epsilon \in \Omega}\{\mathbb{E}_{\mathbb{P}_{0.2N}}[f(\widehat{\boldsymbol{s}}(\epsilon, 0.8N), \boldsymbol{u})]\}$. Finally, we repeat this procedure for 30 random partitions and set ϵ to the average of the $\widehat{\epsilon}_{\text{best}}^N$ obtained from these 30 partitions.

2.4.1.2 Out-of-sample performance

We compare the out-of-sample performance of the (W-DRAS) approach with that of a cross-moment (CM) distributionally robust approach, in which the ambiguity set is characterized by the mean, variance, and correlation information [41]. This moment information is estimated from the data samples $\{\widehat{\boldsymbol{u}}^j\}_{j=1}^N$. We obtain optimal appointment schedules of the CM approach by solving the semidefinite programming approximations; see details in [41]. In addition, we compare with a sample average approximation (SAA) approach that solves model (2.3) with $\mathbb{P}_{\boldsymbol{u}}$ replaced by the empirical distribution $(1/N)\sum_{j=1}^N \delta_{\widehat{\boldsymbol{u}}^j}$. We generate data sets of size $N \in \{5, 10, 50, 100, 500, 1000\}$ from each of the distributions LN, UB, and NG. For each of the generated data sets, we solve an instance of (W-DRAS) with an ϵ set in the cross validation method, an instance of CM approach, and an instance of SAA approach.

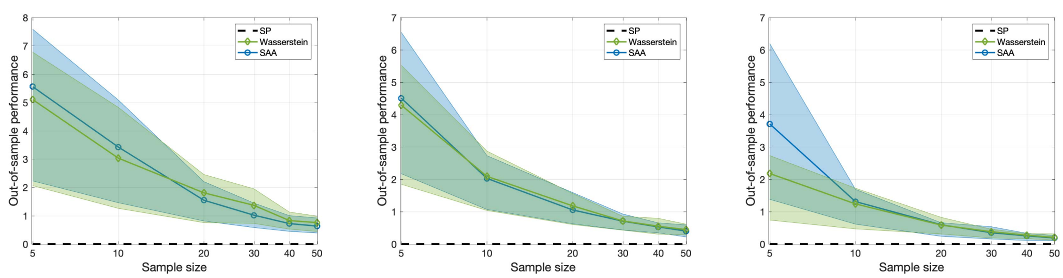
Figure 2.3 displays the tubes between the 20th and 80th percentiles (shaded areas) and the mean values (solid lines) of the out-of-sample performance as a function of the sample



(a) LN (b) UB (c) NG

Figure 2.3: Out-of-sample performance of optimal W-DRAS, CM, and SAA appointment schedules as a function of data size N

size N. The percentiles and mean values are estimated over 30 independent simulation runs. Note that the out-of-sample performance presented in Figure 2.3 are estimated by using the optimal solutions that minimize the W-DRAS, CM, and SAA problems, respectively. The horizontal dashed line represents $Z^\star$, the optimal value of the stochastic appointment scheduling model (2.3), in which $\mathbb{P}_{\boldsymbol{u}}$ is replaced with an empirical distribution based on $10,000$ scenarios.[4] From Figure 2.3, we observe that the out-of-sample performance of W-DRAS and SAA converge to $Z^\star$, while that of CM does not. This is consistent with the theoretical results in Theorem 1, confirming that the (W-DRAS) approach enjoys the asymptotic consistency. In contrast, the CM approach relies on the first two moments of the service durations and so the asymptotic consistency cannot be guaranteed in general.

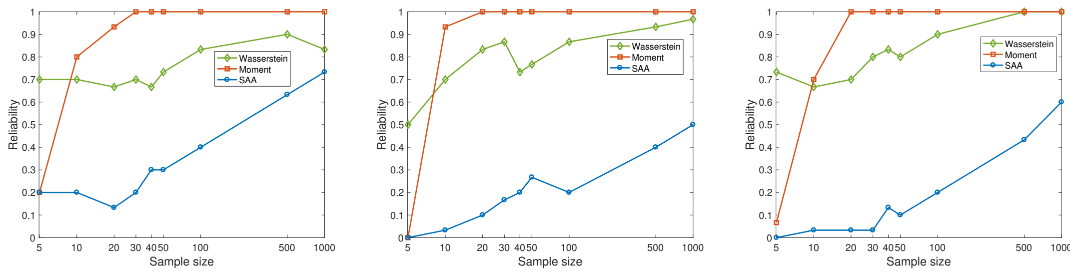


(a) LN (b) UB (c) NG

Figure 2.4: Out-of-sample performance of the W-DRAS and SAA optimal appointment schedules with small data sizes

In addition, we compare the out-of-sample performance of the W-DRAS and SAA approaches in Figure 2.4. From this figure, we observe that W-DRAS (slightly) outperforms

[4]Note that, for the convenience of making comparisons, we shifted the vertical coordinate of all points downwards by $Z^\star$ in Figure 2.3. As a consequence, the horizontal dashed line for $Z^\star$ appears with zero vertical coordinate, and a point with vertical coordinate 1, for example, represents an out-of-sample average cost higher than $Z^\star$ by 1 unit. We applied similar shifting operation in Figures 2.4, 2.6, 2.7a–2.7f, and 2.8.



(a) LN (b) UB (c) NG

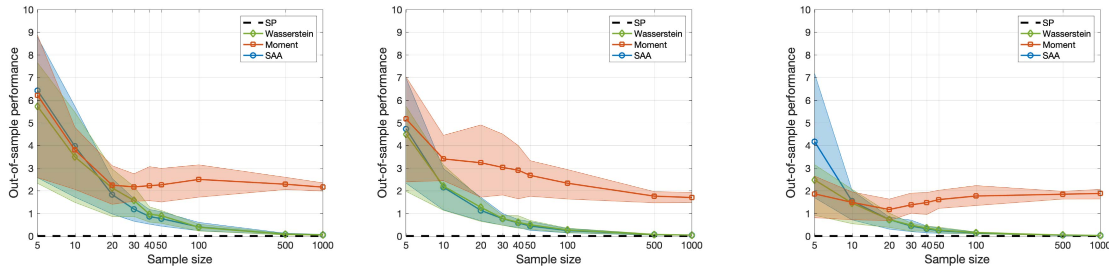
Figure 2.5: Reliability of W-DRAS, CM, and SAA as a function of data size N

SAA. Intuitively, this demonstrates that W-DRAS is capable to effectively learn distributional information even from a very limited amount of data (e.g., when $N = 5$ or 10). As a consequence, the proposed W-DRAS approach is particularly effective in AS systems with scarce service duration data.

Figure 2.5 displays the reliability of the three approaches, which is the empirical probability of the event that the optimal values of W-DRAS, CM, and SAA exceed the out-of-sample performance of the corresponding optimal solutions. The empirical probability is estimated over 30 independent simulation runs. From Figure 2.5, we observe that the reliability of W-DRAS and CM is consistently higher than that of SAA under all tested data sizes and across all tested generating distributions. For example, the reliability of CM increases to 100% once N exceeds 30, and that of W-DRAS is higher than 70% in most instances. In contrast, the reliability of SAA is generally lower than 50%, unless when N becomes large (e.g., $N \geq 500$). This is consistent with the theoretical results in Theorem 2, confirming that (W-DRAS) can provide a safe (upper bound) guarantee on the expected total cost even with a small data size.

2.4.1.3 Misspecified distributions

Another situation of interest is when the distribution of service durations in an AS system quickly varies due to, e.g., changes in service provider and/or appointee mix. As a consequence, the data we rely on to produce the appointment schedule may follow a misspecified distribution, i.e., one that is different from the true distribution. We conduct an experiment to examine the performance of the optimal W-DRAS, CM, and SAA appointment schedules by using data generated from a different distribution. Specifically, we use the same types of distributions as those generating the N in-sample data, but increase or decrease their parameters by $\sigma\%$ with σ uniformly sampled from $[5, 10]$. Figure 2.6 shows the performance of these appointment schedules under misspecified distributions. We observe



(a) Misspecified LN (b) Misspecified UB (c) Misspecified NG

Figure 2.6: Out-of-sample performance of the optimal W-DRAS, CM, and SAA appointment schedules under misspecified distributions

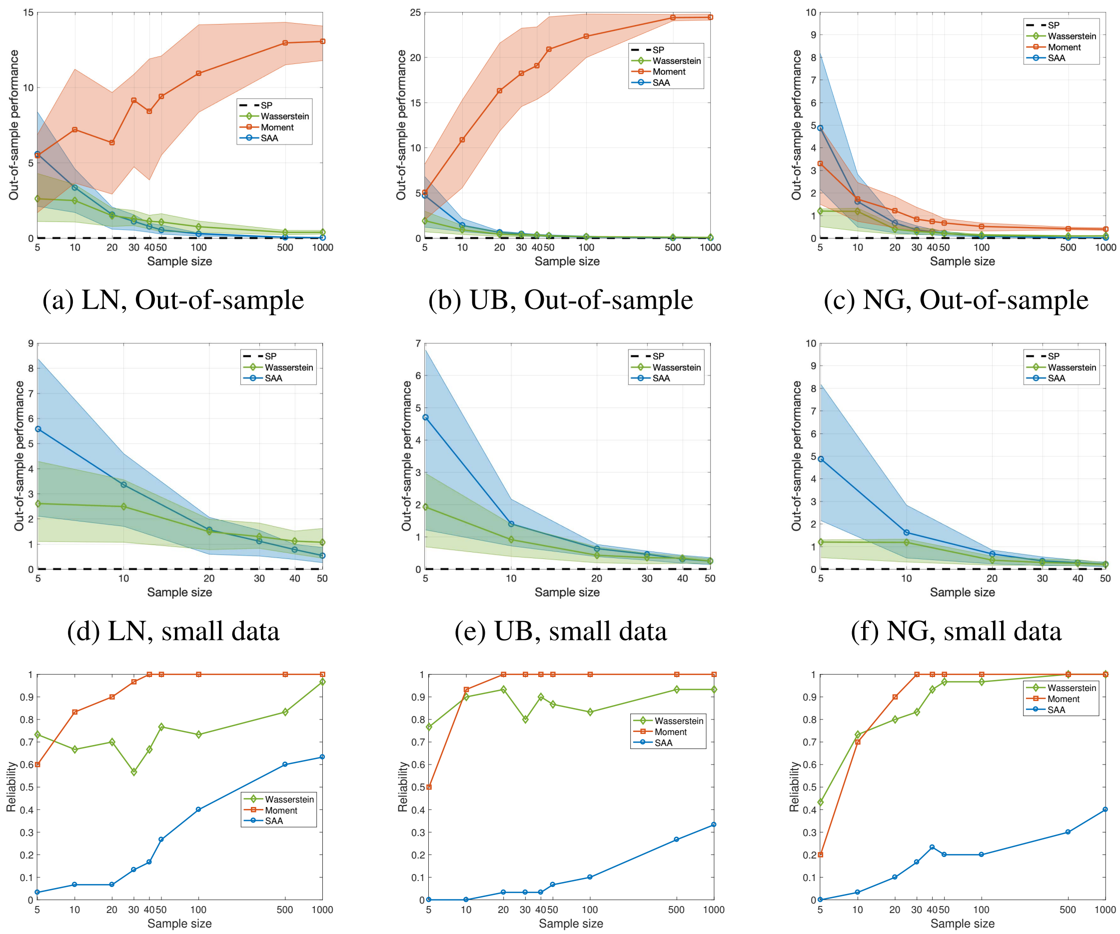
that the W-DRAS and SAA approaches still outperform the CM approach even under misspecified distributions. In addition, the W-DRAS approach still (slightly) outperforms the SAA approach when the data size is small. This demonstrates that the proposed W-DRAS approach is particularly effective in AS systems in quickly varying environments.

2.4.2 Random no-shows and service durations

We conduct numerical experiments to test the (W-NS) model discussed in Section 2.3, where both no-shows and service durations are random. We consider the same distributions (LN, UB, and NG) for the service durations as in Section 2.4.1. Meanwhile, we employ a Bernoulli distribution with parameter 0.4 for no-shows (i.e., each appointment does not show up with a probability of 0.4). In addition, we implement the same cross-validation method to calibrate the Wasserstein ball radius.

We compare the out-of-sample performance of our W-NS approach with a marginal-moment (MM) distributionally robust approach, which characterizes the ambiguity set based on the mean and support information of the random no-shows and service durations (see [44]). In addition, as in Section 2.4.1, we compare with the simple SAA approach.

We report the experiment results in Figure 2.7. In particular, Figures 2.7a–2.7c visualize the out-of-sample performance of optimal W-NS, MM, and SAA appointment schedules. From these figures, we observe that the out-of-sample performance of W-NS and SAA converge to the optimal value of the stochastic schedule model, while that of MM does not. This confirms that the (W-NS) approach enjoys the asymptotic consistency. In contrast, the MM approach does not have such convergence guarantee because its ambiguity set relies only on the mean and support information. Figures 2.7d–2.7f report the out-of-sample performance of the W-NS and SAA approaches when the data size is small. From these figures, we observe that W-NS outperforms SAA. Intuitively, this demonstrates that W-NS

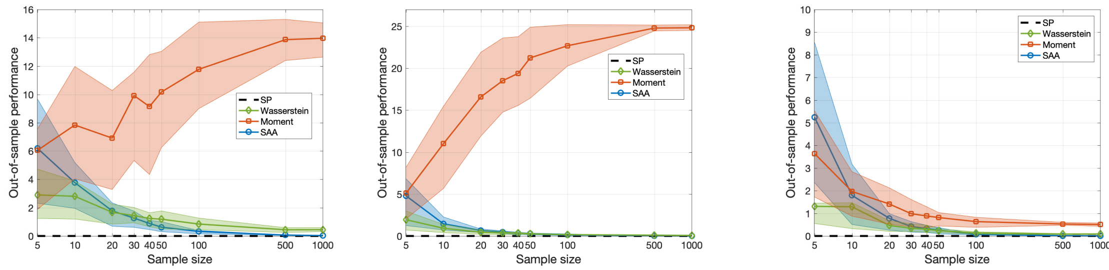


(a) LN, Out-of-sample (b) UB, Out-of-sample (c) NG, Out-of-sample

(d) LN, small data (e) UB, small data (f) NG, small data

(g) LN, reliability (h) UB, reliability (i) NG, reliability

Figure 2.7: Out-of-sample performance and reliability of the optimal W-NS, MM, and SAA appointment schedules as a function of data size N



(a) Misspecified LN (b) Misspecified UB (c) Misspecified NG

Figure 2.8: Out-of-sample performance of the optimal W-NS, MM, and SAA appointment schedules under misspecified distributions

is capable of learning the distributional information even from a limited amount of data (e.g., when $N \leq 20$). As a consequence, the proposed W-NS approach is particularly effective in AS systems with scarce no-show and service duration data. Figures 2.7g–2.7i report the reliability of the three approaches. These figures once again confirm our observations in Section 2.4.1 that the W-NS approach can provide a safe (upper bound) guarantee on the expected total cost even with a small data size.

We also conduct an experiment to examine out-of-sample performance of the optimal W-NS, MM, and SAA appointment schedules under misspecified distributions. This is particularly motivated by a situation where the no-show behaviors depend on the appointment schedule (see [43]). In that case, if we ignore the impact of the schedule and model the random no-shows by using a schedule-*independent* ambiguity set, then the true, schedule-*dependent* distribution may not belong with the ambiguity set. In this experiment, we use the same types of distributions as those generating the N in-sample data, but increase or decrease their parameters by $\sigma\%$ with σ uniformly sampled from [5, 10]. Figure 2.8 shows the performance of the W-NS, MM, and SAA appointment schedules under misspecified distributions. We observe that, once again, the W-NS and SAA approaches outperform the MM approach even under misspecified distributions. In addition, the W-NS approach still outperforms SAA when the data size is small. Finally, we observe that the out-of-sample performance of both W-NS and SAA quickly improve (i.e., tend to the optimal value of the true model) as the data size increases. This demonstrates that the proposed W-NS approach remains effective in AS systems under misspecified distributions, and the cost of ignoring the appointment dependency is limited.

2.5 Concluding Remarks

In this chapter, we studied a distributionally robust appointment scheduling problem over Wasserstein ambiguity sets. We proposed two models, with the first considering random service durations only and the second considering both random no-shows and service durations. We showed that both models can be recast as tractable convex programs under mild conditions on (i) support set of the uncertainties and (ii) penalty costs of the waiting time and server idleness. Through extensive numerical experiments, we demonstrated that the proposed approaches enjoy both asymptotic consistency and finite-data guarantees, and are particularly suitable for the AS systems with scarce data or in quickly varying environments.

CHAPTER 3

Nurse Staffing under Absenteeism: A Distributionally Robust Optimization Approach

3.1 Introductory Remarks

Nurse staffing plays a key role in hospital management. The cost of staffing nurses accounts for over 30% of the overall hospital annual expenditures (see, e.g., [103]). Besides, the nurse staffing level make significant impacts on patient safety, quality of care, and the job satisfaction of nurses (see, e.g., [104]). In view of that, a number of governing agencies (e.g., the California Department of Health [105] and the Victoria Department of Health [106]) have set up minimum nurse-to-patient ratios (NPRs) for various types of hospital units to regulate the staffing decision.

In general, the nurse planning consists of the following four phases: (1) nurse demand forecasting and staffing, (2) nurse shift scheduling, (3) pre-shift staffing and re-scheduling, and (4) nurse-patient assignment (see [107, 108, 109, 110]). In particular, phase (1) takes place weeks or months ahead of a shift and determines the nurse staffing levels based on, e.g., the forecasted patient census and the NPRs; and phase (3) takes place hours before the shift and recruits additional workforce (e.g., temporary or off-duty nurses) if any units are short of nurses. In this chapter, we focus on these two phases and refer the corresponding decision making process as nurse staffing. The outputs of our study (e.g., the nurse staffing levels) can be used in phases (2) and (4) to generate shift schedules and assignments of the nurses.

Nurse staffing is a challenging task, largely because of the uncertainties of nurse demand and absenteeism. The demand uncertainty stems from the random patient census and has been well documented (see, e.g., [111, 112]) and studied in the nurse staffing literature (see, e.g., [45, 36]). In contrast, the absenteeism uncertainty has received relatively less attention in this literature (see, e.g., [37, 113]), albeit commonly observed in practice. For example, according to the U.S. Bureau of Labor Statistics [114], the average absence rate among all nurses in the Veterans Affairs Health Care System is 6.4% [115], significantly higher than that among all occupations (2.9%) and among health-care support occupations (4.3%). For the quality of care, many hospitals have developed float pools of nurses by cross-training, so that in phase (3) a pool nurse can be assigned to the units short of nurses (see, e.g., [116]).

Unlike the demand, the random number of nurses who show up for a shift partially depends on the nurse staffing level, i.e., the absenteeism uncertainty is *endogenous*. For example, if the nurse staffing level is $w \in \mathbb{N}_+$ then the random number of nurses who show up cannot exceed w. Although failing to incorporate such endogeneity may result in understaffing (see [37]), unfortunately, modeling endogeneity usually makes optimization models computationally prohibitive (see, e.g., [8]). Due to this technical difficulty, the endogenous uncertainty has received much less attention in the literature of stochastic optimization than the exogenous uncertainty. Existing works often resort to exogenous uncertainty for an approximate solution. Alternatively, they employ certain parametric probability distributions to model the endogenous uncertainty (see [8]), e.g., the absence of each nurse follows *independent* Bernoulli distribution with the *same* probability (which may depend on the staffing level; see [37]). A basic challenge to adopting parametric models is that a complete and accurate knowledge of the endogenous probability distribution is usually unavailable. Under many circumstances, we only have historical data, including the nurse staffing level and the corresponding absence records, which can be considered as samples taken from the true (but ambiguous) endogenous distribution. As a result, the solution obtained by assuming a parametric model can yield unpleasant out-of-sample performance if the chosen model is biased.

In this chapter (see more details in [117]), we propose an alternative, nonparametric model of both exogenous and endogenous uncertainties based on distributionally robust optimization (DRO). Our approach considers a family of probability distributions, termed an ambiguity set, based only on the support and moment information of these uncertainties. In particular, the number of nurses who show up in a unit/pool is bounded by the corresponding staffing level and its mean value is a function of this level. Then, we employ this ambiguity set in a two-stage distributionally robust nurse staffing (DRNS) model that imitates the decision making process in phases (1) and (3). Building on DRNS, we further search for *sparse* pool structures that result in a minimum amount of cross-training while achieving a specified target staffing cost. To the best of our knowledge, this is the first study of the endogenous uncertainty in nurse staffing by using a DRO approach.

3.1.1 Literature review

A vast majority of the nurse staffing literature focuses on deterministic models that do not take into account the randomness of the nurse demand and/or absenteeism (see [118]). Various (deterministic) optimization models have been employed, including linear programming (see, e.g., [119, 120]) and mixed-integer programming (see, e.g., [121, 122, 123, 110]). For example, [119] assessed the need for hiring permanent staffs and temporary helpers and [121] analyzed the trade-offs among hiring full-time, part-time, and overtime nurses. More recently, [122] compared cross-training and flexible work days and demonstrated that cross-training is far more effective for performance improvement than flexible work days. Similarly, [110] identified cross-training as a promising extension from their deterministic model. Despite the potential benefit of operational flexibility brought by float pools and cross-training, [123] pointed out that the pool design and staffing are often made manually in a qualitative fashion (also see [124]). In addition, when the nurse demand and/or absenteeism is random, the deterministic models may underestimate the total staffing cost (see, e.g., [125]).

Existing stochastic nurse staffing models often consider the demand uncertainty only. For example, [126] studied a two-stage stochastic programming model that integrates the staffing and scheduling of cross-trained workers (e.g., nurses) under demand uncertainty. Through numerical tests, [126] demonstrated that cross-training can be even more valuable than the perfect demand information (i.e., knowing the realization of demand when making staffing decisions). In addition, [127] studied how the mandatory overtime laws can negatively effect the service quality of a nursing home. Using a two-stage stochastic programming model under demand uncertainty, [127] pointed out that these laws result in a lower staffing level of permanent registered nurses and a higher staffing level of temporary registered nurses. Unfortunately, as [37] pointed out, ignoring nurse absenteeism may result in understaffing, which reduces the service quality and increases the operational cost because additional temporary nurses need to be called in.

When the nurse absenteeism is taken into account, the stochastic optimization models become unscalable. [37] considered the staffing of a single unit under both nurse demand and absenteeism uncertainty and successfully derived a closed-form optimal staffing level. In addition, [128] studied the staffing of a single on-call pool that serves multiple units whose staffing levels are fixed and known. In a setting that regular nurses can be absent while pool nurses always show up, the authors successfully derived a closed-form optimal pool staffing level. Unfortunately, the problem becomes computationally prohibitive when multiple units and/or multiple float pools are incorporated. For example, [129] studied a multi-unit and one-pool setting[1]. The author showed that the proposed stochastic optimization model outperforms the (deterministic) mean value approximation. However, the evaluation of this model "does not scale well." More specifically, even when staffing levels are *fixed*, one needs to solve an exponential number (in terms of the staffing level) of linear programs to evaluate the expected total cost of staffing. This renders the search of an optimal staffing level so challenging that one has to resort to heuristics. [115] considered a multi-unit and no-pool setting and analyzed the staffing problem based on a cohort of nurses who have heterogeneous absence rates. The authors showed that the staffing cost is lower when the nurses are heterogenous within each unit but uniform across units. Unfortunately, searching for an optimal staffing strategy is "computationally demanding" with a large number of nurses. Similar to [129], [115] resort to easy-to-use heuristics.

To mitigate the computational challenges of nurse absenteeism, the existing literature often make parametric assumptions on the endogenous probability distribution. For example, [37, 129, 115] assumed that the absences of all nurses are stochastically *independent* and the absence rate in [37, 129] is assumed *homogeneous*. But the nurse absences may be positively correlated during extreme weather (e.g., heavy snow) or during day shifts (e.g., due to conflicting family obligations). In addition, the data analytic in [115] suggests that the nurses actually have heterogeneous absence rates. Furthermore, the absenteeism can be drastically different among different units/hospitals, and even within the same unit/hospital, has high temporal variations. For example, based on the data from different hospitals, [37] concluded that the absence rate depends on the staffing level and ignoring such dependency results in understaffing, while [115] concluded that such dependency is insignificant. A fundamental challenge to adopting parametric models is that the solution thus obtained can yield suboptimal out-of-sample performance if the adopted model is biased. In this chapter, we take into account both nurse demand and absenteeism uncertainty in a multi-unit and multi-pool setting. To address the challenges on computational scalability and out-of-sample performance, we propose an alternative nonparametric model based on DRO. In particular, this model allows dependence or independence between the absence rate and the staffing level. Moreover, our model can be solved to global optimality by a separation algorithm and, in several important special cases, by solving a single mixed-integer linear program (MILP).

DRO models have received increasing attention in the recent literature. In particular, as in this chapter, DRO has been applied to model two-stage stochastic optimization problems

[1]More precisely, the model in [129] allows to re-assign nurses from one unit to any other unit. In the context of this chapter, that is equivalent to having a single float pool that serves all the units and assigning all nurses to this pool.

(see, e.g., [31, 130, 131]). In general, the two-stage DRO models are computationally prohibitive. For example, suppose that the second-stage formulation is linear and continuous with right-hand side uncertainty. Then, even with *fixed* first-stage decision variables, [31] showed that evaluating the objective function of the DRO model is NP-hard. To mitigate the computational challenge, [131, 132] recast the two-stage DRO model as a copositive program, which admits semidefinite programming approximations. In addition, [133, 130] applied linear decision rules (LDRs) to obtain conservative and tractable approximations. In contrast to these work, our second-stage formulation involves integer variables to model the pre-shift staffing. Besides undermining the convexity of our formulation, this prevents us from applying the LDRs because fractional staffing levels are not implementable. To the best of our knowledge, there are only two existing work [27, 26] on DRO with endogenous uncertainty. Specifically, [27] derived equivalent reformulations of the endogenous DRO model under various ambiguity sets, and [26] applied an endogenous DRO model on the machine scheduling problem. In this chapter, we study a two-stage endogenous DRO model for nurse staffing and derive tractable reformulations under several practical float pool structures. We summarize our main contributions as follows:

1. We propose the first DRO approach for nurse staffing, considering both exogenous nurse demand and endogenous nurse absenteeism. The proposed two-stage endogenous DRO model considers multiple units, multiple float pools, and both long-term and pre-shift nurse staffing. For arbitrary pool structures, we derive a min-max reformulation of the model and a separation algorithm that solves this model to global optimality.

2. For multiple pool structures that often arise in practice, including one pool, disjoint pools, and chained pools, we provide a monolithic MILP reformulation of our DRO model by deriving strong valid inequalities. The binary variables of this MILP reformulation arise from the nurse staffing decisions only. That is, under these practical pool structures, the computational burden of our DRO approach is de facto the same as that of the deterministic nurse staffing.

3. Building upon the DRO model, we further study how to design sparse and effective disjoint pools. To this end, we proactively optimize the nurse pool structure to minimize the total number of cross-training, while providing a guarantee on the staffing cost.

4. We conduct extensive case studies based on the data and insights from our collaborating hospital. The results demonstrate the value of modeling nurse absenteeism and the computational efficacy of our DRO approach. In addition, we provide managerial insights on how to design sparse and effective pools.

The remainder of the chapter is organized as follows. In Section 3.2, we describe the two-stage DRO model with endogenous nurse absenteeism. In Section 3.3, we derive a solution approach for this model under arbitrary pool structures. In Section 3.4, we derive strong valid inequalities and tractable reformulations under special pool structures. We extend the DRO model for optimal pool design in Section 3.5, conduct case studies in Section 3.6, and conclude in Section 3.7. To ease the exposition, we relegate all proofs to the appendices.

Notation: We use $\sim$ to indicate random variables and $\wedge$ to indicate realizations of the random variables. For example, $\tilde{d}$ represents a random variable and $\hat{d}^1, \ldots, \hat{d}^N$ represent N realizations of $\tilde{d}$. For $a, b \in \mathbb{Z}$, we define $[a] := \{1, 2, \ldots, a\}$ and $[a, b]_{\mathbb{Z}} := \{n \in \mathbb{Z} : a \leq n \leq b\}$. For $x \in \mathbb{R}$, we define $[x]_+ = \max\{x, 0\}$. For set S, we define its indicator function $\mathbb{1}_S$ such that $\mathbb{1}_S(s) = 1$ if $s \in S$ and $\mathbb{1}_S(s) = 0$ if $s \notin S$, and denote its convex hull by conv(S).

3.2 Distributionally Robust Nurse Staffing

We consider a group of J hospital units, each facing a random demand of nurses denoted by $\tilde{d}_j$ for all $j \in [J]$. To enhance the operational flexibility, the manager forms I nurse float pools. For all $i \in [I]$, pool i is associated with a set P_i of units and each nurse assigned to this pool is capable of working in any unit $j \in P_i$. Due to random absenteeism, if we staff unit j with w_j nurses (termed unit nurses), then there will be a random number $\tilde{w}_j$ of nurses showing up for work, where $\tilde{w}_j \in [0, w_j]_{\mathbb{Z}}$. Likewise, $\tilde{y}_i$ nurses show up if we staff pool i with y_i nurses, where $\tilde{y}_i \in [0, y_i]_{\mathbb{Z}}$. After the uncertain parameters $\tilde{d}_j$, $\tilde{w}_j$, and $\tilde{y}_i$ are realized, the nurses showing up in pool i can be re-assigned to any units in P_i to make up the nurse shortage, if any. After the re-assignment, any remaining shortage will be covered by hiring temporary nurses in order to meet the NPR requirement. Mathematically, for given $\tilde{w} := [\tilde{w}_1, \ldots, \tilde{w}_J]^\top$, $\tilde{y} := [\tilde{y}_1, \ldots, \tilde{y}_I]^\top$, and $\tilde{d} := [\tilde{d}_1, \ldots, \tilde{d}_J]^\top$, the total operational cost can be obtained from solving the following integer program:

$$V(\tilde{w}, \tilde{y}, \tilde{d}) = \min_{z,x,e} \sum_{j=1}^{J} \big(c^{\mathrm{x}} x_j - c^{\mathrm{e}} e_j\big) \tag{3.1a}$$

$$\text{s.t.} \sum_{i:\, j \in P_i} z_{ij} + x_j - e_j = \tilde{d}_j - \tilde{w}_j, \ \ \forall j \in [J], \tag{3.1b}$$

$$\sum_{j \in P_i} z_{ij} \leq \tilde{y}_i, \ \ \forall i \in [I], \tag{3.1c}$$

$$x_j, e_j \in \mathbb{Z}_+, \ \ \forall j \in [J], \ \ z_{ij} \in \mathbb{Z}_+, \ \ \forall i \in [I], \forall j \in P_i, \tag{3.1d}$$

where variables z_{ij} represent the number of nurses re-assigned from pool i to unit j, variables x_j represent the number of temporary nurses hired in unit j, variables e_j represent the excessive number of nurses in unit j, parameter c^{x} represents the unit cost of hiring temporary nurses, and parameter c^{e} represents the unit benefit of having excessive nurses. We can set c^{e} to be zero when such benefit is not taken into account. In the above formulation, objective function (3.1a) minimizes the cost of hiring temporary nurses minus the benefit of having excessive nurses. Constraints (3.1b) describe three ways of satisfying the nurse demand in each unit: (i) assigning unit nurses, (ii) re-assigning pool nurses, and (iii) hiring temporary nurses. Constraints (3.1c) ensure that the number of nurses re-assigned from each pool does not exceed the number of nurses showing up in that pool. Constraints (3.1d) describe integrality restrictions.

In reality, it is often challenging to obtain an accurate estimate of the true probability distribution $\mathbb{P}_{\tilde{w},\tilde{y},\tilde{d}}$ of $(\tilde{w}, \tilde{y}, \tilde{d})$. For example, the historical data of the nurse demand (via patient census and NPRs) can typically be explained by multiple (drastically) different distributions. More importantly, because of the endogeneity of $\tilde{w}$ and $\tilde{y}$, $\mathbb{P}_{\tilde{w},\tilde{y},\tilde{d}}$ is in fact a *conditional distribution* depending on the nurse staffing levels. This further increases the difficulty of estimation. Using a biased estimate of $\mathbb{P}_{\tilde{w},\tilde{y},\tilde{d}}$ can yield post-decision disappointment. For example, if one simply ignores the endogeneity of $\tilde{w}$ and $\tilde{y}$ and employs their empirical distribution based on historical data, then the nurse staffing thus obtained may lead to disappointing out-of-sample performance. In this chapter, we assume that $\mathbb{P}_{\tilde{w},\tilde{y},\tilde{d}}$ is ambiguous and it belongs to the following moment ambiguity set:

$$\mathcal{D} = \Big\{\mathbb{P} \in \mathcal{P}(\Xi) : \ \mathbb{E}_{\mathbb{P}}[\tilde{d}_j^q] = \mu_{jq}, \ \ \forall j \in [J], \ \forall q \in [Q], \tag{3.2a}$$

$$\mathbb{E}_{\mathbb{P}}[\tilde{w}_j] = f_j(w_j), \ \ \forall j \in [J], \tag{3.2b}$$

$$\mathbb{E}_{\mathbb{P}}[\tilde{y}_i] = g_i(y_i), \ \ \forall i \in [I]\Big\}, \tag{3.2c}$$

where Ξ represents the support of $(\tilde{w}, \tilde{y}, \tilde{d})$ and $\mathcal{P}(\Xi)$ represents the set of probability distribution supported on Ξ. We consider a box support $\Xi := \Xi_{\tilde{w}} \times \Xi_{\tilde{y}} \times \Xi_{\tilde{d}}$, where $\Xi_{\tilde{w}} = \Pi_{j=1}^{J}[0, w_j]_{\mathbb{Z}}$, $\Xi_{\tilde{y}} = \Pi_{i=1}^{I}[0, y_i]_{\mathbb{Z}}$, $\Xi_{\tilde{d}} = \Pi_{j=1}^{J}[d_j^{\mathrm{L}}, d_j^{\mathrm{U}}]_{\mathbb{Z}}$, and d_j^{L} and d_j^{U} represent lower and upper bounds of the nurse demand in unit j. In addition, for $Q \in \mathbb{N}_+$, all $q \in [Q]$, and all $j \in [J]$, μ_{jq} represents the q^{th} moment of $\tilde{d}_j$. Furthermore, for all $j \in [J]$ and $i \in [I]$, $f_j : \mathbb{N}_+ \to \mathbb{R}_+$ and $g_i : \mathbb{N}_+ \to \mathbb{R}_+$ represent two functions such that $f_j(0) = g_i(0) = 0$. We note that these functions can model arbitrary dependence of $(\tilde{w}, \tilde{y})$ on the staffing levels, and the assumption $f_j(0) = g_i(0) = 0$ ensures that if we assign no nurses in a unit/pool then nobody will show up.

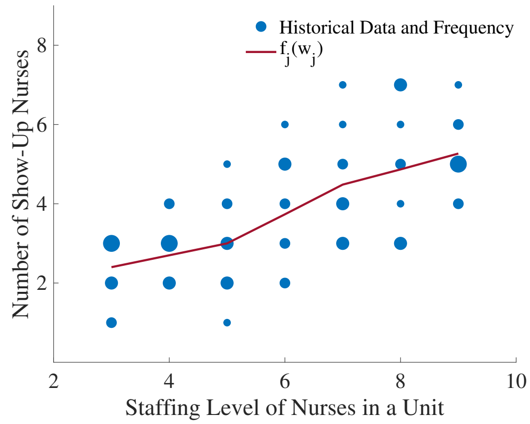


Figure 3.1: An example of segmented linear regression. Dots represent historical data samples and the size of dots indicate frequency.

The ambiguity set $\mathcal{D}$ can be conveniently calibrated. First, suppose that $\mathbb{P}_{\tilde{w},\tilde{y},\tilde{d}}$ is observed through nurse demand data $\{\hat{d}_j^1, \ldots, \hat{d}_j^N\}_{j=1}^J$ and attendance records $\{(w_j^1, \hat{w}_j^1), \ldots, (w_j^N, \hat{w}_j^N)\}_{j=1}^J$ and $\{(y_i^1, \hat{y}_j^1), \ldots, (y_i^N, \hat{y}_i^N)\}_{i=1}^I$ during the past N days, where, in each pair $(w_j^n, \hat{w}_j^n)$, w_j^n represents the staffing level of unit j in day n and $\hat{w}_j^n$ represents the corresponding number of nurses who actually showed up. Then, μ_{jq} can be obtained from empirical estimates (e.g., $\mu_{j1} = (1/N)\sum_{n=1}^N \hat{d}_j^n$, $\mu_{j2} = (1/N)\sum_{n=1}^N (\hat{d}_j^n)^2$, etc.), and f_j and g_i can be obtained by performing segmented linear regression on the attendance data, using the staffing levels $\{w_j^1, \ldots, w_j^N\}$ and $\{y_i^1, \ldots, y_i^N\}$ as breakpoints, respectively (see Figure 3.1 for an example). Second, if $\tilde{w}$ and $\tilde{y}$ are believed to follow certain parametric models, then we can follow such models to calibrate $\{f_j(w_j)\}_{j=1}^J$ and $\{g_i(y_i)\}_{i=1}^I$. For example, if $\tilde{w}_j$ is modeled as a Binomial random variable $B(w_j, 1-a(w_j))$ as in [37], where $a(w_j)$ represents the absence rate, i.e., the probability of any scheduled nurse in unit j being absent from work, then we have $f_j(w_j) = w_j(1 - a(w_j))$.

We seek nurse staffing levels that minimize the expected total cost with regard to the worst-case probability distribution in $\mathcal{D}$, i.e., we consider the following two-stage DRO model:

$$
\begin{aligned}
(\textbf{DRNS}):\quad & \min_{w,y} \ \sum_{j=1}^J c^{\mathrm{w}} w_j + \sum_{i=1}^I c^{\mathrm{y}} y_i + \sup_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}\left[V(\tilde{w},\tilde{y},\tilde{d})\right] && \text{(3.3a)}\\
& \text{s.t. } \ w_j^{\mathrm{L}} \le w_j \le w_j^{\mathrm{U}}, \ \ \forall j \in [J], && \text{(3.3b)}\\
& \qquad y_i^{\mathrm{L}} \le y_i \le y_i^{\mathrm{U}}, \ \ \forall i \in [I], && \text{(3.3c)}\\
& \qquad y, w \in R \cap \mathbb{Z}_+^{I+J}, && \text{(3.3d)}
\end{aligned}
$$

where parameters c^{w} and c^{y} represent the unit cost of hiring unit and pool nurses, respectively, constraints (3.3b)–(3.3c) designate lower and upper bounds on staffing levels, and set R represents all remaining restrictions, which we assume can be represented via mixed-integer linear inequalities. (DRNS) is computationally challenging because (i) $\mathcal{D}$ involves exponentially many probability distributions, all of which depend on the decision variables w_j and y_i and (ii) it is a two-stage DRO model with integer recourse variables. In the next two sections, we shall derive equivalent reformulations of (DRNS) that facilitate a separation algorithm, and identify practical pool structures that admit more tractable solution approaches.

3.3 Solution Approach: Arbitrary Pool Structure

In this section, we consider arbitrary pool structures, recast (DRNS) as a min-max formulation, and derive a separation algorithm that solves this model to global optimality.

We start by noticing that the integrality restrictions (3.1d) in the second-stage formulation of (DRNS) can be relaxed without loss of generality.

Lemma 4. *For any given* $(\tilde{w}, \tilde{y}, \tilde{d}) \in \Xi$*, the value of* $V(\tilde{w}, \tilde{y}, \tilde{d})$ *remains unchanged if constraints* (3.1d) *are replaced by non-negativity restrictions, i.e.,* $x_j, e_j \geq 0, \forall j \in [J]$ *and* $z_{ij} \geq 0, \forall i \in [I],\ \forall j \in P_i$*.*

Thanks to Lemma 4, we are able to rewrite $V(\tilde{w}, \tilde{y}, \tilde{d})$ as the following dual formulation:

$$V(\tilde{w}, \tilde{y}, \tilde{d}) \;=\; \max_{\alpha,\beta} \;\; \sum_{j=1}^{J}(\tilde{d}_j - \tilde{w}_j)\alpha_j + \sum_{i=1}^{I} \tilde{y}_i\beta_i \tag{3.4a}$$

$$\text{s.t.}\;\; \beta_i + \alpha_j \leq 0, \;\; \forall i \in [I],\ \forall j \in P_i, \tag{3.4b}$$

$$c^{\text{e}} \leq \alpha_j \leq c^{\text{x}}, \;\; \forall j \in [J], \tag{3.4c}$$

where dual variables α_j and β_i are associated with primal constraints (3.1b) and (3.1c), respectively, and dual constraints (3.4b) and (3.4c) are associated with primal variables z_{ij} and (x_j, e_j), respectively. We let Λ denote the dual feasible region for variables (α, β), i.e., $\Lambda := \{(\alpha, \beta) : \text{(3.4b)–(3.4c)}\}$. Strong duality between formulations (3.1a)–(3.1d) and (3.4a)–(3.4c) hold valid because (3.1a)–(3.1d) has a finite optimal value.

We are now ready to recast (DRNS) as a min-max formulation. To this end, we consider $\mathbb{P}$ as a decision variable and take the dual of the worst-case expectation in (3.3a). For strong duality, we make the following technical assumption on the ambiguity set $\mathcal{D}$.

Assumption 4. *For any given* $w := [w_1, \dots, w_J]^\top$ *and* $y := [y_1, \dots, y_I]^\top$ *that are feasible to (DRNS),* $\mathcal{D}$ *is non-empty.*

Assumption 4 is mild. For example, it holds valid whenever the moments of demands $\{\mu_{jq} : j \in [J], q \in [Q]\}$ are obtained from empirical estimates and the decision-depend moments $\{g_i(y_i), f_j(w_j) : i \in [I], j \in [J]\}$ lie in the convex hull of their support, i.e., $f_j(w_j) \in [0, w_j]$ and $g_i(y_i) \in [0, y_i]$. In Appendix B.2, we present an approach to verify Assumption 4 by solving J linear programs. The reformulation is summarized in the following proposition.

Proposition 6. *Under Assumption 4, the (DRNS) model* (3.3) *yields the same optimal value and the same set of optimal solutions as the following min-max optimization problem:*

$$\min_{\substack{w,y\\ \gamma,\lambda,\rho}} \max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta) + \sum_{i=1}^{I} (c^{y} y_i + g_i(y_i)\lambda_i) + \sum_{j=1}^{J} \left[c^{w} w_j + \sum_{q=1}^{Q} \mu_{jq}\rho_{jq} + f_j(w_j)\gamma_j \right] \tag{3.5a}$$

$$\text{s.t. } (3.3b)\text{–}(3.3d), \tag{3.5b}$$

where

$$F(\alpha,\beta) := \sum_{j=1}^{J} \Big[(-\alpha_j - \gamma_j) w_j\Big]_+ + \sum_{i=1}^{I} \Big[(\beta_i - \lambda_i) y_i\Big]_+ + \sum_{j=1}^{J} \sup_{\tilde{d}_j \in [d_j^L, d_j^U]_{\mathbb{Z}}} \Big\{ \alpha_j \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\}. \tag{3.5c}$$

In the min-max reformulation (3.5a)–(3.5b), the additional variables γ, λ, ρ are generated in the process of taking dual. In addition, function $F(\alpha, \beta)$ is jointly convex in (α, β) because, as presented in (3.5c), $F(\alpha, \beta)$ is the pointwise maximum of functions affine in (α, β). This min-max reformulation is not directly computable because (i) for fixed $(w, y, \gamma, \lambda, \rho)$, evaluating the objective function (3.5a) needs to solve a *convex maximization* problem $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$, which is in general NP-hard, and (ii) the formulation includes nonlinear and non-convex terms $g_i(y_i)\lambda_i$ and $f_j(w_j)\gamma_j$. We shall address these two challenges before presenting a separation algorithm for solving (DRNS).

First, we analyze the convex maximization problem $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ and derive the following optimality conditions.

Lemma 5. *For fixed* $(w, y, \gamma, \lambda, \rho)$*, there exists an optimal solution* $(\bar{\alpha}, \bar{\beta})$ *to problem* $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ *such that (a)* $\bar{\alpha}_j \in \{c^{e}, c^{x}\}$ *for all* $j \in [J]$ *and (b)* $\bar{\beta}_i = -\max\{\bar{\alpha}_j : j \in P_i\}$ *for all* $i \in [I]$.

Lemma 5 enables us to avoid enumerating the infinite number of elements in Λ and focus only on a finite set of $(\bar{\alpha}, \bar{\beta})$ values. In addition, we introduce binary variables to encode the special structure identified in the optimality conditions. Specifically, for all $j \in [J]$, we define binary variables t_j such that $t_j = 1$ if $\bar{\alpha}_j = c^{\text{x}}$ and $t_j = 0$ if $\bar{\alpha}_j = c^{\text{e}}$; and for all $i \in [I]$ and $j \in P_i$, binary variables $s_{ij} = 1$ if j is the largest index in P_i such that $t_j = 1$ (i.e., $\bar{\alpha}_j = -c^{\text{x}}$ and $\bar{\alpha}_\ell = -c^{\text{e}}$ for all $\ell \in P_i$ and $\ell \geq j+1$) and $s_{ij} = 0$ otherwise. Variables (t, s) need to satisfy the following constraints to make the encoding well-defined:

$$\sum_{j \in P_i} s_{ij} \leq 1, \ \ \forall i \in [I], \tag{3.6a}$$

$$s_{ij} \leq t_j, \ \ \forall i \in [I], \ \forall j \in P_i, \tag{3.6b}$$

$$t_j + s_{i\ell} \leq 1, \ \ \forall i \in [I], \ \forall j, \ell \in P_i \text{ and } j > \ell, \tag{3.6c}$$

$$t_j \leq \sum_{\ell \in P_i} s_{i\ell}, \ \ \forall i \in [I], \ \forall j \in P_i, \tag{3.6d}$$

$$t_j \in \mathbb{B}, \ \ \forall j \in [J], \ \ s_{ij} \in \mathbb{B}, \ \ \forall i \in [I], \ \forall j \in P_i, \tag{3.6e}$$

where constraints (3.6a) describe that, for all $i \in [I]$, $s_{ij} = 1$ holds for at most one $j \in P_i$, constraints (3.6b) designate that if $s_{ij} = 1$ then $t_j = 1$ because of the definition of s_{ij}, constraints (3.6c) describe that, for any two indices $j, \ell \in P_i$ with $j > \ell$, if $s_{i\ell} = 1$ then $t_j = 0$ because ℓ is the largest index such that $t_\ell = 1$, and constraints (3.6d) ensure that $\bar{\alpha}_j = c^{\text{e}}$ for all $j \in P_i$ if all $s_{ij} = 0$. It follows that $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ can be recast as an integer linear program presented in the following theorem. For the ease of exposition, we introduce dependent variables $r_j \equiv 1 - t_j$ and $p_i \equiv 1 - \sum_{j \in P_i} s_{ij}$.

Theorem 9. *For fixed $(w, y, \gamma, \lambda, \rho)$, problem $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ yields the same optimal value as the following integer linear program:*

$$\max_{t,s,r,p} \ \sum_{j=1}^{J} \left(c^{t}_j t_j + c^{r}_j r_j\right) + \sum_{i=1}^{I} \Big(c^{p}_i p_i + \sum_{j \in P_i} c^{s}_i s_{ij}\Big) \tag{3.7a}$$

$$\text{s.t.} \ \ (t, s, r, p) \in \mathcal{H} := \Big\{ \text{(3.6a)–(3.6e)}, \tag{3.7b}$$

$$t_j + r_j = 1, \ \ \forall j \in [J], \tag{3.7c}$$

$$\sum_{j \in P_i} s_{ij} + p_i = 1, \ \ \forall i \in [I] \Big\}, \tag{3.7d}$$

where $c^{t}_j := \left[(-c^{\text{x}} - \gamma_j) w_j\right]_+ + \sup_{\tilde{d}_j \in [d^L_j, d^U_j]_{\mathbb{Z}}} \left\{ c^{\text{x}} \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}^q_j \right\}$, $c^{r}_j := \left[(-c^{\text{e}} - \gamma_j) w_j\right]_+ + \sup_{\tilde{d}_j \in [d^L_j, d^U_j]_{\mathbb{Z}}} \left\{ c^{\text{e}} \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}^q_j \right\}$, $c^{p}_i := \left[(-c^{\text{e}} - \lambda_i) y_i\right]_+$, *and* $c^{s}_i := \left[(-c^{\text{x}} - \lambda_i) y_i\right]_+$.

Second, we linearize the terms $f_j(w_j)\gamma_j$ and $g_i(y_i)\lambda_i$. For all $j \in [J]$, although $f_j(w_j)$ can be nonlinear and non-convex, thanks to the integrality of w_j, we can rewrite $f_j(w_j)$ as an affine function based on a binary expansion of w_j. Specifically, we introduce binary variables $\{u_{jk} : k \in [w_j^{\text{U}} - w_j^{\text{L}}]\}$ such that $w_j = w_j^{\text{L}} + \sum_{k=1}^{w_j^{\text{U}}-w_j^{\text{L}}} u_{jk}$, where we interpret u_{jk} as whether we assign at least $w_j^{\text{L}} + k$ nurses to unit j. That is, $u_{jk} = 1$ if $w_j \geq w_j^{\text{L}} + k$ and $u_{jk} = 0$ otherwise. Then, defining $\Delta_{jk} := f_j(w_j^{\text{L}}+k) - f_j(w_j^{\text{L}}+k-1)$ for all $k \in [w_j^{\text{U}} - w_j^{\text{L}}]$, we have

$$
\begin{aligned}
f_j(w_j) &= f_j(w_j^{\text{L}}) + \sum_{k=1}^{w_j - w_j^{\text{L}}} \left[f_j(w_j^{\text{L}} + k) - f_j(w_j^{\text{L}} + k - 1)\right] \\
&= f_j(w_j^{\text{L}}) + \sum_{k=1}^{w_j^{\text{U}} - w_j^{\text{L}}} \left[f_j(w_j^{\text{L}} + k) - f_j(w_j^{\text{L}} + k - 1)\right] \mathbb{1}_{[w_j^{\text{L}}+k, w_j^{\text{U}}]}(w_j) \\
&= f_j(w_j^{\text{L}}) + \sum_{k=1}^{w_j^{\text{U}} - w_j^{\text{L}}} \Delta_{jk} u_{jk}.
\end{aligned}
$$

It follows that $f_j(w_j)\gamma_j = f_j(w_j^{\text{L}})\gamma_j + \sum_{k=1}^{w_j^{\text{U}}-w_j^{\text{L}}} \Delta_{jk} u_{jk} \gamma_j$. We can linearize the bilinear terms $u_{jk}\gamma_j$ by defining continuous variables $\varphi_{jk} := u_{jk}\gamma_j$ and incorporating the following standard McCormick inequalities (see [134]):

$$
\gamma_j - M(1 - u_{jk}) \leq \varphi_{jk} \leq M u_{jk}, \;\; \forall j \in [J], \; \forall k \in [w_j^{\text{U}} - w_j^{\text{L}}], \tag{3.8a}
$$

$$
-M u_{jk} \leq \varphi_{jk} \leq \gamma_j + M(1 - u_{jk}), \;\; \forall j \in [J], \; \forall k \in [w_j^{\text{U}} - w_j^{\text{L}}], \tag{3.8b}
$$

where M represents a sufficiently large positive constant. Likewise, for all $i \in [I]$, we rewrite $g_i(y_i)\lambda_i$ as $g_i(y_i^{\text{L}})\lambda_i + \sum_{\ell=1}^{y_i^{\text{U}}-y_i^{\text{L}}} \delta_{i\ell} v_{i\ell} \lambda_i$ by using constants $\delta_{i\ell} := g_i(y_i^{\text{L}} + \ell) - g_i(y_i^{\text{L}} + \ell - 1)$ for all $\ell \in [y_i^{\text{U}} - y_i^{\text{L}}]$ and binary variables $\{v_{i\ell} : \ell \in [y_i^{\text{U}} - y_i^{\text{L}}]\}$, where $v_{i\ell} = 1$ if $y_i \geq y_i^{\text{L}} + \ell$ and $v_{i\ell} = 0$ otherwise. We linearize the bilinear terms $v_{i\ell}\lambda_i$ by continuous variables $\nu_{i\ell} := v_{i\ell}\lambda_i$ and the McCormick inequalities

$$
\lambda_i - M(1 - v_{i\ell}) \leq \nu_{i\ell} \leq M v_{i\ell}, \;\; \forall i \in [I], \; \forall \ell \in [y_i^{\text{U}} - y_i^{\text{L}}], \tag{3.8c}
$$

$$
-M v_{i\ell} \leq \nu_{i\ell} \leq \lambda_i + M(1 - v_{i\ell}), \;\; \forall i \in [I], \; \forall \ell \in [y_i^{\text{U}} - y_i^{\text{L}}]. \tag{3.8d}
$$

In computation, a large big-M coefficient M can significantly slow down the solution of (DRNS). Theoretically, for the correctness of the linearization (3.8a)–(3.8d), M needs to be larger than $|\gamma_j|$ and $|\lambda_i|$ for all $j \in [J]$ and $i \in [I]$, respectively. The following proposition derives uniform lower and upper bounds of γ_j and λ_i, leading to a small value of M.

Proposition 7. *For fixed w and y, there exists an optimal solution $(\gamma^*, \lambda^*, \rho^*)$ to formulation* (3.5a)–(3.5b) *such that $\gamma_j^* \in [-c^x, 0]$ for all $j \in [J]$ and $\lambda_i^* \in [-c^x, 0]$ for all $i \in [I]$.*

Proposition 7 indicates that (i) we can set $M := c^x$ in the McCormick inequalities (3.8a)–(3.8d) without loss of optimality and (ii) as all γ_j and λ_i are non-positive at optimality, we can replace McCormick inequalities (3.8a) and (3.8c) as $\gamma_j \leq \varphi_{jk} \leq 0$ and $\lambda_i \leq \nu_{i\ell} \leq 0$ respectively, both of which are now big-M-free. In addition, we incorporate the following constraints to break the symmetry among binary variables:

$$u_{jk} \geq u_{j(k+1)}, \ \ \forall j \in [J], \ \forall k \in [w_j^{\mathrm{U}} - w_j^{\mathrm{L}} - 1], \tag{3.8e}$$

$$v_{i\ell} \geq v_{i(\ell+1)}, \ \ \forall i \in [I], \ \forall \ell \in [y_i^{\mathrm{U}} - y_i^{\mathrm{L}} - 1]. \tag{3.8f}$$

The above analysis recasts (DRNS) into a mixed-integer program, which is summarized in the following theorem without proof.

Theorem 10. *Under Assumption 4, the (DRNS) model* (3.3) *yields the same optimal value as the following mixed-integer program:*

$$\begin{aligned} \min_{\substack{u,v,\varphi,\nu \\ \gamma,\lambda,\rho,\theta}} \ & \theta + \sum_{i=1}^{I} \Big(c^y y_i^L + g_i(y_i^L)\lambda_i + \sum_{\ell=1}^{y_i^U - y_i^L} \big(\delta_{i\ell}\nu_{i\ell} + c^y v_{i\ell}\big)\Big) \\ & + \sum_{j=1}^{J} \Bigg[\sum_{q=1}^{Q} \mu_{jq}\rho_{jq} + c^w w_j^L + f_j(w_j^L)\gamma_j + \sum_{k=1}^{w_j^U - w_j^L} \big(c^w u_{jk} + \Delta_{jk}\varphi_{jk}\big)\Bigg] \end{aligned} \tag{3.9a}$$

$$\text{s.t.} \ \ (3.8a)\text{–}(3.8f), \tag{3.9b}$$

$$\theta \geq \sum_{j=1}^{J} \big(c_j^t t_j + c_j^r r_j\big) + \sum_{i=1}^{I} \big(c_i^p p_i + \sum_{j \in P_i} c_i^s s_i\big), \ \ \forall (t,s,r,p) \in \mathcal{H}, \tag{3.9c}$$

$$u_{jk} \in \mathbb{B}, \ \ \forall j \in [J], \ \forall k \in [w_j^U - w_j^L], \ \ v_{i\ell} \in \mathbb{B}, \ \ \forall i \in [I], \ \forall \ell \in [y_i^U - y_i^L], \tag{3.9d}$$

where set $\mathcal{H}$ is defined in (3.7b)–(3.7d) *and coefficients c_j^t, c_i^s, c_j^r, and c_i^p are represented through*

$$c_j^t = \sup_{\tilde{d}_j \in [d_j^L, d_j^U]_{\mathbb{Z}}} \Big\{ c^x \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\}, \tag{3.9e}$$

$$c_i^s = 0, \tag{3.9f}$$

$$c_j^r = \Big[(-c^e - \gamma_j) w_j^L - \sum_{k=1}^{w_j^U - w_j^L} (\varphi_{jk} + c^e u_{jk}) \Big]_+ + \sup_{\tilde{d}_j \in [d_j^L, d_j^U]_{\mathbb{Z}}} \Big\{ c^e \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\}, \tag{3.9g}$$

$$\textit{and } c_i^p = \Big[(-c^e - \lambda_i)y_i^L - \sum_{\ell=1}^{y_i^U - y_i^L} (\nu_{i\ell} + c^e v_{i\ell})\Big]_+. \tag{3.9h}$$

The reformulation (3.9a)–(3.9d) facilitates the separation algorithm (see, e.g., [135]), also known as delayed constraint generation. We notice that (3.9c) involve 2^J many constraints, making it computationally prohibitive to solve (3.9a)–(3.9d) in one shot. Instead, the separation algorithm incorporates constraints (3.9c) on-the-fly. Specifically, this algorithm first solves a relaxation of the reformulation by overlooking constraints (3.9c). Then, we check if the optimal solution thus obtained violates any of (3.9c). If yes, then we add one violated constraint back into the relaxation and re-solve. We call this added constraint a "cut" and note that each cut describes a convex feasible region. This procedure is repeat until an optimal solution is found to satisfy all of constraints (3.9c). We present the pseudo code in Algorithm 3.1.

1: Initialization: Set the set of cuts $\mathcal{H}_{\text{sep}} = \emptyset$.

2: Solve the master problem

$$(\mathbf{MP}): \min_{\substack{u,v,\varphi,\nu \\ \gamma,\lambda,\rho}} \ \theta + \sum_{i=1}^{I}\Big(c^{\text{y}} y_i^{\text{L}} + g_i(y_i^{\text{L}})\lambda_i + \sum_{\ell=1}^{y_i^{\text{U}} - y_i^{\text{L}}} \big(\delta_{i\ell}\nu_{i\ell} + c^{\text{y}} v_{i\ell}\big)\Big)$$
$$+ \sum_{j=1}^{J}\Big[\sum_{q=1}^{Q} \mu_{jq}\rho_{jq} + c^{\text{w}} w_j^{\text{L}} + f_j(w_j^{\text{L}})\gamma_j + \sum_{k=1}^{w_j^{\text{U}} - w_j^{\text{L}}} \big(c^{\text{w}} u_{jk} + \Delta_{jk}\varphi_{jk}\big)\Big]$$
$$\text{s.t. } (3.9\text{b}), (3.9\text{d}),$$
$$\theta \geq \sum_{j=1}^{J}\big(c_j^{\text{t}} t_j + c_j^{\text{r}} r_j\big) + \sum_{i=1}^{I}\big(c_i^{\text{p}} p_i + \sum_{j \in P_i} c_i^{\text{s}} s_{ij}\big), \ \ \forall (t,s,r,p) \in \mathcal{H}_{\text{sep}},$$

and record an optimal solution $(u^*, v^*, \varphi^*, \nu^*, \gamma^*, \lambda^*, \rho^*, \theta^*)$.

3: Compute $c_j^{\text{t}*}$, $c_i^{\text{s}*}$, $c_j^{\text{r}*}$, and $c_i^{\text{p}*}$ based on (3.9e)–(3.9h) and the values of $(u^*, v^*, \varphi^*, \nu^*, \gamma^*, \lambda^*, \rho^*)$.

4: Solve the integer linear program (3.7a)–(3.7d) using objective coefficients $c_j^{\text{t}*}$, $c_i^{\text{s}*}$, $c_j^{\text{r}*}$, and $c_i^{\text{p}*}$. Record an optimal solution (t^*, s^*, r^*, p^*).

5: **if** $\theta^* \geq \sum_{j=1}^{J}(c_j^{\text{t}*} t_j^* + c_j^{\text{r}*} r_j^*) + \sum_{i=1}^{I}(c_i^{\text{p}*} p_i^* + \sum_{j \in P_i} c_i^{\text{s}*} s_{ij}^*)$ **then**

6: Stop and return (u^*, v^*) as an optimal solution to (DRNS).

7: **else**

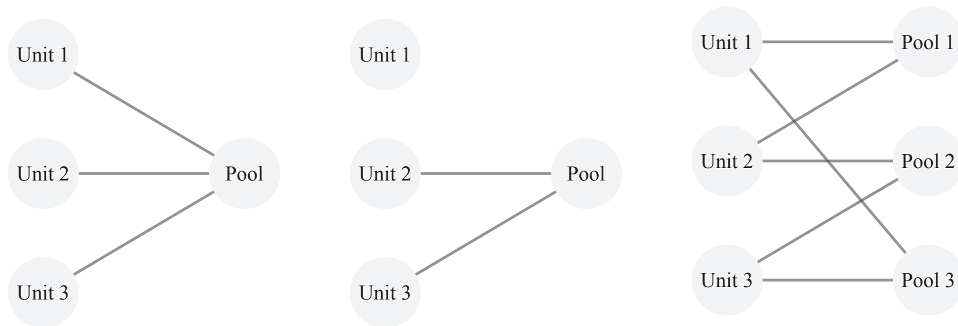
8: Add a cut in the form of (3.9c) into (**MP**) by setting $\mathcal{H}_{\text{sep}} \leftarrow \mathcal{H}_{\text{sep}} \cup \{(t^*, s^*, r^*, p^*)\}$. Go To Step 2.

9: **end if**

Algorithm 3.1: A separation algorithm for solving the (DRNS) model (3.3)

We close this section by confirming the correctness of Algorithm 3.1.

Theorem 11. *Algorithm 3.1 finds a globally optimal solution to the (DRNS) model* (3.3) *in a finite number of iterations.*



(a) Structure [1] (b) Structure [D] (c) Structure [C]

Figure 3.2: Examples of practical pool structures

3.4 Tractable Cases: Practical Pool Structures

In this section, we consider the following three nurse pool structures that often arise in reality.

Structure [1]. (One Pool) $I = 1$*, i.e., there is one single nurse pool shared among all units* (see Figure 3.2a for an example).

Structure [D]. (Disjoint Pools) *All nurse pools are disjoint, i.e., for all* $i_1, i_2 \in [I]$ *and* $i_1 \neq i_2$*, it holds that* $P_{i_1} \cap P_{i_2} = \emptyset$ (see Figure 3.2b for an example).

Structure [C]. (Chained Pools) *The nurse pools form a long chain, i.e., there are* $I = J$ *pools with* $P_i = \{i, i+1\}$ *for all* $i \in [I-1]$ *and* $P_I = \{I, 1\}$ (see Figure 3.2c for an example).

Structure [1] can be utilized when all units have similar functionalities and so they can all share one nurse pool. Accordingly, every nurse assigned to this pool should be cross-trained for all units so that he/she is able to undertake the tasks in them. Structure [D] is less demanding than one pool, as each pool covers only a subset of units which, e.g., have distinct functionalities. Accordingly, the amount of cross-training under this structure significantly decreases from that under one pool. Structure [C] has been applied in the production systems to increase the operational flexibility (see, e.g., [136, 137, 138, 139]). Under this structure, every unit is covered by two nurse pools. Accordingly, every pool nurse needs to be cross-trained for only two units. All three structures have been considered and compared in a nurse staffing context (see, e.g., [116]). Under these practical pool structures, we derive tractable reformulations of the (DRNS) model (3.3). Our derivation leads to monolithic MILP reformulations that facilitate off-the-shelf software like GUROBI.

3.4.1 One pool

We derive a valid inequality to strengthen feasible region $\mathcal{H}$ of the integer program (3.7a)–(3.7d).

Lemma 6. *Under any nurse pool structure, the following inequalities hold valid for all* $(t, s, r, p) \in \mathcal{H}$*:*

$$t_j \leq \sum_{\ell \in P_i : \ell \geq j} s_{i\ell}, \ \ \forall i \in [I], \ \forall j \in P_i. \tag{3.10}$$

Under Structure [1], we show that inequalities (3.10), in conjunction with the existing constraints (3.7b)–(3.7d), are sufficient to describe the convex hull of $\mathcal{H}$. Better still, this yields a closed-form solution to the convex maximization problem $\max_{(\alpha,\beta)} F(\alpha, \beta)$.

Theorem 12. *Under Structure [1], it holds that conv*$(\mathcal{H}) = \overline{\mathcal{H}}$*, where*

$$\overline{\mathcal{H}} \ = \ \big\{(t, s, r, p) \geq 0 : \ \text{(3.6a)–(3.6b), (3.7c)–(3.7d), (3.10)}\big\}.$$

In addition, for fixed $(u, v, \gamma, \lambda, \rho)$*, problem* $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ *yields the same optimal value as the linear program* $\max_{t,s,r,p}\big\{\sum_{j=1}^{J}\big(c^t_j t_j + c^r_j r_j\big) + c^p_1 p_1 : \ (t, s, r, p) \in \overline{\mathcal{H}}\big\}$ *and*

$$\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta) \ = \ \max\left\{c^p_1 + \sum_{\ell=1}^{J} c^r_\ell, \ \max_{j\in[J]}\left\{c^t_j + \sum_{\ell=1}^{j-1} \max\{c^t_\ell, c^r_\ell\} + \sum_{\ell=j+1}^{J} c^r_\ell\right\}\right\},$$

where c^t_j*,* c^r_j*, and* c^p_1 *are computed by* (3.9e)–(3.9h).

Theorem 12 enables us to reduce the 2^J many constraints (3.9c) in the reformulation of (DRNS) to $(J + 1)$ many, thanks to the closed-form solution of $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$. This leads to the following monolithic MILP reformulation of (DRNS).

Proposition 8. *Under Assumption 4 and Structure [1], the (DRNS) model* (3.3) *yields the same optimal objective value as the following MILP:*

$$Z^\star_{[1]} := \min \ \theta + c^y y^L_1 + g_1(y^L_1)\lambda_1 + \sum_{\ell=1}^{y^U_1 - y^L_1} \big(\delta_{1\ell}\nu_{1\ell} + c^y v_{1\ell}\big)$$

$$+\sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\mu_{jq}\rho_{jq}+c^{w}w_j^{L}+f_j(w_j^{L})\gamma_j+\sum_{k=1}^{w_j^{U}-w_j^{L}}\left(c^{w}u_{jk}+\Delta_{jk}\varphi_{jk}\right)\right]$$

$$\textit{s.t.}\;\; \text{(3.8a)–(3.8f)},\; \text{(3.9d)},$$

$$\theta \geq \phi_1+\sum_{\ell=1}^{J}(\zeta_\ell+\eta_\ell^{e}),\;\; \theta\geq \eta_j^{x}+\sum_{\ell}^{j-1}\chi_\ell+\sum_{\ell=j+1}^{J}(\zeta_\ell+\eta_\ell^{e}),\;\; \forall j\in[J],$$

$$\left.\begin{array}{l} \chi_j\geq \zeta_j+\eta_j^{e},\;\; \chi_j\geq\eta_j^{x} \\ \zeta_j\geq(-c^{e}-\gamma_j)w_j^{L}-\sum_{k=1}^{w_j^{U}-w_j^{L}}(\varphi_{jk}+c^{e}u_{jk}),\;\; \zeta_j\geq 0 \\ \eta_j^{x}\geq c^{x}\tilde{d}_j-\sum_{q=1}^{Q}\tilde{d}_j^{q}\rho_{jq},\;\; \eta_j^{e}\geq c^{e}\tilde{d}_j-\sum_{q=1}^{Q}\tilde{d}_j^{q}\rho_{jq},\;\; \forall\tilde{d}_j\in[d_j^{L},d_j^{U}]_{\mathbb{Z}} \end{array}\right\}\;\forall j\in[J], \tag{3.11}$$

$$\phi_1\geq(-c^{e}-\lambda_1)y_1^{L}-\sum_{\ell=1}^{y_1^{U}-y_1^{L}}(\nu_{1\ell}+c^{e}v_{1\ell}),\;\; \phi_1\geq 0.$$

A special case of Structure [1] is when there are no nurse float pools. Mathematically, this is equivalent to assigning all units to one single pool with no pool nurses. We hence call it Structure [0] as there is zero pool nurse. Under this structure, $y_1^{\mathrm{L}}=y_1^{\mathrm{U}}=0$ and accordingly $g_1(y_1^{\mathrm{L}})=0$. A MILP reformulation of (DRNS) under Structure [0] follows from Proposition 8:

$$Z_{[0]}^{\star}\;:=\;\min\;\theta+\sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\mu_{jq}\rho_{jq}+c^{\mathrm{w}}w_j^{\mathrm{L}}+f_j(w_j^{\mathrm{L}})\gamma_j+\sum_{k=1}^{w_j^{\mathrm{U}}-w_j^{\mathrm{L}}}\left(c^{\mathrm{w}}u_{jk}+\Delta_{jk}\varphi_{jk}\right)\right]$$

$$\text{s.t.}\;\; \text{(3.8a)–(3.8b)},\; \text{(3.8e)},\; \text{(3.9d)},\; \text{(3.11)},$$

$$\theta\geq\sum_{\ell=1}^{J}(\zeta_\ell+\eta_\ell^{\mathrm{e}}),\;\; \theta\geq\eta_j^{\mathrm{x}}+\sum_{\ell}^{j-1}\chi_\ell+\sum_{\ell=j+1}^{J}(\zeta_\ell+\eta_\ell^{\mathrm{e}}),\;\; \forall j\in[J].$$

We notice that, whenever $y_1^{\mathrm{L}}=0$, any feasible nurse staffing levels under Structure [0] are also feasible to (DRNS) under Structure [1]. It then follows that $Z_{[1]}^{\star}\leq Z_{[0]}^{\star}$. In addition, as Structure [1] provides the most operational flexibility and Structure [0] has zero flexibility, we may interpret the difference $Z_{[0]}^{\star}-Z_{[1]}^{\star}$ as the (maximum) value of operational flexibility.

3.4.2 Disjoint pools

Under Structure [D], we can once again obtain the convex hull of $\mathcal{H}$ by incorporating inequalities (3.10). Intuitively, as the nurse pools are disjoint, $\mathcal{H}$ becomes separable in index i, i.e., separable among the nurse pools and the units under each pool. Hence, conv($\mathcal{H}$) can be obtained by convexifying the projection of $\mathcal{H}$ in each pool and then taking their Cartesian product. It follows that, once again, the convex maximization problem $\max_{(\alpha,\beta)} F(\alpha,\beta)$ admits a closed-form solution and (DRNS) can be recast as a monolithic MILP. In particular, we reduce the exponentially many constraints (3.9c) in the reformulation of (DRNS) to $(I+J)$ many. We summarize these results in the following proposition.

Proposition 9. *Under Structure [D], it holds that conv($\mathcal{H}$) $=$ $\overline{\mathcal{H}}$. In addition, for fixed $(u,v,\gamma,\lambda,\rho)$, it holds that*

$$\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta) \;=\; \sum_{i\in[I]}\max\left\{c^{p}_i+\sum_{\ell\in P_i}c^{r}_{\ell},\ \max_{j\in P_i}\left\{c^{l}_j+\sum_{\ell\in P_i:\ell<j}\max\{c^{l}_{\ell},c^{r}_{\ell}\}+\sum_{\ell\in P_i:\ell>j}c^{r}_{\ell}\right\}\right\},$$

where c^{l}_j, c^{r}_j, and c^{p}_i are computed by (3.9e)–(3.9h). *Furthermore, under Assumption 4, the (DRNS) model* (3.3) *yields the same optimal objective value as the following MILP:*

$$\begin{aligned}
Z^{\star}_{[D]} \;:=\; \min\;\; & \sum_{i=1}^{I}\Big(\theta_i+c^{y}y^{L}_i+g_i(y^{L}_i)\lambda_i+\sum_{\ell=1}^{y^{U}_i-y^{L}_i}\big(\delta_{i\ell}\nu_{i\ell}+c^{y}v_{i\ell}\big)\Big)\\
&+\sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\mu_{jq}\rho_{jq}+c^{w}w^{L}_j+f_j(w^{L}_j)\gamma_j+\sum_{k=1}^{w^{U}_j-w^{L}_j}\big(c^{w}u_{jk}+\Delta_{jk}\varphi_{jk}\big)\right]\\
\textit{s.t.}\;\; & \text{(3.8a)–(3.8f)},\ \text{(3.9d)},\ \text{(3.11)},\ \theta_i\ge\phi_i+\sum_{\ell\in P_i}(\zeta_\ell+\eta^{e}_{\ell}),\ \ \forall i\in[I],\\
& \theta_i\ge\eta^{x}_j+\sum_{\ell\in P_i:\ell<j}\chi_\ell+\sum_{\ell\in P_i:\ell>j}(\zeta_\ell+\eta^{e}_{\ell}),\ \ \forall i\in[I],\ \forall j\in P_i,\\
& \phi_i\ge(-c^{e}-\lambda_i)y^{L}_i-\sum_{\ell=1}^{y^{U}_i-y^{L}_i}(\nu_{i\ell}+c^{e}v_{i\ell}),\ \ \forall i\in[I], && \text{(3.12a)}\\
& \phi_i\ge 0,\ \ \forall i\in[I],\ \ \zeta_j\ge 0,\ \ \forall j\in[J]. && \text{(3.12b)}
\end{aligned}$$

3.4.3 Chained pools

Under Structure [C], the valid inequalities (3.10) can still be incorporated to strengthen and simplify the mixed-integer set $\mathcal{H}$. Specifically, as $P_i=\{i,i+1\}$ for all $i\in[I-1]$, inequalities (3.10) imply that $t_{i+1}\le s_{i(i+1)}$. But constraints (3.6b) designate $s_{i(i+1)}\le t_{i+1}$, implying that $t_{i+1}=s_{i(i+1)}$. Similarly, we obtain $s_{I1}=t_1$ and simplify $\mathcal{H}$ as follows:

$$\begin{aligned}
\mathcal{H} \;=\; \Big\{(t,s,r,p)\in\mathbb{B}^{4I}:\ & s_{ii}\le t_i\le s_{ii}+t_{\sigma(i)}\le 1,\ \ \forall i\in[I], && \text{(3.13)}\\
& t_i+r_i=1,\ \ \forall i\in[I],
\end{aligned}$$

$$p_i + s_{ii} + t_{\sigma(i)} = 1, \ \forall i \in [I]\Big\},$$

where $\sigma(i) := i+1$ for all $i \in [I-1]$ and $\sigma(I) := 1$. Unfortunately, unlike under Structures [1] and [D], the strengthened $\mathcal{H}$ is no longer integral, i.e., conv$(\mathcal{H}) \neq \overline{\mathcal{H}}$. We demonstrate this fact in the following example.

Example 1. *Consider an example of 3 chained pools, i.e., $I = J = 3$, $P_1 = \{1, 2\}$, $P_2 = \{2, 3\}$, and $P_3 = \{3, 1\}$. Incorporating valid inequalities* (3.10) *and relaxing integrality restrictions in $\mathcal{H}$ yields*

$$\overline{\mathcal{H}} = \Big\{(t, s, r, p) \geq 0 : s_{11} \leq t_1 \leq s_{11} + t_2 \leq 1, \tag{3.14a}$$

$$s_{22} \leq t_2 \leq s_{22} + t_3 \leq 1, \tag{3.14b}$$

$$s_{33} \leq t_3 \leq s_{33} + t_1 \leq 1, \tag{3.14c}$$

$$t_1 + r_1 = t_2 + r_2 = t_3 + r_3 = 1,$$

$$p_1 + s_{11} + t_2 = p_2 + s_{22} + t_3 = p_3 + s_{33} + t_1 = 1\Big\}.$$

We observe that polyhedron $\overline{\mathcal{H}}$ is 6-dimensional. Hence, replacing the first and last inequalities in constraints (3.14a)–(3.14c) *with equalities yields the following extreme point:*

$$(t_1, t_2, t_3, s_{11}, s_{22}, s_{33}, r_1, r_2, r_3, p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0, 0\right),$$

which is fractional. Therefore, $\overline{\mathcal{H}}$ is not integral and $\overline{\mathcal{H}} \neq$ conv$(\mathcal{H})$.

Despite the loss of integrality, we adopt an alternative approach to recast the integer program (3.7a)–(3.7d), and hence the convex maximization problem $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$, as a linear program. We start by noticing that inequalities (3.13) allow us to represent variables s_{ii} as $s_{ii} = t_i(1 - t_{\sigma(i)})$ for all $i \in [I]$. In fact, (3.13) are exactly the McCormick inequalities that linearize this (nonlinear) representation. It follows that $p_i = 1 - s_{ii} - t_{\sigma(i)} = (1 - t_i)(1 - t_{\sigma(i)})$ for all $i \in [I]$. Plugging these representations into formulation (3.7a)–(3.7d) yields a reformulation of $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ based on variables t only:

$$\begin{aligned}\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta) &= \max_{t\in\mathbb{B}^I} \sum_{i=1}^{I} \Big[c_i^{\mathrm{t}} t_i + c_i^{\mathrm{r}}(1 - t_i) + c_i^{\mathrm{p}}(1 - t_i)(1 - t_{\sigma(i)})\Big] \\ &= \max_{t\in\mathbb{B}^I} c_1^{\mathrm{t}} t_1 + c_1^{\mathrm{r}}(1 - t_1) \\ &\quad + \sum_{i=2}^{I} \Big[c_i^{\mathrm{t}} + (c_i^{\mathrm{r}} - c_i^{\mathrm{t}})(1 - t_i) + c_{i-1}^{\mathrm{p}}(1 - t_{i-1})(1 - t_i)\Big]\end{aligned}$$

$$+ c_I^{\mathrm{p}}(1 - t_I)(1 - t_1). \tag{3.15}$$

The reformulation (3.15) decomposes objective function based on index $i \in [I]$ and enables us to solve $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ by a dynamic program (DP), i.e., we sequentially optimize $t_1, t_2, \ldots, t_I$. To this end, we define the state of the DP in stage 1 as $\widehat{t}_1 \in \mathbb{B}$ and the states in stage i as $(\widehat{t}_1, \widehat{t}_i) \in \mathbb{B}^2$ for all $i \in [2, I]_{\mathbb{Z}}$. In addition, we formulate the DP as $\max_{(\widehat{t}_1,\widehat{t}_I)\in\mathbb{B}^2}\{V_I(\widehat{t}_1, \widehat{t}_I) + c_I^{\mathrm{p}}(1-\widehat{t}_I)(1-\widehat{t}_1)\}$, where the value functions $V_i(\cdot)$ are recursively defined through

$$\begin{aligned} V_1(\widehat{t}_1) &= c_1^{\mathrm{t}}\widehat{t}_1 + c_1^{\mathrm{r}}(1 - \widehat{t}_1), \\ \text{and } V_i(\widehat{t}_1, \widehat{t}_i) &= \max_{\widehat{t}_{i-1}\in\mathbb{B}} \Big\{ V_{i-1}(\widehat{t}_1, \widehat{t}_{i-1}) + c_i^{\mathrm{t}} + (c_i^{\mathrm{r}} - c_i^{\mathrm{t}})(1 - \widehat{t}_i) \\ &\qquad + c_{i-1}^{\mathrm{p}}(1 - \widehat{t}_{i-1})(1 - \widehat{t}_i) \Big\}, \ \forall i \in [2, I]_{\mathbb{Z}}, \ \forall (\widehat{t}_1, \widehat{t}_i) \in \mathbb{B}^2. \end{aligned}$$

For all $i \in [I]$, value function $V_i(\widehat{t}_1, \widehat{t}_i)$ represents the "cumulative reward" up to stage i, i.e., the terms in (3.15) that involve $t_1, \ldots, t_i$ only. We note that, as $\widehat{t}_1$ is involved in the final-stage reward, the DP stores the value of $\widehat{t}_1$ in the state throughout stages $2, \ldots, I$.

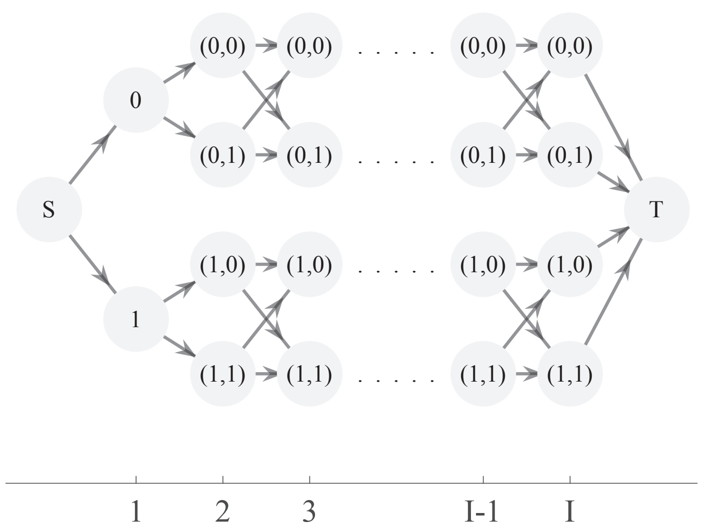


Figure 3.3: Longest-path problem on an acyclic direct network

We further interpret the DP as a longest-path problem on an acyclic directed network $(\mathcal{N}, \mathcal{A})$. Specifically, the set of nodes $\mathcal{N}$ consists of I layers, denoted by $\{\mathcal{N}_i\}_{i=1}^N$. For all $i \in [I]$, layer i consists of the states of the DP in stage i, i.e., $\mathcal{N}_1 = \{0, 1\}$, and $\mathcal{N}_i = \{(0,0), (0,1), (1,0), (1,1)\}$ for all $i \in [2, I]_{\mathbb{Z}}$. In addition, $\mathcal{A}$ consists of arcs that connect two nodes in neighboring layers, as long as the two nodes share a common $\widehat{t}_1$ value, i.e., $\mathcal{A} = \{[\widehat{t}_1, (\widehat{t}_1, \widehat{t}_2)] : \widehat{t}_1, \widehat{t}_2 \in \mathbb{B}\} \cup \{[(\widehat{t}_1, \widehat{t}_{i-1}), (\widehat{t}_1, \widehat{t}_i)] : \widehat{t}_{i-1}, \widehat{t}_i \in \mathbb{B}, \forall i \in [3, I]_{\mathbb{Z}}\}$.

Finally, we incorporate into $\mathcal{N}$ a starting node $\texttt{S}$ and a terminal node $\texttt{T}$, and into $\mathcal{A}$ arcs from $\texttt{S}$ to all nodes in $\mathcal{N}_1$ and from all nodes in $\mathcal{N}_I$ to $\texttt{T}$. We depict $(\mathcal{N}, \mathcal{A})$ in Figure 3.3. Then, the DP is equivalent to the longest-path problem from $\texttt{S}$ to $\texttt{T}$ on $(\mathcal{N}, \mathcal{A})$. We formally state this result in the following theorem.

Theorem 13. *Define* $\{c_{[m,n]} : [m,n] \in \mathcal{A}\}$*, the length of the arcs in network* $(\mathcal{N}, \mathcal{A})$*, such that*

$$\begin{aligned} c_{[\texttt{S},\widehat{t}_1]} &= c^t_1\widehat{t}_1 + c^r_1(1-\widehat{t}_1), \ \ \forall \widehat{t}_1 \in \mathbb{B}, \\ c_{[(\widehat{t}_1,\widehat{t}_{i-1}),(\widehat{t}_1,\widehat{t}_i)]} &= c^t_i + (c^r_i - c^t_i)(1-\widehat{t}_i) + c^p_{i-1}(1-\widehat{t}_{i-1})(1-\widehat{t}_i), \\ &\quad \forall \widehat{t}_{i-1}, \widehat{t}_i \in \mathbb{B}, \ \forall i \in [2, I]_{\mathbb{Z}}, \\ \textit{and} \ \ c_{[(\widehat{t}_1,\widehat{t}_I),\texttt{T}]} &= c^p_I(1-\widehat{t}_I)(1-\widehat{t}_1), \ \ \forall \widehat{t}_1, \widehat{t}_I \in \mathbb{B}. \end{aligned}$$

Then, for fixed $(u, v, \gamma, \lambda, \rho)$*,* $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ *equals the length of the longest* $\texttt{S}$*-*$\texttt{T}$ *path on* $(\mathcal{N}, \mathcal{A})$*, that is,*

$$\begin{aligned} \max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta) \ = \ \max_{x\in[0,1]^{\mathcal{A}}} & \sum_{[m,n]\in\mathcal{A}} c_{[m,n]}x_{[m,n]} \\ \textit{s.t.} & \sum_{n:[m,n]\in\mathcal{A}} x_{[m,n]} - \sum_{n:[n,m]\in\mathcal{A}} x_{[n,m]} = \begin{cases} 1, & \textit{if } m = \texttt{S} \\ 0, & \textit{if } m \neq \texttt{S}, \ \texttt{T} \\ -1, & \textit{if } m = \texttt{T}, \end{cases} \quad \forall m \in \mathcal{N}. \end{aligned}$$

We note that $(\mathcal{N}, \mathcal{A})$ is acyclic and it consists of $4I$ nodes and $8I - 6$ arcs. Hence, the longest-path problem, as well as $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$, can be solved in time polynomial of the problem input. Accordingly, we are able to replace the exponentially many constraints (3.9c) in the reformulation of (DRNS) with $\mathcal{O}(I)$ many linear constraints. This yields the following monolithic MILP reformulation.

Proposition 10. *Under Structure [C] and Assumption 4, the (DRNS) model* (3.3) *yields the same optimal objective value as the following MILP:*

$$\begin{aligned} \min \ \theta &+ \sum_{i=1}^{I}\Big(c^y y^L_i + g_i(y^L_i)\lambda_i + \sum_{\ell=1}^{y^U_i - y^L_i}\big(\delta_{i\ell}\nu_{i\ell} + c^y v_{i\ell}\big)\Big) \\ &+ \sum_{j=1}^{I}\left[\sum_{q=1}^{Q}\mu_{jq}\rho_{jq} + c^w w^L_j + f_j(w^L_j)\gamma_j + \sum_{k=1}^{w^U_j - w^L_j}\big(c^w u_{jk} + \Delta_{jk}\varphi_{jk}\big)\right] \\ \textit{s.t.} \ \ & \text{(3.8a)–(3.8f)}, \ \text{(3.9d)}, \ \text{(3.11)}, \ \text{(3.12a)–(3.12b)}, \\ & \theta \geq \pi_{\texttt{S}} - \pi_{\texttt{T}}, \end{aligned}$$

$$
\begin{aligned}
&\pi_{\mathrm{S}} - \pi_{\widehat{t}_1} \geq \widehat{t}_1 \eta_1^x + (1-\widehat{t}_1)(\zeta_1 + \eta_1^e), \;\; \forall \widehat{t}_1 \in \mathbb{B}, \\
&\pi_{(\widehat{t}_1, \widehat{t}_{i-1})} - \pi_{(\widehat{t}_1, \widehat{t}_i)} \geq \eta_i^x + (1-\widehat{t}_i)(\zeta_i + \eta_i^e - \eta_i^x) \\
&\qquad\qquad\qquad\qquad + (1-\widehat{t}_{i-1})(1-\widehat{t}_i)\phi_{i-1}, \;\; \forall \widehat{t}_{i-1}, \widehat{t}_i \in \mathbb{B}, \; \forall i \in [2, I]_{\mathbb{Z}}, \\
&\pi_{(\widehat{t}_1, \widehat{t}_I)} - \pi_{\mathrm{T}} \geq (1-\widehat{t}_I)(1-\widehat{t}_1)\phi_I, \;\; \forall \widehat{t}_1, \widehat{t}_I \in \mathbb{B}.
\end{aligned}
$$

3.5 Optimal Nurse Pool Design

Of all the three practical nurse pool structures, Structure [1] is most flexible as every pool nurse is capable of working in all units. However, this incurs a high need for cross-training. For example, to enable a nurse working in a unit to be a pool nurse, he/she needs to be cross-trained for all the remaining $J-1$ units. As a result, enabling all nurses needs as many as $J(J-1)/2$ pairs of cross-training. In contrast, Structure [C] needs J pairs of cross-training because every pool consists of exactly two units. Structure [D] needs even less cross-training if we adopt a "sparse" design, e.g., pooling together a small subset of units. In this section, we examine how to design a sparse but effective pool structure that is disjoint. Specifically, we search for a disjoint pool structure that needs as few cross-training as possible, while achieving a pre-specified performance guarantee in terms of DR staffing cost.[2] To this end, we define binary variables a_{ij} such that $a_{ij} = 1$ if unit j is assigned to pool i and $a_{ij} = 0$ otherwise, binary variables o_i such that $o_i = 1$ if any units are assigned to pool i (i.e., if pool i is "opened") and $o_i = 0$ otherwise, and binary variables p_{jk} such that $p_{jk} = 1$ if units j and k are assigned to the same pool and $p_{jk} = 0$ otherwise. Then, the total amount of needed cross-training equals $\sum_{j=1}^{J} \sum_{k=j+1}^{J} p_{jk}$. In addition, these binary variables satisfy the following constraints:

$$\sum_{i=1}^{I+1} a_{ij} = 1, \;\; \forall j \in [J], \tag{3.16a}$$

$$a_{ij} \leq o_i, \;\; \forall i \in [I], \; \forall j \in [J], \tag{3.16b}$$

$$p_{jk} \geq a_{ij} + a_{ik} - 1, \;\; \forall i \in [I], \; \forall j, k \in [J] \text{ and } j < k, \tag{3.16c}$$

where constraints (3.16a) designate that each unit is assigned to one and only one pool (we create a dummy pool $I+1$ that collects all units that are not covered by any existing pools), constraints (3.16b) ensure that no units can be assigned to a pool if it is not opened, and constraints (3.16c) designate that $p_{jk} = 1$ if there is a pool i such that $a_{ij} = a_{ik} = 1$. If

[2] We notice that there exist multiple alternative quantities that can be used to quantify the effort of cross-training. In this chapter, we pick the number of pairs of cross-training as a representative objective function. Alternative objectives can be similarly modeled and computed.

no such a pool i exists, then constraints (3.16b) reduce to $p_{jk} \geq 0$ and p_{jk} equals zero at optimality due to the objective function (3.17a). Based on Proposition 9, the optimal nurse pool design (OPD) model is formulated as

$$(\textbf{OPD}): \quad \min \sum_{j=1}^{J} \sum_{k=j+1}^{J} p_{jk} \tag{3.17a}$$

$$\text{s.t.} \quad (3.8\text{a})\text{–}(3.8\text{f}),\ (3.9\text{d}),\ (3.11),\ (3.12\text{a})\text{–}(3.12\text{b}),\ (3.16\text{a})\text{–}(3.16\text{c}), \tag{3.17b}$$

$$\sum_{i=1}^{I} \Big(\theta_i + c^{\text{y}} y_i^{\text{L}} + g_i(y_i^{\text{L}})\lambda_i + \sum_{\ell=1}^{y_i^{\text{U}}-y_i^{\text{L}}} \big(\delta_{i\ell}\nu_{i\ell} + c^{\text{y}} v_{i\ell} \big) \Big) o_i$$
$$+ \sum_{j=1}^{J} \left[\sum_{q=1}^{Q} \mu_{jq}\rho_{jq} + c^{\text{w}} w_j^{\text{L}} + f_j(w_j^{\text{L}})\gamma_j + \sum_{k=1}^{w_j^{\text{U}}-w_j^{\text{L}}} \big(c^{\text{w}} u_{jk} + \Delta_{jk}\varphi_{jk} \big) \right] \leq T, \tag{3.17c}$$

$$\theta_i \geq \phi_i + \sum_{\ell=1}^{J} (\zeta_\ell + \eta_\ell^{\text{e}}) a_{i\ell}, \quad \forall i \in [I], \tag{3.17d}$$

$$\theta_i \geq \eta_j^{\text{x}} a_{ij} + \sum_{\ell=1}^{j-1} \chi_\ell a_{i\ell} + \sum_{\ell=j+1}^{J} (\zeta_\ell + \eta_\ell^{\text{e}}) a_{i\ell}, \quad \forall i \in [I],\ \forall j \in [J], \tag{3.17e}$$

where constraint (3.17c) ensures that the DR staffing cost does not exceed a given target T. If $y_i^{\text{L}} = 0$ for all $i \in [I]$, i.e., if there is no minimum staffing requirement for pool nurses, then we shall pick T from the interval $[Z^\star_{[1]}, Z^\star_{[0]}]$, where $Z^\star_{[1]}$ represents the DR staffing cost with maximum flexibility and $Z^\star_{[0]}$ represents that with minimum flexibility. By gradually decreasing this target from $Z^\star_{[0]}$ to $Z^\star_{[1]}$, the amount of cross-training grows and accordingly we obtain a cost-training frontier that can clearly illustrate the trade-off between these two performance measures (see Section 3.6.4 for the numerical demonstration).

To effectively solve the (OPD) model, we recast it as a MILP in the following proposition.

Proposition 11. *Under Assumption 4, the (OPD) model* (3.17) *yields the same optimal objective value and the same set of optimal solutions as the following MILP:*

$$\min \sum_{j=1}^{J} \sum_{k=j+1}^{J} p_{jk}$$

$$\textit{s.t.} \quad (3.8\text{a})\text{–}(3.8\text{f}),\ (3.9\text{d}),\ (3.11),\ (3.12\text{b}),\ (3.16\text{a})\text{–}(3.16\text{c}),$$

$$\sum_{i=1}^{I} \Big(\theta_i + c^{\text{y}} y_i^{L} o_i + g_i(y_i^{L})\lambda_i + \sum_{\ell=1}^{y_i^{U}-y_i^{L}} \big(\delta_{i\ell}\nu_{i\ell} + c^{\text{y}} v_{i\ell} \big) \Big)$$

$$
+\sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\mu_{jq}\rho_{jq}+c^{w}w_j^{L}+f_j(w_j^{L})\gamma_j+\sum_{k=1}^{w_j^U-w_j^L}\left(c^{w}u_{jk}+\Delta_{jk}\varphi_{jk}\right)\right] \le T,
$$

$$
\left.\begin{array}{l}
\theta_i \ge \eta^{x}_{ij}+\sum_{\ell=1}^{j-1}\chi_{i\ell}+\sum_{\ell=j+1}^{J}(\zeta_{i\ell}+\eta^{e}_{i\ell})\\
\chi_{ij}\ge \zeta_{ij}+\eta^{e}_{ij},\ \ \chi_{ij}\ge \eta^{x}_{ij}\\
0\le \zeta_{ij}\le Ka_{ij},\ \ -Ka_{ij}\le \eta^{x}_{ij}\le Ka_{ij},\ \ -Ka_{ij}\le \eta^{e}_{ij}\le Ka_{ij}
\end{array}\right\}\ \forall i\in[I+1],\ \forall j\in[J],
$$

$$
\left.\begin{array}{l}
\theta_i\ge \phi_i+\sum_{\ell=1}^{J}(\zeta_{i\ell}+\eta^{e}_{i\ell})\\
\phi_i\ge -c^{e}y_i^{L}o_i-y_i^{L}\lambda_i-\sum_{\ell=1}^{y_i^U-y_i^L}(\nu_{i\ell}+c^{e}v_{i\ell})\\
-c^{x}o_i\le \lambda_i\le 0,\ \ v_{i1}\le o_i
\end{array}\right\}\ \forall i\in[I+1],
$$

$$
\zeta_j=\sum_{i=1}^{I+1}\zeta_{ij},\ \ \eta^{x}_j=\sum_{i=1}^{I+1}\eta^{x}_{ij},\ \ \eta^{e}_j=\sum_{i=1}^{I+1}\eta^{e}_{ij},\ \ \forall j\in[J],
$$

where K represents a sufficiently large positive constant.

In addition, the above formulation involves symmetric binary solutions. In Appendix B.15, we derive symmetry breaking inequalities to enhance its computational efficacy.

3.6 Numerical Experiments

In this section, we report numerical experiments on (DRNS) and (OPD) models. We summarize our main findings as follows:

1. Under the practical nurse pool structures as introduced in Section 3.4, the monolithic MILP reformulations of (DRNS) lead to significant speed-up over the separation algorithm.

2. Modeling nurse absenteeism improves the out-of-sample performance of staffing decisions. The improvement becomes more significant as the value of operational flexibility increases.

3. Even a very sparse nurse pool design can harvest most of the operational flexibility.

4. An optimal nurse pool design tends to pool together the units with higher variability, e.g., higher standard deviation of nurse demand and/or higher absence rate. In particular, the variability of nurse absenteeism plays a more important role in optimal pool design.

In all experiments, we solve optimization models by GUROBI 7.0.1 via Python 2.7 on a personal laptop with an Intel(R) Core(TM) i7-4850HQ CPU@2.3GHz and 16GB RAM.

3.6.1 Instance design

We design test instances based on the data and insights provided by our collaborating hospital and existing literature [113]. Specifically, we set $Q = 2$ for nurse demand uncertainty. That is, we consider the nurse demand mean value μ_{j1}, which is randomly extracted from the interval $[5, 20]$, and the standard deviation sd_j, which is randomly extracted from the interval $[0, 20]$. In addition, we assume a constant nurse absence rate such that $f_j(w_j) = A_j^{\text{u}} w_j$ and $g_i(y_i) = A_i^{\text{p}} y_i$, where A_j^{u} denotes absence rate of unit nurses and is randomly extracted from the interval $[0.60, 0.98]$ and A_i^{p} denotes that of pool nurses and is randomly extracted from the interval $[0.98, 1.00]$. For (DRNS), we set $w_j^{\text{L}} = \lfloor \underline{S} A_j^{\text{u}} \mu_{j1} \rfloor$, $w_j^{\text{U}} = 200$ for all $j \in [J]$ and $y_i^{\text{L}} = 0$, $y_i^{\text{U}} = 200$ for all $i \in [I]$, where $\underline{S}$ represents a safety constant. In practice, a positive w_j^{L} helps to maintain a constant roster in each unit to promote teamwork. We also incorporate an integrative staffing upper bound by specifying that

$$R := \left\{ (w, y) : \sum_{j=1}^{J} w_j + \sum_{i=1}^{I} y_i \leq \left\lceil \overline{S} \sum_{j=1}^{J} \left(\frac{\sum_{j=1}^{J} A_j^{\text{u}} + \sum_{i=1}^{I} A_i^{\text{p}}}{I + J} \right) \mu_{j1} \right\rceil \right\},$$

where $\overline{S}$ represents another safety constant that describes an upper limit on the human resource.

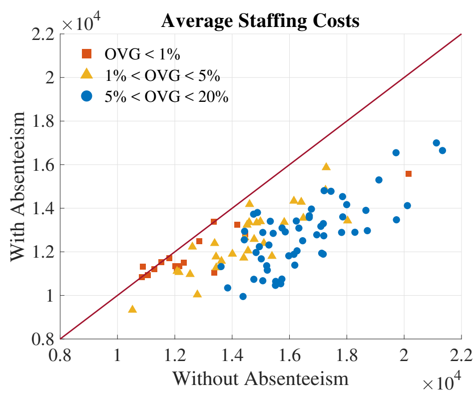
Table 3.1: Average wall-clock seconds used to solve (DRNS)

$[I, J]$	Separation	MILP	$[I, J]$	Separation	MILP
$[1, 5]$	2.28	0.09	$[3, 5]$	1.93	0.09
$[1, 7]$	8.24	0.10	$[3, 7]$	9.82	0.10
$[1, 10]$	44.42	0.13	$[3, 10]$	68.33	0.16
$[1, 20]$	> 3600	0.37	$[3, 20]$	> 3600	0.37
$[1, 50]$	> 3600	1.51	$[3, 50]$	> 3600	1.13

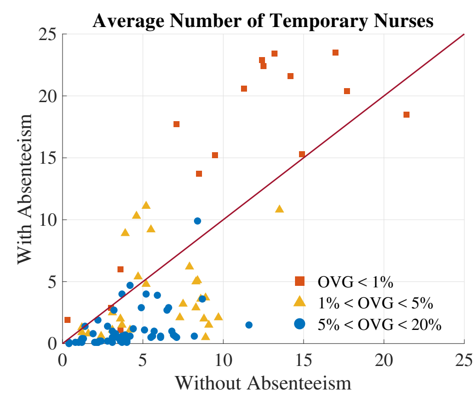
3.6.2 Computational performance

We compare the computational efficacy of the separation algorithm and the monolithic MILP reformulations on solving (DRNS) under practical pool structures. Specifically, we create 10 random test instances with various $[I, J]$ combinations, where $I = 1$ indicates one single pool (i.e., Structure [1]) and $I = 3$ indicates three disjoint pools (e.g., Structure [D]). We report the computing time (in wall-clock seconds) in Table 3.1. From this table, we observe that the time spent by the separation algorithm quickly increases and hits the 1-hour time limit as J increases. In contrast, the MILP reformulations are significantly more scalable and can be solved to global optimality within 2 seconds in all instances.

3.6.3 Value of modeling nurse absenteeism



(a) Average staffing costs



(b) Average number of temporary nurses

Figure 3.4: An out-of-sample comparison of considering versus overlooking nurse absenteeism

As discussed in Section **??**, modeling nurse absenteeism incurs endogenous uncertainty and computational challenges. It is hence worth examining what (DRNS) buys us, i.e., the value of modeling nurse absenteeism. To this end, we consider a test instance with 7 units and one single pool (i.e., under Structure [1]). In addition, we consider a variant of (DRNS) that overlooks the nurse absenteeism, in which we assume that all assigned nurses show up. Then, we compare the out-of-sample performance of the optimal nurse staffing decisions produced by (DRNS) and that produced by overlooking absenteeism. Fixing the nurse staffing levels as in a (DRNS) optimal solution (w^*, y^*), we generate a large number of scenarios for nurse demand and absenteeism, where the demands follow log-normal distribution, i.e., $\tilde{d}_j \sim \text{LN}(\mu_{j1}, \text{sd}_j)$, and the numbers of present nurses follow binomial distribution, i.e., $\tilde{w}_j \sim B(w^*_j, A^{\text{u}}_j)$ and $\tilde{y}_i \sim B(y^*_i, A^{\text{p}}_i)$. Exposing (w^*, y^*) under these scenarios produces an out-of-sample estimate of the average staffing cost with absenteeism, which we denote by Z_{abs}. Using the same set of scenarios, we examine the optimal solution produced by overlooking absenteeism and obtain an out-of-sample average cost without absenteeism, denoted by $Z_{\text{w/o}}$. Using the same out-of-sample procedures, we compute the average number of temporary nurses hired when considering absenteeism (denoted by x_{abs}) and when overlooking it (denoted by $x_{\text{w/o}}$).

We depict the values of $Z_{\text{w/o}}$ (x-coordinate) and Z_{abs} (y-coordinate) obtained in 100 replications in Figure 3.4a. From this figure, we observe that most dots are below the 45-degree line, indicating that $Z_{\text{w/o}} - Z_{\text{abs}} > 0$, i.e., modeling nurse absenteeism yields nurse staffing levels with better out-of-sample performance. In addition, we group the dots based on the relative value of operational flexibility OVG $:= (Z^\star_{[0]} - Z^\star_{[1]})/Z^\star_{[0]} \times 100\%$. From Figure 3.4a, we observe that the difference $Z_{\text{w/o}} - Z_{\text{abs}}$ shows an increasing trend as OVG increases. That is, modeling nurse absenteeism becomes more valuable as the value of operational flexibility increases. This makes sense because when a unit is short of supply due to nurse absenteeism, making it up with pool nurses are less expensive than doing so with temporary nurses. As a result, setting up nurse pools can effectively mitigate the impacts of nurse absenteeism. In Figure 3.4b, we depict the values of $x_{\text{w/o}}$ and x_{abs} obtained in the 100 replications and make similar observations.

3.6.4 Comparison among various pool structures

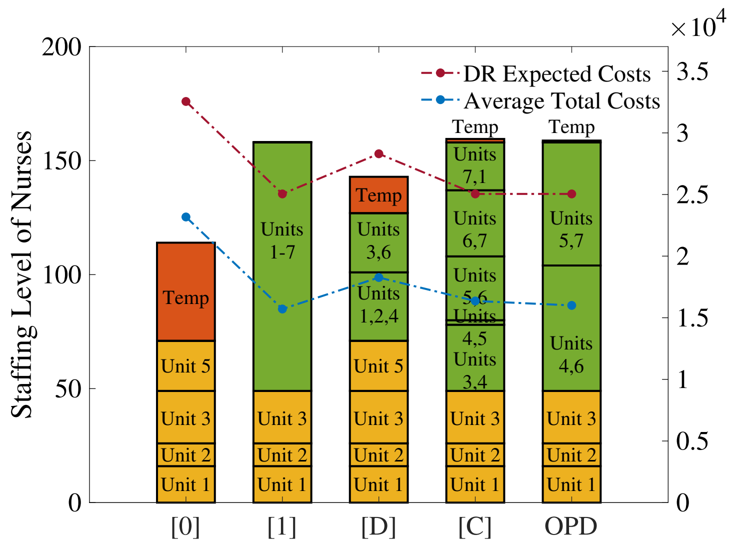


Figure 3.5: Staffing levels and out-of-sample performance under various pool structures

We compare the operational cost of nurse staffing under Structures [0], [1], [D], [C], and under the optimal pool design obtained from the (OPD) model. To this end, we generate a set of random test instances and solve each instance under all the structures. Then, by fixing the nurse staffing levels at the obtained optimal solution under each structure, we conduct an out-of-sample simulation to compute the average staffing cost of each solution based on scenarios of nurse demand and absenteeism. In this experiment, we observe that a sparse nurse pool design can often achieve similar out-of-sample performance as under Structure [1]. We report the input parameters of a representative instance in Appendix B.16 and the results of this instance in Figure 3.5. From this figure, we observe that the total number of unit nurses and that of pool nurses hired under Structures [1], [C], and

(OPD) are similar. Likewise, their DR staffing costs and out-of-sample average staffing cost are close. Nevertheless, these structures are drastically different in the amount of cross-training. For example, Structures [1] and (OPD) cross-train 21 and 2 pairs of units, respectively. That is, by cross-training two pairs of units, the optimal pool design produced by (OPD) harvests nearly all the operational flexibility of cross-training all the units.

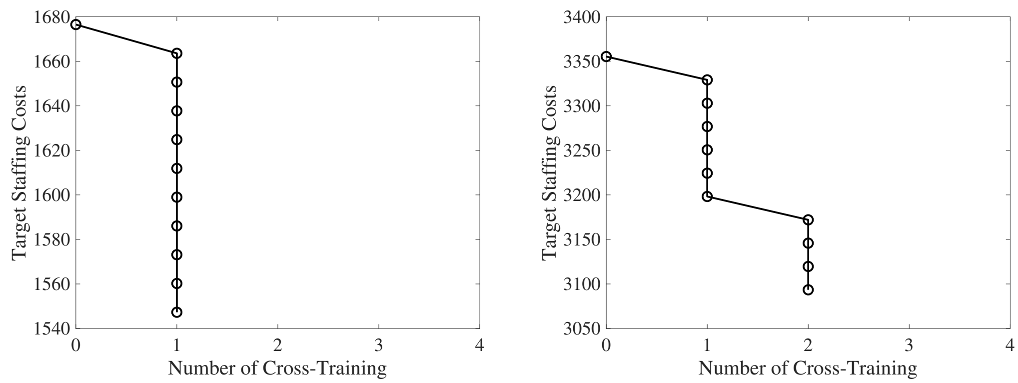


Figure 3.6: Amount of cross-training as a function of the target operational cost

To further verify this observation, we generate another set of random test instances and conduct sensitivity analysis on the target staffing cost T in the (OPD) model (the input parameters are specified later in Section 3.6.5). Specifically, we uniformly pick ten values of T between $Z^\star_{[0]}$ (i.e., the optimal value of (DRNS) with no nurse pools) and $Z^\star_{[1]}$ (i.e., the optimal value of (DRNS) under Structure [1]). For each value of T, we solve (OPD) to obtain the minimum amount of cross-training #(T) that guarantees that the DR staffing cost is no larger than T. We report the curve of #(T) in two representative instances in Figure 3.6. The results confirm our observations from Figure 3.5.

3.6.5 Patterns of the optimal nurse pool design

We notice from Figure 3.5 that a sparse pool design does not simply yield good out-of-sample performance. For example, in this figure, Structure [D] pools together units 3, 6 and units 1, 2, 4, and yields a considerably higher out-of-sample average staffing cost than that of (OPD), which pools together units 4, 6 and units 5, 7. From the input parameters of this instance (see Appendix B.16), we observe that units 1, 2, and 3 have lower variability in nurse demand and lower nurse absence rate, while the remaining units are more variable in both nurse demand and absenteeism. We hence conjecture that an optimal design tends to pool together units with higher variability (i.e., higher standard deviation in demand and/or higher absence rate). We numerically verify this conjecture in the following experiments.

We generate a set of random test instances with 4 or 8 units, $c^{\mathrm{x}}/c^{\mathrm{w}} = 2$, $c^{\mathrm{y}}/c^{\mathrm{w}} = 1.1$, $A_i^{\mathrm{p}} = 0.99$, $\overline{S} \in \{0.5, 1.0, 1.5\}$, and $\underline{S} \in \{0.1, 0.2, 0.3\}$. In addition, we divide the units into two disjoint subsets $\mathcal{A}$ and $\mathcal{B}$, where units in $\mathcal{A}$ have lower variability and those in $\mathcal{B}$ have higher variability. We consider the following three cases depending on what variability refers to:

Table 3.2: Three cases on constructing subsets $\mathcal{A}$ and $\mathcal{B}$

	Units in $\mathcal{A}$	Units in $\mathcal{B}$
Case 1	Low sd_j, Low A_j^{u}	High sd_j, Low A_j^{u}
Case 2	Low sd_j, Low A_j^{u}	Low sd_j, High A_j^{u}
Case 3	Low sd_j, Low A_j^{u}	High sd_j, High A_j^{u}

Note that (i) a value of low (respectively, high) standard deviation of nurse demand is randomly extracted from the interval $[7.24, 7.92]$ (respectively, $[17.14, 18.42]$), (ii) a value of low (respectively, high) absence rate is randomly extracted from the interval $[0.02, 0.04]$ (respectively, $[0.20, 0.40]$), and (iii) the mean value of nurse demand is randomly extracted from the interval $[25, 27]$. Finally, we set $T = Z^{\star}_{[1]}$, i.e., we are interested in the most sparse pool structures that produce equally good DR staffing cost as under Structure [1].

Table 3.3: OPD and OVG $(\%)$ the instances

4-Unit System			8-Unit System		
Case 1	#Pools for each type	OVG (%)	Case 1	#Pools for each type	OVG (%)
Instance 1-1	[0,0,2]	14.17	Instance 1-1	[1,1,2]	13.45
Instance 1-2	[1,1,0]	14.36	Instance 1-2	[1,1,2]	14.72
Instance 1-3	[0,0,2]	15.08	Instance 1-3	[1,1,2]	13.42
Instance 1-4	[1,1,0]	14.58	Instance 1-4	[1,1,2]	14.35
Instance 1-5	[0,0,2]	13.92	Instance 1-5	[1,1,2]	14.07
Case 2	#Pools for each type	OVG (%)	Case 2	#Pools for each type	OVG (%)
Instance 2-1	[0,1,0]	20.04	Instance 2-1	[0,2,0]	15.48
Instance 2-2	[0,1,0]	19.68	Instance 2-2	[0,2,0]	16.67
Instance 2-3	[0,1,0]	15.14	Instance 2-3	[0,2,0]	17.48
Instance 2-4	[0,1,0]	15.86	Instance 2-4	[0,2,0]	17.11
Instance 2-5	[0,1,0]	20.86	Instance 2-5	[0,2,0]	16.06
Case 3	#Pools for each type	OVG (%)	Case 3	#Pools for each type	OVG (%)
Instance 3-1	[0,1,0]	11.98	Instance 3-1	[0,2,0]	11.55
Instance 3-2	[0,1,0]	12.89	Instance 3-2	[0,2,0]	12.21
Instance 3-3	[0,1,0]	11.01	Instance 3-3	[0,2,0]	11.43
Instance 3-4	[0,1,0]	11.23	Instance 3-4	[0,2,0]	12.69
Instance 3-5	[0,1,0]	11.37	Instance 3-5	[0,2,0]	12.07

We classify the pools produced by (OPD) into three types based on the variability of the units a pool covers. We call a pool “Type-1” if all the units in this pool come from subset

$\mathcal{A}$, "Type-2" if all the units come from $\mathcal{B}$, and "Type-3" if the units come from both $\mathcal{A}$ and $\mathcal{B}$. We report the frequencies of each type appearing in an optimal pool design and the corresponding OVG in Table 3.3. [3] Take Instance 1-1 in Case 1 with the 8-unit system for example. The optimal design of this instance (see $[1, 1, 2]$ in the fifth column) consists of one Type-1 pool, one Type-2 pool, and two Type-3 pools. From this table, we observe that the optimal pool design diversifies among all three types in Case 1, i.e., the pools include both units with low variability and those with high variability. In contrast, in Cases 2 and 3, Type-2 pools become dominant, i.e., a majority of the pools include only the units with high variability. This observation numerically confirms our conjecture.

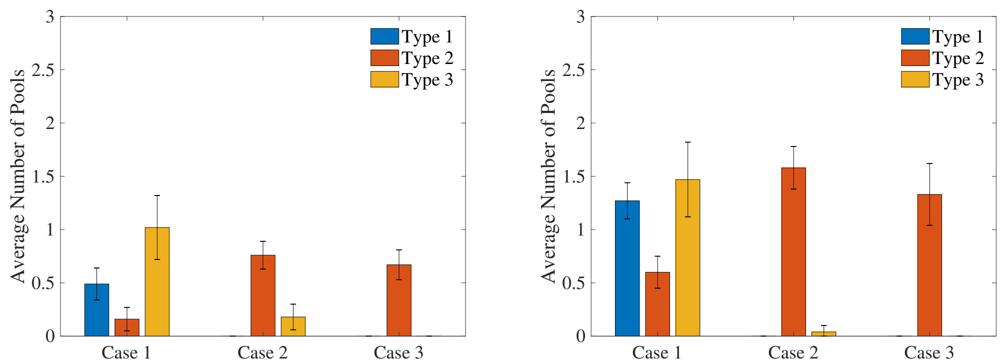


Figure 3.7: Types of the optimal pool design

We report the average number of each type appearing in an optimal pool design among all instances in Figure 3.7, where the error bars represent the corresponding 80%-confidence interval. From this figure, we observe that the Type-2 pools become dominant as we move to Cases 2 or 3. This once again confirms our conjecture numerically. In addition, we notice that the dominance of the Type-2 pools vanishes when moving from Case 3 to Case 1, i.e., when the variability of absenteeism decreases and that of demand remains unchanged. In contrast, the dominance of the Type-2 pools stays the same when moving from Case 3 to Case 2, i.e., when the variability of absenteeism remains unchanged and that of demand decreases. This indicates that the variability of nurse absenteeism plays a more important role in deciding the pattern of the optimal pool design. Hence, this result suggests that we should prioritize pooling together the units with higher variability, and especially those with higher nurse absence rates.

[3]The results in this table are associated with $\overline{S} = 1.5$ and $\underline{S} = 0.1$, but the observations remain the same for all other $\overline{S}$, $\underline{S}$ combinations.

3.7 Concluding Remarks

We studied a two-stage (DRNS) model for nurse staffing under both exogenous demand uncertainty and endogenous absenteeism uncertainty. We derived a min-max reformulation for (DRNS) under arbitrary nurse pool structures, leading to a separation algorithm that provably finds a globally optimal solution within a finite number of iterations. Under practical pool structures including one pool, disjoint pools, and chained pools, we derived monolithic MILP reformulations for (DRNS) and significantly improved the computational efficacy. Via numerical case studies, we found that modeling absenteeism improves the out-of-sample performance of staffing decisions, and such improvement is positively correlated with the value of operational flexibility. For nurse pool design, we found that sparse pool structures can already harvest most of the operational flexibility. More importantly, it is particularly effective to pool together the units with higher nurse absence rates.

CHAPTER 4

Distributionally Robust Transmission Grid Operation under Geomagnetic Disturbances

4.1 Introductory Remarks

Geomagnetic disturbances (GMD) event is a natural disaster caused by solar storms. During these storms, charged particles escape from the sun, travel to the earth, and cause variation in the geo-electric field, i.e., the electric field at the surface of the earth. The change in geo-electric field drives geomagnetically-induced currents (GIC) in the transmission grid, which has detrimental impacts, such as current distortions, a saturation of transformers inducing hot-spot heating, and increased reactive power losses [1, 140, 141], each of which causes unreliability of the power system (see Figure 4.1).

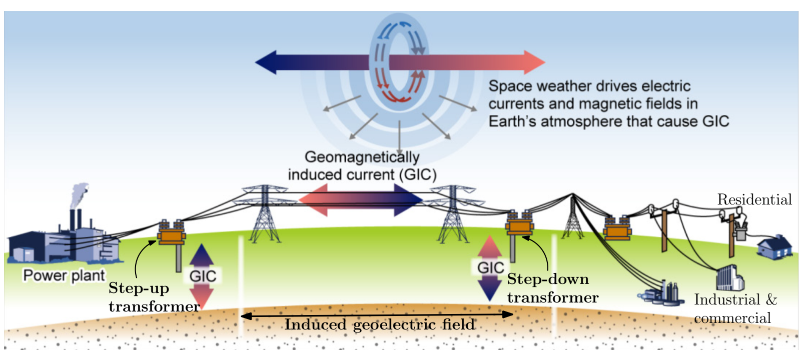


Figure 4.1: Effect of GMD on the transmission grid. (Source: [1])

The effect of GMD on ground-based electrical systems has been known for over 170 years. The first impacts of the GMD on the telegraph systems were observed in the 1840s. After the power grid came into existence in the 1880s, the GMD-induced blackouts and the other negative impacts were experienced later in the 1990s. For example, in 1989, the Hydro-Quebec power system was shut down for 9 hours due to GMD, which led to a net loss of \$13.2 million [142]. Recently in the early 2010s, the U.S. Department of Energy and the North American Electric Reliability Corporation (NERC) published a joint report [143] pointing out that the GMD is one of the high-impact, low-frequency event risks to the North American bulk power system. Also, NERC published an application guide [144], which is about computing GIC in the bulk power system, to better understand how GMD affects our power system. Likewise, enhancing power system resiliency to GMD becomes a very important topic.

There exist several approaches that mitigate the harmful effects of GMD on the power system. For example, one can install Direct Current (DC) blocking devices at regional substations to prevent the GIC from entering the power network through transformer neutrals [145]. This has led to work like [146], which posed the GIC blocking device placement problem that minimizes the costs of selecting appropriate locations to install these devices. Recently, numerous researchers have suggested that the risk of GIC could be reduced by the use of existing controls such as generator dispatch, line switching, and load shedding. In [147], for the first time, the authors proposed an optimal transmission line switching (OTS) model under GMD based on Alternating Current (AC) power flow equations and a set of constraints that captures GIC effects on various types of transformers. They utilized state-of-the-art convex relaxations and formulated the problem as a mixed-integer quadratic convex program, which could be solved using commercial optimization solvers. Recently, [148] presented heuristic-based algorithms to mitigate the effect of GIC on transformers by using line switching strategies on large-scale grids. Given that the GMD events are hard to predict in advance as the probability distribution of the magnitude and orientation of GMD is not known precisely due to the insufficient historical data, [149] proposed a two-stage DRO model with a mean-support ambiguity set and applied the column-and-constraint generation (CCG) algorithm [150] to solve on a small-scale instance, but with an average run time up to 3 hours.

In this chapter (see more details in [151]), we formulate a modified and improved version of the two-stage DRO formulation presented in [149]. The first-stage problem models the AC Optimal Transmission Switching (AC-OTS) which determines active transmission lines and the set-points for generators which minimize the worst-case expected costs occurred by GMD, i.e., taking expectation over the worst-case probability distribution among all the distributions in the ambiguity set. Given the solutions of the first-stage, the second-stage problem consists of linear constraints that capture the GIC effects. We assume the mean-support ambiguity set is provided and is uniform throughout the grid. The support set of the uncertain parameters is convex and can be approximated by a polytope with N extreme points. With these assumptions, the contributions of this chapter are:

1. We first reformulate the two-stage DRO model as a min-max-min problem that can

be solved by a CCG algorithm which solves a sequence of mixed integer second order cone programs (MISOCPs).

2. For the special case of the support set with *three* extreme points, we prove that the two-stage DRO model can be equivalently reformulated as a two-stage stochastic program with 3 scenarios. The resulting monolithic MISOCP can be solved efficiently using commercial solvers.

3. We propose a modified CCG algorithm that utilizes the MISOCP reformulation to enhance computation.

4. We present a detailed numerical analysis on the 'epri-21' and 'uiuc-150' systems, which are designed specifically for the GMD studies.

For the rest of this Chapter, we describe our mathematical models in Section 4.2, solution methodologies in Section 4.3, and numerical studies in Section 4.4.

4.2 Mathematical Formulation

In this section, we describe mathematical models that find a set of corrective actions which mitigates the negative impacts of GMD on the transmission grid. Note that the formulation presented in this section focuses on the quasi-static case (single time period).

4.2.1 GIC modeling

In this section, we describe how to calculate the GIC when the GMD occur. More details can be found in [141].

Geo-electric field is an electric field at the surface of the earth, whose unit is [V/km] and it has a direction. Based on the geomagnetic data and the earth conductivity models, one can calculate the geo-electric field, which can be used as a measure of the induction hazard to artificial conductors such as transmission grid. Historically severe GMD lead to the 4 – 20 V/km of geo-electric field. For example, the 1989 Quebec storm recorded the maximum of 8 V/Km. Based on the common assumption in the literature that the geo-electric field is uniform within an interconnected power system, we consider two components of the geo-electric field, which are the eastward component $\tilde{\xi}_{\text{E}}$ and the northward component $\tilde{\xi}_{\text{N}}$.

Variation in the geo-electric field due to GMD induce current on the conductors, such as transmission lines and transformers. This is known as GIC which is considered as quasi-DC due to its low frequencies (< 0.1 Hz) compared to the AC (50 – 60 Hz). From a perspective of power system modeling, the GIC can be treated as a DC which allows us to utilize the DC flow analysis to calculate the GIC. We describe how to convert the AC power network into its equivalent DC power network in Section 4.2.1.1, and how to calculate the GIC in the DC power network and its impact on the AC power network in Section 4.2.1.2.

4.2.1.1 AC and DC power network representation

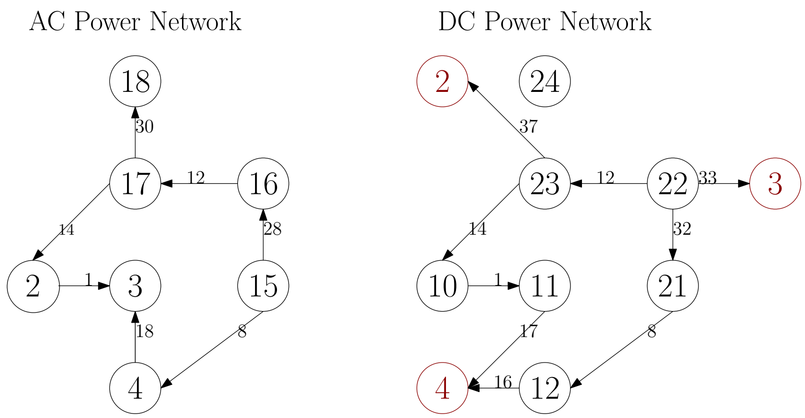


Figure 4.2: An example of AC to DC network mapping

The AC power network is represented by a graph $(\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is a set of nodes and $\mathcal{E}$ is a set of arcs. This network is the standard representation used for modeling AC power flow physics in power system applications. The set $\mathcal{E}$ is composed of $\mathcal{E}^\tau$, a set of transformers, and $\mathcal{E} \setminus \mathcal{E}^\tau$, a set of transmission lines. The set $\mathcal{N}$ is composed of buses that are adjacent to transmission lines and/or transformers. For calculating GIC (details in subsequent sections), which is a quasi-DC flow, we construct a DC power network $(\mathcal{N}^{\mathrm{d}}, \mathcal{E}^{\mathrm{d}})$. $\mathcal{N}^{\mathrm{d}}$ includes buses in $\mathcal{N}$ and additional nodes that model the neutral grounding points of transformers. The set $\mathcal{E}^{\mathrm{d}}$ includes transmission lines in $\mathcal{E}$ and additional lines between the end points of transformers and their neutrals. The transformer configurations depend on the type of the transformer. In this chapter, we consider 3 types of transformers, (i) Gwye-Gwye, (ii) GWye-GWye Auto, and (iii) GWye-Delta GSU. Detailed descriptions of these transformer types are found in [147, 149].

The mapping between the AC and DC network representations are described in Figure 4.2. In the AC network, $\mathcal{N} = \{2, 3, 4, 15, 16, 17, 18\}$, $\mathcal{E} = \{1, 8, 12, 14, 18, 28, 30\}$, and $\mathcal{E}^\tau = \{18, 28, 30\}$ (red arrows in Figure 4.2). The DC network $(\mathcal{N}^{\mathrm{d}}, \mathcal{E}^{\mathrm{d}})$ is constructed by adding information on different types of transformers and their connection to neutral (red circles in Figure 4.2). We first add neutral nodes $\{S_1, S_2, S_3\}$, each of which is connected to a transformer. The set $\mathcal{N}$ is relabeled with $\{2^d, 3^d, 4^d, 15^d, 16^d, 17^d, 18^d\}$ in the DC network. Topological mappings for different types of transformers are highlighted in Figure

4.2.

To link the two networks, we define functions E and E^{-1} that map $\ell \in \mathcal{E}^{\text{d}}$ to an edge $e \in \mathcal{E}$ and vice versa, i.e., if $E_\ell = e$, then $E_e^{-1} = \{\ell \in \mathcal{E}^{\text{d}} \ : \ E_\ell = e\}$. For example, line 17 in the DC network maps to the transformer line 18 in the AC network, thus $E_{17} = 18$. Transformer line 18 maps to lines 16 and 17 in DC network, thus $E_{18}^{-1} = \{16, 17\}$.

4.2.1.2 GIC calculation

This section describes how to calculate GIC based on the DC power flow analysis and how GIC affects different types of transformers in the AC power network. Under the assumption of an uniformly induced geo-electric field within an interconnected transmission grid, GIC that flow on a line in the DC network are given by

$$I_\ell^{\text{d}} = \gamma_\ell (v_m^{\text{d}} - v_n^{\text{d}} + \tilde{\xi}_\ell), \ \forall \ell_{mn} \in \mathcal{E}^{\text{d}}$$

where, γ_ℓ is a conductance of line ℓ, $v_m^{\text{d}}, v_n^{\text{d}}$ are voltage magnitude at bus m and n, and $\tilde{\xi}_\ell$ is the GIC-induced voltage source, given by:

$$\tilde{\xi}_\ell = \begin{cases} \tilde{\xi}_{\text{E}} L_\ell^{\text{E}} + \tilde{\xi}_{\text{N}} L_\ell^{\text{N}}, \ \forall \ell \in \mathcal{E}^{\text{d}} : E_\ell \in \mathcal{E} \setminus \mathcal{E}^\tau, \\ 0, \ \forall \ell \in \mathcal{E}^{\text{d}} : E_\ell \in \mathcal{E}^\tau \end{cases}$$

where, $\tilde{\xi}_{\text{E}}$ and $\tilde{\xi}_{\text{N}}$ are geo-electric fields [V/km] in the eastward and northward direction, respectively, and L_ℓ^{E} and L_ℓ^{N} are the length [km] of transmission line ℓ in the eastward and northward direction, respectively.

Based on the GIC in the DC network, the effective GIC of transformers in the AC network are calculated by

$$I_e^{\text{eff}} = |\Theta(I_\ell^{\text{d}}, \ \forall \ell \in E_e^{-1})|, \ \forall e \in \mathcal{E}^\tau$$

where $\Theta(I_\ell^{\text{d}}, \ \forall \ell \in E_e^{-1})$ is a linear function of GIC (I_ℓ^{d}). This function depends on the type of transformers as described in Table 4.1. Note that (N_h, N_l, N_s, N_c) are parameters which indicate the number of turns in the high-side/low-side/series/common winding, respectively.

Lastly, given a set $\mathcal{E}_i^\tau$ of transformers connected to node $i \in \mathcal{N}$, the reactive power loss due to GIC at node i in the AC network is calculated by

$$\sum_{e \in \mathcal{E}_i^\tau} k_e v_i I_e^{\text{eff}}, \ \forall i \in \mathcal{N}$$

Table 4.1: Effective GIC for each type of transformers

Type of transformer e	E_e^{-1}	$\Theta(I^{\text{d}}_\ell,\ \forall \ell \in E_e^{-1})$
Gwye-Gwye	$\{h, l\}$	$\Theta(I^{\text{d}}_h, I^{\text{d}}_l) = \frac{N_h I^{\text{d}}_h + N_l I^{\text{d}}_l}{N_h}$
GWye-GWye Auto	$\{s, c\}$	$\Theta(I^{\text{d}}_s, I^{\text{d}}_c) = \frac{N_s I^{\text{d}}_s + N_c I^{\text{d}}_c}{N_s + N_c}$
GWye-Delta GSU	$\{h\}$	$\Theta(I^{\text{d}}_h) = I^{\text{d}}_h$

where k_e is a loss factor of transformer e and v_i is a voltage magnitude at bus i.

4.2.2 Deterministic model

In this section we propose deterministic models that provide a set of corrective actions, such as line switching, generation dispatch, and load shedding, that mitigates the negative impacts of GMD on the transmission grid under the assumption that the geo-electric field $(\tilde{\xi}_{\text{E}}, \tilde{\xi}_{\text{N}})$ is given precisely. Note that Table 4.2 displays nomenclature used in our formulation.

4.2.2.1 Mixed-integer non-linear program

We present the following mixed-integer non-linear program:

$$\min \sum_{k \in \mathcal{G}} (c^{\text{F0}}_k z^{\text{g}}_k + c^{\text{F1}}_k f^{\text{p}}_k + c^{\text{F2}}_k (f^{\text{p}}_k)^2) + \sum_{i \in \mathcal{N}} \kappa^{\text{l}} (l^{\text{p+}}_i + l^{\text{p-}}_i + l^{\text{q+}}_i + l^{\text{q-}}_i) + \sum_{i \in \mathcal{N}} \kappa^{\text{s}} s_i$$

$$\text{s.t. } z^{\text{a}}_e \in \mathbb{B},\ \forall e \in \mathcal{E},\ \ z^{\text{g}}_k \in \mathbb{B},\ \forall k \in \mathcal{G}, \tag{4.1a}$$

$$\underline{g}^{\text{p}}_k z^{\text{g}}_k \le f^{\text{p}}_k \le \overline{g}^{\text{p}}_k z^{\text{g}}_k,\ \ \underline{g}^{\text{q}}_k z^{\text{g}}_k \le f^{\text{q}}_k \le \overline{g}^{\text{q}}_k z^{\text{g}}_k,\ \forall k \in \mathcal{G}, \tag{4.1b}$$

$$\sum_{e \in \mathcal{E}_k} z^{\text{a}}_e \ge z^{\text{g}}_k,\ \forall k \in \mathcal{G}, \tag{4.1c}$$

$$\underline{v}_i \le v_i \le \overline{v}_i,\ \ l^{\text{p+}}_i, l^{\text{p-}}_i, l^{\text{q+}}_i, l^{\text{q-}}_i \ge 0,\ \forall i \in \mathcal{N}, \tag{4.1d}$$

$$\sum_{e \in \mathcal{E}_i} p_{ei} = \sum_{k \in \mathcal{G}_i} f^{\text{p}}_k - d^{\text{p}}_i + l^{\text{p+}}_i - l^{\text{p-}}_i - g^{\text{s}}_i w_i,\ \forall i \in \mathcal{N}, \tag{4.1e}$$

$$\sum_{e \in \mathcal{E}_i} q_{ei} = \sum_{k \in \mathcal{G}_i} f^{\text{q}}_k - d^{\text{q}}_i + l^{\text{q+}}_i - l^{\text{q-}}_i + b^{\text{s}}_i w_i - d^{\text{qloss}}_i,\ \forall i \in \mathcal{N}, \tag{4.1f}$$

$$p^2_{ei} + q^2_{ei} \le z^{\text{a}}_e (\overline{s}^2_e),\ \ p^2_{ej} + q^2_{ej} \le z^{\text{a}}_e (\overline{s}^2_e),\ \forall e_{ij} \in \mathcal{E}, \tag{4.1g}$$

$$p_{ei} = z^{\text{a}}_e \Big\{ \frac{1}{\alpha^2_{ij}} g_e w_i - \frac{1}{\alpha_{ij}} (g_e w^{\text{c}}_e + b_e w^{\text{s}}_e) \Big\},\ \forall e_{ij} \in \mathcal{E}, \tag{4.1h}$$

$$p_{ej} = z^{\text{a}}_e \Big\{ g_e w_j - \frac{1}{\alpha_{ij}} (g_e w^{\text{c}}_e - b_e w^{\text{s}}_e) \Big\},\ \forall e_{ij} \in \mathcal{E}, \tag{4.1i}$$

$$q_{ei} = z^{\text{a}}_e \Big\{ -\frac{1}{\alpha^2_{ij}} (b_e + \frac{b^{\text{c}}_e}{2}) w_i + \frac{1}{\alpha_{ij}} (b_e w^{\text{c}}_e - g_e w^{\text{s}}_e) \Big\},\ \forall e_{ij} \in \mathcal{E}, \tag{4.1j}$$

Table 4.2: Nomenclature

Sets and parameters	
$\mathcal{G},\ \mathcal{N},\ \mathcal{E}$	a set of generators, buses, and lines in AC network
$\mathcal{E}^{\tau} \subseteq \mathcal{E}$	a set of transformers
$\mathcal{E}_i \subseteq \mathcal{E}$	a set of lines connected to $i \in \mathcal{N}$
c_k^{F0}	fixed cost when turning on $k \in \mathcal{G}$
$c_k^{\text{F1}}, c_k^{\text{F2}}$	fuel cost coefficients of power generation of $k \in \mathcal{G}$
$\underline{g}_k^{\text{p}}, \overline{g}_k^{\text{p}}$	bounds on the real power generation of $k \in \mathcal{G}$
$\underline{g}_k^{\text{q}}, \overline{g}_k^{\text{q}}$	bounds on the reactive power generation of $k \in \mathcal{G}$
κ^{l}	penalty cost for the power unbalance at $i \in \mathcal{N}$
κ^{s}	penalty cost for the exceeding amount of reactive power loss due to GIC at $i \in \mathcal{N}$
$d_i^{\text{p}}, d_i^{\text{q}}$	real and reactive power demand at $i \in \mathcal{N}$
$\underline{v}_i, \overline{v}_i$	voltage limits at $i \in \mathcal{N}$
$g_i^{\text{s}}, b_i^{\text{s}}$	shunt conductance and susceptance at $i \in \mathcal{N}$
g_e, b_e	conductance, susceptance of $e \in \mathcal{E}$
b_e^{c}	line charging susceptance of $e \in \mathcal{E}$
$\overline{s}_e$	apparent power limit of line $e \in \mathcal{E}$
α_{ij}	tap ratio of $e_{ij} \in \mathcal{E}$
k_e	loss factor of transformer $e \in \mathcal{E}^{\tau}$
$\overline{I}_e^{\text{eff}}$	upper limit of the effective GIC on $e \in \mathcal{E}^{\tau}$
$\mathcal{N}^{\text{d}}, \mathcal{E}^{\text{d}}$	a set of nodes and arcs in DC network
$\mathcal{E}_m^{\text{d}-}, \mathcal{E}_m^{\text{d}+}$	a set of incoming and outgoing arcs connected to $m \in \mathcal{N}^{\text{d}}$
γ_ℓ	conductance of $\ell \in \mathcal{E}^{\text{d}}$
a_m	inverse of ground resistance at $m \in \mathcal{N}^{\text{d}}$
$\overline{v}^{\text{d}}$	bound on the GIC-induced voltage magnitude
$\tilde{\xi}_\ell$	(random) GIC-induced voltage sources on $\ell \in \mathcal{E}^{\text{d}}$
Variables	
$z_e^{\text{a}} \in \mathbb{B}$	$z_e^{\text{a}} = 1$ if $e \in \mathcal{E}$ is turned on, and $z_e^{\text{a}} = 0$ otherwise
$z_k^{\text{g}} \in \mathbb{B}$	$z_k^{\text{g}} = 1$ if $k \in \mathcal{G}$ is turned on, and $z_k^{\text{g}} = 0$ otherwise
$l_i^{\text{p+}}, l_i^{\text{p-}}$	real power shedding at $i \in \mathcal{N}$
$l_i^{\text{q+}}, l_i^{\text{q-}}$	reactive power shedding at $i \in \mathcal{N}$
v_i	voltage magnitude at $i \in \mathcal{N}$
θ_i	phase angle at $i \in \mathcal{N}$
$f_k^{\text{p}},\ f_k^{\text{q}}$	real and reactive power generated by $k \in \mathcal{G}$
$p_{ei},\ p_{ej}$	real power flow on $e_{ij} \in \mathcal{E}$ *from node* i and *to node* j
$q_{ei},\ q_{ej}$	reactive power flow on $e_{ij} \in \mathcal{E}$ *from node* i and *to node* j
w_i	$w_i = v_i^2,\ \forall i \in \mathcal{N}$
w_e^{c}	$w_e^{\text{c}} = v_i v_j \cos(\theta_i - \theta_j),\ \forall e_{ij} \in \mathcal{E}$
w_e^{s}	$w_e^{\text{s}} = v_i v_j \sin(\theta_i - \theta_j),\ \forall e_{ij} \in \mathcal{E}$
d_i^{qloss}	allowable reactive power loss $i \in \mathcal{N}$
I_ℓ^{d}	GIC that flow on $\ell \in \mathcal{E}^{\text{d}}$
I_e^{eff}	effective GIC on $e \in \mathcal{E}^{\tau}$
v_m^{d}	GIC-induced voltage magnitude at $m \in \mathcal{N}^{\text{d}}$
s_i	the exceeding amount of reactive power loss due to GIC at $i \in \mathcal{N}$

$$q_{ej} = z^{\text{a}}_{e}\Big\{ - (b_e + \frac{b^{\text{c}}_{e}}{2})w_j + \frac{1}{\alpha_{ij}}(b_e w^{\text{c}}_{e} + g_e w^{\text{s}}_{e})\Big\}, \; \forall e_{ij} \in \mathcal{E}, \tag{4.1k}$$

$$z^{\text{a}}_{e} \left\{(w^{\text{c}}_{e})^2 + (w^{\text{s}}_{e})^2 - w_i w_j\right\} = 0, \; \forall e_{ij} \in \mathcal{E}, \tag{4.1l}$$

$$z^{\text{a}}_{e}\{\tan(\theta_i - \theta_j) - \frac{w^{\text{s}}_{e}}{w^{\text{c}}_{e}}\} = 0, \; \underline{\theta}_{ij} \leq \theta_i - \theta_j \leq \overline{\theta}_{ij}, \; \forall e_{ij} \in \mathcal{E}, \tag{4.1m}$$

$$w_i = v_i^2, \; \forall i \in \mathcal{N}, \tag{4.1n}$$

$$I^{\text{d}}_{\ell} = z^{\text{a}}_{\text{E}_\ell}\{\gamma_\ell(v^{\text{d}}_{m} - v^{\text{d}}_{n} + \tilde{\xi}_\ell)\}, \; -\overline{v}^{\text{d}} \leq v^{\text{d}}_{m} - v^{\text{d}}_{n} \leq \overline{v}^{\text{d}}, \; \forall \ell_{mn} \in \mathcal{E}^{\text{d}}, \tag{4.1o}$$

$$\sum_{\ell \in \mathcal{E}^{d-}_{m}} I^{\text{d}}_{\ell} - \sum_{\ell \in \mathcal{E}^{d+}_{m}} I^{\text{d}}_{\ell} = a_m v^{\text{d}}_{m}, \; \forall m \in \mathcal{N}^{\text{d}}, \tag{4.1p}$$

$$I^{\text{eff}}_{e} = |\Theta(I^{\text{d}}_{\ell}, \; \forall \ell \in E^{-1}_{e})|, \;\; 0 \leq I^{\text{eff}}_{e} \leq \overline{I}^{\text{eff}}_{e}, \; \forall e \in \mathcal{E}^{\tau}, \tag{4.1q}$$

$$d^{\text{qloss}}_{i} = \sum_{e \in \mathcal{E}^{\tau}_{i}} k_e v_i I^{\text{eff}}_{e} - s_i, \; \forall i \in \mathcal{N}. \tag{4.1r}$$

Constraints (4.1a) ensure that line switching and generator dispatch variables to be binary. Constraints (4.1b) ensure that the real and reactive power generation are within their bounds if the generator is turned on. Constraints (4.1c) ensure that a generator is turned off if all the lines and transformers connected to the generator is off. Constraints (4.1d) ensure that the voltage magnitude and the shedding variables to be within their bounds. Constraints (4.1e) and (4.1f) are the real and reactive power balance equation, where d^{qloss}_{i} in (4.1f) represents the *allowable* reactive power loss. Note that if the reactive power loss due to GIC, $\sum_{e \in \mathcal{E}^{\tau}_{i}} k_e v_i I^{\text{eff}}_{e}$, exceeds the d^{qloss}_{i}, then it can lead to the unreliability of the power grid. Constraints (4.1g) ensure that the apparent power flow does not exceed its limit if the line is switched on. Constraints (4.1h) – (4.1n) are the AC power flow equations with the line switching variables z^{a}, which are mixed-integer non-linear equations. Constraints (4.1o) calculate GIC in the DC network. Constraints (4.1p) represents the GIC balance equations. Constraints (4.1q) calculate the effective GIC for each type of transformer in the AC network. Constraints (4.1r) calculate the reactive power losses due to GIC in the AC network, $\sum_{e \in \mathcal{E}^{\tau}_{i}} k_e v_i I^{\text{eff}}_{e}$, and the exceeding amount of reactive power losses due to GIC, i.e., s_i. The objective function is to minimize the power generation costs, the penalty for shedding loads, and the penalty for the exceeding amount of reactive power loss due to GIC.

4.2.2.2 Mixed-integer second-order Cone Program

In this section, we propose a mixed integer second-order cone program (MISOCP) which is obtained by relaxing a subset of constraints in (4.1). Specifically, we utilize the convex relaxation of the AC power flow equation with line switching variables (4.1h) – (4.1n), suggested by [152], and the McCormick relaxation, i.e., $w \in \langle x, y \rangle^{\text{MC}}$ is represented by

$$
\begin{aligned}
& x \in [\underline{x}, \overline{x}], \;\; y \in [\underline{y}, \overline{y}], \\
& w \geq \underline{x}y + x\underline{y} - \underline{x}\underline{y}, \;\; w \geq \overline{x}y + x\overline{y} - \overline{x}\overline{y}, \\
& w \leq \overline{x}y + x\underline{y} - \overline{x}\underline{y}, \;\; w \leq \underline{x}y + x\overline{y} - \underline{x}\overline{y},
\end{aligned}
$$

where $\underline{x}, \overline{x}, \underline{y}, \overline{y}$ are lower and upper bounds on the x and y, respectively, which are given. Note that if at least one of the variables is binary, then the McCormick relaxation is an exact reformulation.

First of all, we consider constraints (4.1h) – (4.1n), which is mainly to calculate the AC power flow on transmission lines. Based on the assumption that $0 < \underline{v}_i < \overline{v}_i$ and $-90^\circ \leq \underline{\theta}_{ij} \leq 0 \leq \overline{\theta}_{ij} \leq 90^\circ$, which holds true for numerous practical instances, we obtain the following bounds for $w_i, w^{\text{c}}_e, w^{\text{s}}_e$:

$$
\begin{aligned}
& \underline{w}_i = \underline{v}_i^2, \;\; \overline{w}_i = \overline{v}_i^2, \; \forall i \in \mathcal{N}, \\
& \underline{w}^{\text{c}}_e = 0, \;\; \overline{w}^{\text{c}}_e = \overline{v}_i \overline{v}_j, \; \forall e_{ij} \in \mathcal{E}, \\
& \underline{w}^{\text{s}}_e = \overline{v}_i \overline{v}_j \sin \underline{\theta}_{ij}, \;\; \overline{w}^{\text{s}}_e = \overline{v}_i \overline{v}_j \sin \overline{\theta}_{ij}, \; \forall e_{ij} \in \mathcal{E}.
\end{aligned}
$$

Let us introduce new variables $w^{\text{z}}_{ei} = z^{\text{a}}_e w_i$, $w^{\text{cz}}_e = z^{\text{a}}_e w^{\text{c}}_e$, $w^{\text{sz}}_e = z^{\text{a}}_e w^{\text{s}}_e$. According to [152], constraints (4.1h) – (4.1n) can be relaxed as follows:

$$p_{ei} = \frac{1}{\alpha_{ij}^2} g_e w^{\text{z}}_{ei} - \frac{1}{\alpha_{ij}} (g_e w^{\text{cz}}_e + b_e w^{\text{sz}}_e), \; \forall e_{ij} \in \mathcal{E}, \tag{4.2a}$$

$$p_{ej} = g_e w^{\text{z}}_{ej} - \frac{1}{\alpha_{ij}} (g_e w^{\text{cz}}_e - b_e w^{\text{sz}}_e), \; \forall e_{ij} \in \mathcal{E}, \tag{4.2b}$$

$$q_{ei} = -\frac{1}{\alpha_{ij}^2} (b_e + \frac{b^{\text{c}}_e}{2}) w^{\text{z}}_{ei} + \frac{1}{\alpha_{ij}} (b_e w^{\text{cz}}_e - g_e w^{\text{sz}}_e), \; \forall e_{ij} \in \mathcal{E}, \tag{4.2c}$$

$$q_{ej} = -(b_e + \frac{b^{\text{c}}_e}{2}) w^{\text{z}}_{ej} + \frac{1}{\alpha_{ij}} (b_e w^{\text{cz}}_e + g_e w^{\text{sz}}_e), \; \forall e_{ij} \in \mathcal{E}, \tag{4.2d}$$

$$w^{\text{z}}_{ei} \in \langle z^{\text{a}}_e, w_i \rangle^{\text{MC}}, \;\; w^{\text{z}}_{ej} \in \langle z^{\text{a}}_e, w_j \rangle^{\text{MC}}, \; \forall e_{ij} \in \mathcal{E}, \tag{4.2e}$$

$$\underline{w}^{\text{c}}_e z^{\text{a}}_e \leq w^{\text{cz}}_e \leq \overline{w}^{\text{c}}_e z^{\text{a}}_e, \;\; \underline{w}^{\text{s}}_e z^{\text{a}}_e \leq w^{\text{sz}}_e \leq \overline{w}^{\text{s}}_e z^{\text{a}}_e, \; \forall e \in \mathcal{E}, \tag{4.2f}$$

$$(w^{\text{cz}}_e)^2 + (w^{\text{sz}}_e)^2 \leq w^{\text{z}}_{ei} w^{\text{z}}_{ej}, \; \forall e_{ij} \in \mathcal{E}, \tag{4.2g}$$

$$\tan(\underline{\theta}_{ij}) w^{\text{cz}}_e \leq w^{\text{sz}}_e \leq \tan(\overline{\theta}_{ij}) w^{\text{cz}}_e, \; \forall e_{ij} \in \mathcal{E}, \tag{4.2h}$$

$$(v_i)^2 \leq w_i \leq (\overline{v}_i + \underline{v}_i) v_i - \overline{v}_i \underline{v}_i, \; \forall i \in \mathcal{N}, \tag{4.2i}$$

Next, we consider constraints (4.1o) – (4.1r), which mainly calculate the GIC and the reactive power loss due to GIC. We first rewrite (4.1o) by introducing Big-M coefficients, i.e., $M_\ell^+ = \overline{v}^{\text{d}} + \tilde{\xi}_\ell$, $M_\ell^- = \overline{v}^{\text{d}} - \tilde{\xi}_\ell$. Then, we relax equality in constraints (4.1q) into inequality and utilize the McCormick relaxation on the bilinear terms, $u^{\text{d}}_{ei} \in \langle I^{\text{eff}}_e, v_i \rangle^{\text{MC}}$, in constraints (4.1r). Based on that, constraints (4.1o) – (4.1r) can be relaxed as follows:

$$(v^{\text{d}}_m - v^{\text{d}}_n + \tilde{\xi}_\ell) - M_\ell^+(1 - z^{\text{a}}_{E_\ell}) \le \frac{I^{\text{d}}_\ell}{\gamma_\ell} \le (v^{\text{d}}_m - v^{\text{d}}_n + \tilde{\xi}_\ell) + M_\ell^-(1 - z^{\text{a}}_{E_\ell}),\ \forall \ell_{mn} \in \mathcal{E}^{\text{d}}, \tag{4.3a}$$

$$-z^{\text{a}}_{E_\ell} M_\ell^- \le \frac{I^{\text{d}}_\ell}{\gamma_\ell} \le z^{\text{a}}_{E_\ell} M_\ell^+,\ \forall \ell_{mn} \in \mathcal{E}^{\text{d}}, \tag{4.3b}$$

$$\sum_{\ell \in \mathcal{E}^{d-}_m} I^{\text{d}}_\ell - \sum_{\ell \in \mathcal{E}^{d+}_m} I^{\text{d}}_\ell = a_m v^{\text{d}}_m,\ \forall m \in \mathcal{N}^{\text{d}}, \tag{4.3c}$$

$$I^{\text{eff}}_e \ge \Theta(I^{\text{d}}_\ell,\ \forall \ell \in E_e^{-1}),\ \ I^{\text{eff}}_e \ge -\Theta(I^{\text{d}}_\ell,\ \forall \ell \in E_e^{-1}),\ \ 0 \le I^{\text{eff}}_e \le \overline{I}^{\text{eff}}_e,\ \forall e \in \mathcal{E}^\tau, \tag{4.3d}$$

$$d^{\text{qloss}}_i = \sum_{e \in \mathcal{E}^\tau_i} k_e u^{\text{d}}_{ei},\ \ u^{\text{d}}_{ei} \in \langle I^{\text{eff}}_e, v_i \rangle^{\text{MC}},\ \forall i \in \mathcal{N}. \tag{4.3e}$$

Finally, we obtain the following MISOCP:

$$\min \sum_{k \in \mathcal{G}} (c^{\text{F0}}_k z^{\text{g}}_k + c^{\text{F1}}_k f^{\text{p}}_k + c^{\text{F2}}_k (f^{\text{p}}_k)^2) + \sum_{i \in \mathcal{N}} \kappa^{\text{l}} (l^{\text{p+}}_i + l^{\text{p−}}_i + l^{\text{q+}}_i + l^{\text{q−}}_i) + \sum_{i \in \mathcal{N}} \kappa^{\text{s}} s_i \tag{4.4a}$$

$$\text{s.t. } (4.1\text{a}) - (4.1\text{g}), \tag{4.4b}$$

$$(4.2\text{a}) - (4.2\text{i}), \tag{4.4c}$$

$$(4.3\text{a}) - (4.3\text{e}). \tag{4.4d}$$

The MISOCP can be solved by commercial softwares to find a solution that mitigate the negative impact brought by the deterministic geo-electric field $\tilde{\xi}$ in constraints (4.3a). Note that the formulation (4.4) provides a lower bound on the optimal value of the formulation (4.1).

4.2.3 Distributionally robust optimization model

In this section we assume that the geo-electric field $\tilde{\xi}$ is uncertain when we make decision on the transmission grid operations. Moreover, we assume that the probability distribution of uncertain parameters $\tilde{\xi}$ is not known precisely due to the insufficient historical data, but the mean values and the support set of the uncertain parameters $\tilde{\xi}$ can be estimated relatively easily. Taking these view points into consideration, we propose a two-stage DRO model to find a set of corrective actions that operates the transmission grid so as to mitigate the future damage due to GMD. We describe the mean-support ambiguity set $\mathcal{D}$ in Section

4.2.3.1, and propose a two-stage DRO model in 4.2.3.2.

4.2.3.1 Ambiguity set

The geo-electric field $(\tilde{\xi}_{\text{E}}, \tilde{\xi}_{\text{N}})$ is uncertain when planning decisions are made to mitigate the risk of GMD. We assume that the support set and mean values of $(\tilde{\xi}_{\text{E}}, \tilde{\xi}_{\text{N}})$ are provided or are extracted from a set of historical data, but their joint probability distribution is unknown. Generally, there is not enough information to construct reasonable probability distributions [153]. In this setting, we construct the following mean-support ambiguity set $\mathcal{D}$:

$$\mathcal{D} := \{\mathbb{P} \;:\; \mathbb{E}_{\mathbb{P}_{\text{E}}}[\tilde{\xi}_{\text{E}}] = \mu_{\text{E}}, \;\; \mathbb{E}_{\mathbb{P}_{\text{N}}}[\tilde{\xi}_{\text{N}}] = \mu_{\text{N}}, \;\; \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \Xi\} = 1\}$$

where $(\mathbb{P}_{\text{E}}, \mathbb{P}_{\text{N}})$ and $(\mu_{\text{E}}, \mu_{\text{N}})$ are the marginal distributions and mean values of $(\tilde{\xi}_{\text{E}}, \tilde{\xi}_{\text{N}})$, respectively. The support set Ξ will be described later.

4.2.3.2 Two-stage DRO model

Based on the mean-support ambiguity set defined in 4.2.3.1, we propose a two-stage DRO model that aims to find how to operate the transmission grid via a set of corrective action that minimizes the expected costs due to GMD events over the worst-case distribution within the predefined ambiguity set. For the computational tractability, the proposed DRO model is built upon the MISOCP (4.4). Specifically, constraints (4.4b) and (4.4c) belong to the first-stage problem while constraints (4.4c) belong to the second-stage problem, i.e., GIC and its corresponding reactive power loss are calculated after uncertain geo-electric field $\tilde{\xi}$ is realized. We propose the following two-stage DRO model:

$$\min \sum_{k \in \mathcal{G}} (c_k^{\text{F0}} z_k^{\text{g}} + c_k^{\text{F1}} f_k^{\text{p}} + c_k^{\text{F2}} (f_k^{\text{p}})^2) + \sum_{i \in \mathcal{N}} \kappa^{\text{l}} (l_i^{\text{p+}} + l_i^{\text{p-}} + l_i^{\text{q+}} + l_i^{\text{q-}}) + \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(\boldsymbol{z}^{\text{a}}, \boldsymbol{v}, \boldsymbol{d}^{\text{qloss}}, \tilde{\boldsymbol{\xi}}] \tag{4.5a}$$

$$\text{s.t.} \sum_{e \in \mathcal{E}_i} p_{ei} = \sum_{k \in \mathcal{G}_i} f_k^{\text{p}} - d_i^{\text{p}} + l_i^{\text{p+}} - l_i^{\text{p-}} - g_i^{\text{s}} w_i, \; \forall i \in \mathcal{N}, \tag{4.5b}$$

$$\sum_{e \in \mathcal{E}_i} q_{ei} = \sum_{k \in \mathcal{G}_i} f_k^{\text{q}} - d_i^{\text{q}} + l_i^{\text{q+}} - l_i^{\text{q-}} + b_i^{\text{s}} w_i - d_i^{\text{qloss}}, \; \forall i \in \mathcal{N}, \tag{4.5c}$$

$$p_{ei}^2 + q_{ei}^2 \leq z_e^{\text{a}} (\overline{s}_e^2), \;\; p_{ej}^2 + q_{ej}^2 \leq z_e^{\text{a}} (\overline{s}_e^2), \; \forall e_{ij} \in \mathcal{E}, \tag{4.5d}$$

$$p_{ei} = \frac{1}{\alpha_{ij}^2} g_e w_{ei}^{\text{z}} - \frac{1}{\alpha_{ij}} (g_e w_e^{\text{cz}} + b_e w_e^{\text{sz}}), \; \forall e_{ij} \in \mathcal{E}, \tag{4.5e}$$

$$p_{ej} = g_e w_{ej}^{\text{z}} - \frac{1}{\alpha_{ij}} (g_e w_e^{\text{cz}} - b_e w_e^{\text{sz}}), \; \forall e_{ij} \in \mathcal{E}, \tag{4.5f}$$

$$q_{ei} = -\frac{1}{\alpha_{ij}^2} (b_e + \frac{b_e^{\text{c}}}{2}) w_{ei}^{\text{z}} + \frac{1}{\alpha_{ij}} (b_e w_e^{\text{cz}} - g_e w_e^{\text{sz}}), \; \forall e_{ij} \in \mathcal{E}, \tag{4.5g}$$

$$q_{ej} = -(b_e + \frac{b^{\mathrm{c}}_e}{2})w^{\mathrm{z}}_{ej} + \frac{1}{\alpha_{ij}}(b_e w^{\mathrm{cz}}_e + g_e w^{\mathrm{sz}}_e), \forall e_{ij} \in \mathcal{E}, \tag{4.5h}$$

$$w^{\mathrm{z}}_{ei} \in \langle z^{\mathrm{a}}_e, w_i \rangle^{\mathrm{MC}},\ \ w^{\mathrm{z}}_{ej} \in \langle z^{\mathrm{a}}_e, w_j \rangle^{\mathrm{MC}},\ \forall e_{ij} \in \mathcal{E}, \tag{4.5i}$$

$$\underline{w}^{\mathrm{c}}_e z^{\mathrm{a}}_e \le w^{\mathrm{cz}}_e \le \overline{w}^{\mathrm{c}}_e z^{\mathrm{a}}_e,\ \ \underline{w}^{\mathrm{s}}_e z^{\mathrm{a}}_e \le w^{\mathrm{sz}}_e \le \overline{w}^{\mathrm{s}}_e z^{\mathrm{a}}_e,\ \forall e \in \mathcal{E}, \tag{4.5j}$$

$$(w^{\mathrm{cz}}_e)^2 + (w^{\mathrm{sz}}_e)^2 \le w^{\mathrm{z}}_{ei} w^{\mathrm{z}}_{ej},\ \forall e_{ij} \in \mathcal{E}, \tag{4.5k}$$

$$\tan(\underline{\theta}_{ij}) w^{\mathrm{cz}}_e \le w^{\mathrm{sz}}_e \le \tan(\overline{\theta}_{ij}) w^{\mathrm{cz}}_e,\ \forall e_{ij} \in \mathcal{E}, \tag{4.5l}$$

$$(v_i)^2 \le w_i \le (\overline{v}_i + \underline{v}_i) v_i - \overline{v}_i \underline{v}_i,\ \forall i \in \mathcal{N}, \tag{4.5m}$$

$$\sum_{e \in \mathcal{E}_k} z^{\mathrm{a}}_e \ge z^{\mathrm{g}}_k,\ \forall k \in \mathcal{G}, \tag{4.5n}$$

$$\underline{v}_i \le v_i \le \overline{v}_i,\ \ l^{\mathrm{p+}}_i, l^{\mathrm{p-}}_i, l^{\mathrm{q+}}_i, l^{\mathrm{q-}}_i \ge 0,\ \forall i \in \mathcal{N}, \tag{4.5o}$$

$$\underline{g}^{\mathrm{p}}_k z^{\mathrm{g}}_k \le f^{\mathrm{p}}_k \le \overline{g}^{\mathrm{p}}_k z^{\mathrm{g}}_k,\ \ \underline{g}^{\mathrm{q}}_k z^{\mathrm{g}}_k \le f^{\mathrm{q}}_k \le \overline{g}^{\mathrm{q}}_k z^{\mathrm{g}}_k,\ \forall k \in \mathcal{G}, \tag{4.5p}$$

$$z^{\mathrm{a}}_e \in \{0,1\},\ \forall e \in \mathcal{E},\ \ z^{\mathrm{g}}_k \in \{0,1\},\ \forall k \in \mathcal{G}. \tag{4.5q}$$

where, given the values of $\boldsymbol{z}^{\mathrm{a}}, \boldsymbol{v}, \boldsymbol{d}^{\mathrm{qloss}}$ from the first-stage problem (4.5) and the realization of $\tilde{\boldsymbol{\xi}}$, $\mathcal{Q}(\boldsymbol{z}^{\mathrm{a}}, \boldsymbol{v}, \boldsymbol{d}^{\mathrm{qloss}}, \tilde{\boldsymbol{\xi}})$ is the optimal value of the following second-stage problem which evaluates GIC in the DC network:

$$\min \sum_{i \in \mathcal{N}} \kappa^{\mathrm{s}} s_i \tag{4.6a}$$

$$\text{s.t.}\ \frac{I^{\mathrm{d}}_\ell}{\gamma_\ell} \le (v^{\mathrm{d}}_m - v^{\mathrm{d}}_n + \tilde{\xi}_\ell) + M^-_\ell (1 - z^{\mathrm{a}}_{E_\ell}),\ \forall \ell_{mn} \in \mathcal{E}^{\mathrm{d}}, \tag{4.6b}$$

$$\frac{I^{\mathrm{d}}_\ell}{\gamma_\ell} \ge (v^{\mathrm{d}}_m - v^{\mathrm{d}}_n + \tilde{\xi}_\ell) - M^+_\ell (1 - z^{\mathrm{a}}_{E_\ell}),\ \forall \ell_{mn} \in \mathcal{E}^{\mathrm{d}}, \tag{4.6c}$$

$$- z^{\mathrm{a}}_{E_\ell} M^-_\ell \le \frac{I^{\mathrm{d}}_\ell}{\gamma_\ell} \le z^{\mathrm{a}}_{E_\ell} M^+_\ell,\ \forall \ell_{mn} \in \mathcal{E}^{\mathrm{d}}, \tag{4.6d}$$

$$\sum_{\ell \in \mathcal{E}^{d-}_m} I^{\mathrm{d}}_\ell - \sum_{\ell \in \mathcal{E}^{d+}_m} I^{\mathrm{d}}_\ell = a_m v^{\mathrm{d}}_m,\ \forall m \in \mathcal{N}^{\mathrm{d}}, \tag{4.6e}$$

$$I^{\mathrm{eff}}_e \ge \Theta(I^{\mathrm{d}}_\ell,\ \forall \ell \in E^{-1}_e),\ \ I^{\mathrm{eff}}_e \ge -\Theta(I^{\mathrm{d}}_\ell,\ \forall \ell \in E^{-1}_e), \tag{4.6f}$$

$$0 \le I^{\mathrm{eff}}_e \le \overline{I}^{\mathrm{eff}}_e,\ \forall e \in \mathcal{E}^\tau,\quad s_i \ge 0,\ \forall i \in \mathcal{N} \tag{4.6g}$$

$$u^{\mathrm{d}}_{ei} \in \langle v_i, I^{\mathrm{eff}}_e \rangle^{\mathrm{MC}},\ d^{\mathrm{qloss}}_i \ge \sum_{e \in \mathcal{E}^\tau_i} k_e u^{\mathrm{d}}_{ei} - s_i,\ \forall i \in \mathcal{N}. \tag{4.6h}$$

Constraints (4.5b)–(4.5p) model the relaxed AC power flow equations. These constraints include mitigation actions such as transmission line switching and switching of generators. Note that generator switching in this formulation *does not* imply that we model economic unit commitment. Constraints (4.5b) and (4.5c) model real and reactive power balance constraints, including allowable reactive power loss (d^{qloss}_i). Constraints (4.5d) ensure that the apparent power flow does not exceed its limit when the line is closed. Constraints (4.5e)–

(4.5h) model Ohm's law. Constraints (4.5i)–(4.5m) model the MISOCP relaxation of the AC power flow equations using the formulations discussed in [152]. Notation $\langle x_i, x_j \rangle^{\text{MC}}$ models the standard McCormick relaxation of a bilinear function $x_i \cdot x_j$, which is very effective for ACOPF problems [152, 154]. Constraints (4.5n) ensure that a generator is turned off when all the lines and transformers connected to that generator are off.

In the second-stage problem, constraints (4.6b)–(4.6d) calculate valid GIC in the DC network for lines that are switched on. Else, they are deactivated with the big-M coefficients, where $M_\ell^+ = \overline{v}^{\text{d}} + \tilde{\xi}_\ell$, $M_\ell^- = \overline{v}^{\text{d}} - \tilde{\xi}_\ell$. Constraints (4.6e) model nodal balance equation for GIC in the DC network. Note that, in (4.6e), $a_m = 0$ when m is not a grounded neutral node. Constraints (4.6f) and (4.6g) calculate the effective GIC for each type of transformer (see Table 4.1), which in-turn is used to calculate the reactive power losses induced in the AC network, as shown in constraints (4.6h).

Interpretation of the two-stage DRO formulation Once the uncertain geo-electric field ($\tilde{\boldsymbol{\xi}}$) is realized and the decisions in the AC network on line switching ($\boldsymbol{z}^{\text{a}}$), voltage magnitude ($\boldsymbol{v}$), and allowable reactive power losses $\boldsymbol{d}^{\text{qloss}}$ (interpreted as the amount of reactive power loss that will not cause excessive voltage drops. Exceeding this value would correspond to the cost of installing or using a device to counteract the reactive losses), the second-stage problem calculates the effective GIC and the actual reactive power losses, given by $\sum_{e \in \mathcal{E}_i^\tau} k_e u_{ei}^{\text{d}}$. If the calculated reactive power losses in second-stage exceeds the allowable $\boldsymbol{d}^{\text{qloss}}$ from the first-stage, then appropriate mitigation actions are updated in the AC network to mitigate the negative effects on the transformers. Implicitly, this formulation assumes that reactive losses smaller than $\boldsymbol{d}^{\text{qloss}}$ will not cause voltage problems (i.e., a high voltage).

In summary, the proposed two-stage DRO formulation minimizes the power generation cost, penalty cost for shedding loads and the worst-case expected cost occurred by damaged transformers.

4.3 Solution Approaches

In this section, we describe our solution approaches to the two-stage DRO formulation according to the definition of the support set Ξ. In Section 4.3.1, we focus on the half-circle support set described in [153]. In Section 4.3.2, we focus on the polyhedral support set which is a polytope with N extreme points.

For ease of exposition, we rewrite the formulation using matrix notation as follows:

$$\min_{\boldsymbol{z} \in \mathcal{F}} \boldsymbol{q}^{\text{T}} \boldsymbol{z} + \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \tag{4.7a}$$

where $\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})$ is the optimal value of the following problem:

$$\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) = \min_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})} \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}. \tag{4.7b}$$

Note that $\boldsymbol{z}$ and $\mathcal{F}$ are the solution vector and the feasible region, respectively, of the first-stage problem (4.5) and $\mathcal{X}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}(\tilde{\boldsymbol{\xi}})\boldsymbol{z} \geq \boldsymbol{d}\}$ is the feasible region of the second-stage problem.

Proposition 12. *Problem* (4.7) *is equivalent to the following* min-max-min *problem:*

$$\min_{\boldsymbol{z} \in \mathcal{F}} \boldsymbol{q}^T \boldsymbol{z} + \mu_E \lambda_E + \mu_N \lambda_N + \left\{ \max_{\tilde{\boldsymbol{\xi}} \in \Xi} \left(\min_{x \in \mathcal{X}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})} \boldsymbol{c}^T \boldsymbol{x} \right) - \lambda_E \tilde{\xi}_E - \lambda_N \tilde{\xi}_N \right\}.$$

Proof. The worst-case expected value $\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})]$ can be written as

$$\begin{aligned} \max \ & \int_{\tilde{\boldsymbol{\xi}} \in \Xi} \mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) \, d\mathbb{P} \\ \text{s.t. } & \int_{\tilde{\boldsymbol{\xi}} \in \Xi} \tilde{\xi}_{\mathrm{E}} \, d\mathbb{P} = \mu_{\mathrm{E}}, \ \int_{\tilde{\boldsymbol{\xi}} \in \Xi} \tilde{\xi}_{\mathrm{N}} \, d\mathbb{P} = \mu_{\mathrm{N}}, \ \int_{\tilde{\boldsymbol{\xi}} \in \Xi} d\mathbb{P} = 1. \end{aligned}$$

By taking the dual, we obtain

$$\begin{aligned} & \min \mu_{\mathrm{E}} \lambda_{\mathrm{E}} + \mu_{\mathrm{N}} \lambda_{\mathrm{N}} + \eta \\ & \text{s.t. } \ \lambda_{\mathrm{E}} \tilde{\xi}_{\mathrm{E}} + \lambda_{\mathrm{N}} \tilde{\xi}_{\mathrm{N}} + \eta \geq \ \mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}), \ \forall \tilde{\boldsymbol{\xi}} \in \Xi. \end{aligned}$$

where λ_{E}, λ_{N}, and η are dual variables. At optimality, we have $\eta^* = \max_{\tilde{\boldsymbol{\xi}} \in \Xi} \{\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) - \lambda_{\mathrm{E}} \tilde{\xi}_{\mathrm{E}} - \lambda_{\mathrm{N}} \tilde{\xi}_{\mathrm{N}}\}$, which leads to the proposed formulation. □

4.3.1 Half-circle support set

In this section, we consider the support set Ξ of $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_{\mathrm{E}}, \tilde{\xi}_{\mathrm{N}})$, which is defined as

$$\Xi := \left\{ \tilde{\boldsymbol{\xi}} = (\tilde{\xi}_{\mathrm{E}}, \tilde{\xi}_{\mathrm{N}}) \in \mathbb{R}^2 : -R \leq \tilde{\xi}_{\mathrm{E}} \leq R, \ \ 0 \leq \tilde{\xi}_{\mathrm{N}} \leq R, \ \ (\tilde{\xi}_{\mathrm{E}})^2 + (\tilde{\xi}_{\mathrm{N}})^2 \leq R^2 \right\}$$

where R is a radius of half-circle support set as shown in Figure 4.3. Practically speaking, the support set is bounded by an estimate of the worst-case magnitude of the geo-electric field. Since the square of the magnitude is equal to the sum of the squares of the northward and eastward field strength, a magnitude bound yields a half circle support set for $(\tilde{\xi}_{\mathrm{E}}, \tilde{\xi}_{\mathrm{N}})$ [153].

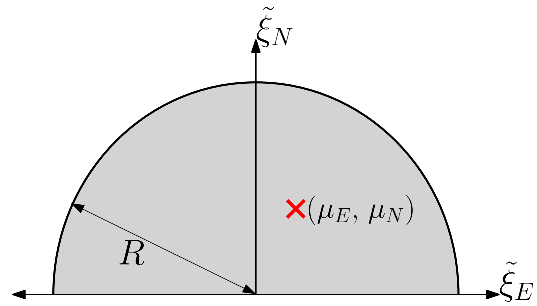


Figure 4.3: Support set and mean of uncertain GMD $(\tilde{\xi}_{\mathrm{E}}, \tilde{\xi}_{\mathrm{N}})$

Proposition 13. *Given solution ($\boldsymbol{z}$, λ_{E}, λ_{N}), the inner* max-min *problem can be reformulated as the following* max *problem:*

$$\mathcal{Z}(\boldsymbol{z}, \lambda_{E}, \lambda_{N}, \tilde{\boldsymbol{\xi}}) = \max_{\tilde{\boldsymbol{\xi}} \in \Xi} \boldsymbol{c}^{T}\boldsymbol{x} - \lambda_{E}\tilde{\xi}_{E} - \lambda_{N}\tilde{\xi}_{N} \tag{4.10a}$$

$$s.t.\ 0 \leq \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}(\tilde{\boldsymbol{\xi}})\boldsymbol{z} - \boldsymbol{d} \leq M(\boldsymbol{1} - \boldsymbol{\alpha}), \tag{4.10b}$$

$$0 \leq \boldsymbol{\beta} \leq M\boldsymbol{\alpha},\ \boldsymbol{A}^{T}\boldsymbol{\beta} = \boldsymbol{c},\ \boldsymbol{\alpha} \in \mathbb{B}^{m}. \tag{4.10c}$$

where M is a sufficiently large value.

Proof. Consider the following max-min problem:

$$\max_{\tilde{\boldsymbol{\xi}} \in \Xi} \left(\min_{x \in \mathcal{X}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})} \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \right) - \lambda_{\mathrm{E}}\tilde{\xi}_{\mathrm{E}} - \lambda_{\mathrm{N}}\tilde{\xi}_{\mathrm{N}}$$

Using the complementary slackness condition, we obtain

$$\max_{\tilde{\boldsymbol{\xi}} \in \Xi, \boldsymbol{x} \in \mathbb{R}^{n}, \boldsymbol{\beta} \in \mathbb{R}_{+}^{m}} \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \lambda_{\mathrm{E}}\tilde{\xi}_{\mathrm{E}} - \lambda_{\mathrm{N}}\tilde{\xi}_{\mathrm{N}} \tag{4.11a}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}(\tilde{\boldsymbol{\xi}})\boldsymbol{z} \geq \boldsymbol{d},\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\beta} = \boldsymbol{c}, \tag{4.11b}$$

$$\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}(\tilde{\boldsymbol{\xi}})\boldsymbol{z} - \boldsymbol{d}) = 0. \tag{4.11c}$$

where $\boldsymbol{\beta}$ is a dual vector, (4.11b) is primal and dual feasibility and (4.11c) is the complementary slackness. By introducing a big-M coefficient, constraints (4.11c) can be expressed as linear constraints, which leads to the formulation in (4.10). □

Based on proposition (13), the two-stage DRO problem with a half-circle support set is exactly solvable with the column-and-constraint generation (CCG) algorithm [150], as described in Algorithm 4.1. Note that the CCG algorithm is a cutting plane method, which iteratively refines the feasible domain of the two-stage DRO problem by sequentially generating a set of recourse variables and their associated constraints. In Algorithm 4.1, LB and

UB denote the incumbent lower and upper bounds of problem (4.7), respectively. Step-2 evaluates the lower bound for problem (4.7) by solving the relaxed master problem. Step-3 finds the worst-case value of $\tilde{\boldsymbol{\xi}}^*$ by solving problem (4.10). Step-4 updates the upper bound for the worst-case value of $\tilde{\boldsymbol{\xi}}^*$. Steps 5-7 terminate the algorithm if the optimality gap is within a specified ϵ, else steps 8-9 augment the master problem with the constraints and variables associated with extreme point $\tilde{\boldsymbol{\xi}}^*$ and continues the iteration until an optimal solution for problem (4.7) is found. Since the support set has infinitely many extreme points, the CCG approach can be computationally intensive.

1: Set LB=$-\infty$, UB=∞, $t = 0$ and $\mathcal{T} = \emptyset$.

2: Update LB by solving the following master problem:

$$\text{LB} = \min_{\boldsymbol{z} \in \mathcal{F}} \boldsymbol{q}^{\mathrm{T}} \boldsymbol{z} + \mu_{\mathrm{E}} \lambda_{\mathrm{E}} + \mu_{\mathrm{N}} \lambda_{\mathrm{N}} + \eta \tag{4.12a}$$

$$\text{s.t. } \eta \geq \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}^{\ell} - \lambda_{\mathrm{E}} \xi_{\mathrm{E}}^{\ell} - \lambda_{\mathrm{N}} \xi_{\mathrm{N}}^{\ell}, \ \forall \ell \in \mathcal{T}, \tag{4.12b}$$

$$\boldsymbol{A}\boldsymbol{x}^{\ell} + \boldsymbol{B}(\tilde{\boldsymbol{\xi}}^{\ell}) \boldsymbol{z} \geq \boldsymbol{d}, \ \forall \ell \in \mathcal{T}. \tag{4.12c}$$

– Record an optimal solution $\boldsymbol{z}^*$, λ_{E}^*, λ_{N}^*, and η^*.

3: Solve problem (4.10):

– Record an optimal $\tilde{\boldsymbol{\xi}}^*$ and the optimal value $\mathcal{Z}(\boldsymbol{z}^*, \lambda_{\mathrm{E}}^*, \lambda_{\mathrm{N}}^*, \tilde{\boldsymbol{\xi}}^*)$

4: Update UB by

$$\text{UB} = \min\{\text{UB}, \ \boldsymbol{q}^{\mathrm{T}} \boldsymbol{z}^* + \mu_{\mathrm{E}} \lambda_{\mathrm{E}}^* + \mu_{\mathrm{N}} \lambda_{\mathrm{N}}^* + \mathcal{Z}(\boldsymbol{z}^*, \lambda_{\mathrm{E}}^*, \lambda_{\mathrm{N}}^*, \tilde{\boldsymbol{\xi}}^*)\}$$

5: **if** $(\text{UB} - \text{LB})/\text{UB} \leq \epsilon$ **then**

6: Stop and return $\boldsymbol{z}^*$ as an optimal solution.

7: **else**

8: Update $\xi_{\mathrm{E}}^{t+1} = \tilde{\xi}_{\mathrm{E}}^*$, $\xi_{\mathrm{N}}^{t+1} = \tilde{\xi}_{\mathrm{N}}^*$, $\mathcal{T} = \mathcal{T} \cup \{t+1\}$ and $t = t+1$.

9: Go to step 2 and solve the updated master problem.

10: **end if**

Algorithm 4.1: CCG for 2-stage DRO with half-circle support set

4.3.2 Polyhedral support set

We consider a polyhedral support set, i.e., a polytope Ξ^N with N extreme points. When the polytope Ξ^N is a subset (resp. superset) of the support set (see Figure 4.4), it lower (resp. upper) bounds the optimal value of problem (4.7). In this section, we solve problem (4.7) with the polytope Ξ^N defined as $\text{conv}\{(\hat{\xi}_{\mathrm{E}}^1, \hat{\xi}_{\mathrm{N}}^1), \ldots, (\hat{\xi}_{\mathrm{E}}^N, \hat{\xi}_{\mathrm{N}}^N)\}$. Since Ξ^N is also a convex set, strong duality (proposition 12) holds. Thus, problem (4.7) with support set Ξ^N is also solvable using the CCG algorithm (see Algorithm 4.2). Note that the CCG algorithm is guaranteed to converge in a finite number of iterations since the number of extreme points of Ξ^N is finite.

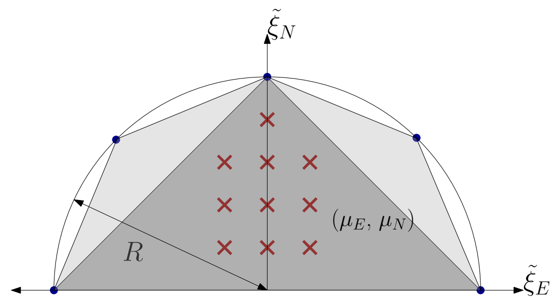


Figure 4.4: Polyhedral support set with 3 and 5 extreme points.

1: Set LB$=-\infty$, UB$=\infty$, $t = 0$ and $\mathcal{T} = \emptyset$.
2: Update LB by solving the following master problem:

$$\text{LB} = \min_{\boldsymbol{z} \in \mathcal{F}} \boldsymbol{q}^{\mathrm{T}} \boldsymbol{z} + \mu_{\mathrm{E}} \lambda_{\mathrm{E}} + \mu_{\mathrm{N}} \lambda_{\mathrm{N}} + \eta \tag{4.13a}$$

$$\text{s.t. } \eta \geq \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}^{\ell} - \lambda_{\mathrm{E}} \xi_{\mathrm{E}}^{\ell} - \lambda_{\mathrm{N}} \xi_{\mathrm{N}}^{\ell}, \ \forall \ell \in \mathcal{T}, \tag{4.13b}$$

$$\boldsymbol{A}\boldsymbol{x}^{\ell} + \boldsymbol{B}(\tilde{\boldsymbol{\xi}}^{\ell})\boldsymbol{z} \geq \boldsymbol{d}, \ \forall \ell \in \mathcal{T}. \tag{4.13c}$$

– Record an optimal solution $\boldsymbol{z}^*$, λ_{E}^*, λ_{N}^*, and η^*.
3: Solve the following problem:

$$\max_{\tilde{\boldsymbol{\xi}} \in \Xi^N} \left\{ \mathcal{Q}(\boldsymbol{z}^*, \tilde{\boldsymbol{\xi}}) - \lambda_{\mathrm{E}}^* \tilde{\xi}_{\mathrm{E}} - \lambda_{\mathrm{N}}^* \tilde{\xi}_{\mathrm{N}} \right\}$$

– Record an optimal $\tilde{\boldsymbol{\xi}}^*$ and the optimal value $\mathcal{Z}(\boldsymbol{z}^*, \lambda_{\mathrm{E}}^*, \lambda_{\mathrm{N}}^*, \tilde{\boldsymbol{\xi}}^*)$
4: Update UB by

$$\text{UB} = \min\{\text{UB}, \ \boldsymbol{q}^{\mathrm{T}} \boldsymbol{z}^* + \mu_{\mathrm{E}} \lambda_{\mathrm{E}}^* + \mu_{\mathrm{N}} \lambda_{\mathrm{N}}^* + \mathcal{Z}(\boldsymbol{z}^*, \lambda_{\mathrm{E}}^*, \lambda_{\mathrm{N}}^*, \tilde{\boldsymbol{\xi}}^*)\}$$

5: **if** $(\text{UB} - \text{LB})/\text{UB} \leq \epsilon$ **then**
6: Stop and return $\boldsymbol{z}^*$ as an optimal solution.
7: **else**
8: Update $\xi_{\mathrm{E}}^{t+1} = \tilde{\xi}_{\mathrm{E}}^*$, $\xi_{\mathrm{N}}^{t+1} = \tilde{\xi}_{\mathrm{N}}^*$, $\mathcal{T} = \mathcal{T} \cup \{t+1\}$ and $t = t+1$.
9: Go to step 2 and solve the updated master problem.
10: **end if**

Algorithm 4.2: CCG for 2-stage DRO with polyhedral support set

4.3.3 Triangle support set

As a special case, we further approximate the convex support set with a polytope that has *three* extreme points, i.e., Ξ^3 (see Figure 4.4). Once again, the CCG algorithm (Algorithm (4.2)) can be used to solve model (4.7). However, with a triangle support set, we derive an exact monolithic reformulation as discussed below. This reformulation is solved efficiently using off-the-shelf commercial solvers such as CPLEX or Gurobi. In the numerical results section, we demonstrate the computational efficacy of this exact reformulation in detail.

Proposition 14. *For fixed $z \in \mathcal{F}$, the worst-case expected value $\sup_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})]$ with Ξ^3 is equivalent to*

$$\max_{p\in\mathbb{R}^3_+} \sum_{k=1}^{3} \mathcal{Q}(\boldsymbol{z}, \hat{\boldsymbol{\xi}}^k) p_k \tag{4.14a}$$

$$s.t. \sum_{k=1}^{3} \hat{\xi}^k_E p_k = \mu_E, \ \sum_{k=1}^{3} \hat{\xi}^k_N p_k = \mu_N, \ \sum_{k=1}^{3} p_k = 1. \tag{4.14b}$$

where $\{\hat{\boldsymbol{\xi}}^1, \hat{\boldsymbol{\xi}}^2, \hat{\boldsymbol{\xi}}^3\}$ are the extreme points of the set Ξ^3.

Proof. For fixed $z \in \mathcal{F}$, the function $\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})$ is convex in $\tilde{\boldsymbol{\xi}}$. Let $\mathcal{H}$ be a convex hull of $\{(\tilde{\boldsymbol{\xi}}, y) : \tilde{\boldsymbol{\xi}} \in \Xi^3, \ y = \mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})\}$. Since taking expectation can be viewed as a convex combination, it follows that $(\boldsymbol{\mu}, \mathcal{Q}(\boldsymbol{z}, \boldsymbol{\mu})) \in \mathcal{H}$. Since the set $\mathcal{D}$ is a mean-support ambiguity set where the support set Ξ^3 is a simplex with 3 extreme points $\hat{\boldsymbol{\xi}}^1, \hat{\boldsymbol{\xi}}^2, \hat{\boldsymbol{\xi}}^3$, there is an unique convex combination of the extreme points which yields $\boldsymbol{\mu}$. Therefore, we have $\sup_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] = \sup\{y | (\boldsymbol{\mu}, y) \in \mathcal{H}\}$, where $\boldsymbol{\mu}$ is a convex combination of the 3 extreme points, $\boldsymbol{\mu} = \hat{\boldsymbol{\xi}}^1 p_1 + \hat{\boldsymbol{\xi}}^2 p_2 + \hat{\boldsymbol{\xi}}^3 p_3$, and $y = \mathcal{Q}(\boldsymbol{z}, \hat{\boldsymbol{\xi}}^1) p_1 + \mathcal{Q}(\boldsymbol{z}, \hat{\boldsymbol{\xi}}^2) p_2 + \mathcal{Q}(\boldsymbol{z}, \hat{\boldsymbol{\xi}}^3) p_3$. $\square$

Remark 5. *The optimal solution (p_1^*, p_2^*, p_3^*) to problem (4.14) is uniquely determined due to the unique convex combination of the extreme points which yields $\boldsymbol{\mu}$. In other words, we solve the following linear system of equations:*

$$\begin{bmatrix} \hat{\xi}^1_E & \hat{\xi}^2_E & \hat{\xi}^3_E \\ \hat{\xi}^1_N & \hat{\xi}^2_N & \hat{\xi}^3_N \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \mu_E \\ \mu_N \\ 1 \end{bmatrix}$$

As long as $\boldsymbol{\mu} \in \Xi^3$, we have $p_1^*, p_2^*, p_3^* \geq 0$.

Proposition 15. *Problem (4.7) with triangle support set (Ξ^3) is equivalent to the following* *stochastic program:*

$$\min_{\boldsymbol{z}\in\mathcal{F}} \boldsymbol{q}^{T}\boldsymbol{z} + \sum_{k=1}^{3} p_k^{*}(\boldsymbol{c}^{T}\boldsymbol{x}^{k}) \tag{4.15a}$$

$$\textit{s.t. } \boldsymbol{A}\boldsymbol{x}^{k} \geq \boldsymbol{d} - \boldsymbol{B}(\hat{\boldsymbol{\xi}}^{k})\boldsymbol{z},\ \forall k \in \{1,2,3\}. \tag{4.15b}$$

4.3.4 Modified CCG algorithm

Problem (4.15) is a monolithic MISOCP which can be solved by the off-the-shelf commercial softwares, and its optimal solution can be used as a warm start for the CCG algorithm that solves two-stage DRO model with a polyhedral support set described in Section 4.3.2, if the polyhedral support set contains the triangle support set. One way to incorporate the traingle support set during the process of the CCG algorithm under the polyhedral set is to utilize a dual formulation of problem (4.14) and add the resulting dual constraints to the relaxed master problem. Specifically, the dual formulation of problem (4.14) can be written as

$$\begin{aligned} &\min \lambda_{\text{E}}\mu_{\text{E}} + \lambda_{\text{N}}\mu_{\text{N}} + \eta \\ &\text{s.t. } \hat{\xi}^{k}_{\text{E}}\lambda_{\text{E}} + \hat{\xi}^{k}_{\text{N}}\lambda_{\text{N}} + \eta \geq \mathcal{Q}(\boldsymbol{z}, \hat{\boldsymbol{\xi}}^{k}),\ \forall k \in [3]. \end{aligned}$$

where $\lambda_{\text{E}}, \lambda_{\text{N}}, \eta \in \mathbb{R}$ are dual variables for each constraint in (4.14b). We conclude this section by providing a modified CCG algorithm (see Algorithm 4.3) that solves two-stage DRO model under a polyhedral support set with N extreme points based on utilizing the monolithic reformulation.

4.4 Numerical Experiments

In this section, we conduct numerical experiments based on the transmission grid instances including the epri-21 [155] and the uiuc-150 systems [156], which are specifically designed for GMD studies. Computations were performed with the HPC resources at Los Alamos National Laboratory with Intel Xeon CPU E5-2660v3, 2.60GHz and 120GB of memory. Optimization models were solved using Gurobi v8.1.0 and were implemented in C++.

4.4.1 Instance description

Most parameters in our models can be obtained from the aforementioned literature [155], [156]. The remaining parameters include $\kappa^{\text{l}} = 50,000[\$/pu]$, $\kappa^{\text{s}} = 100,000[\$/pu]$, $\overline{v}^{\text{d}} =$

1: Set LB=$-\infty$, UB=∞, $t = 3$ and $\mathcal{T} = \emptyset$.

2: Update LB by solving the following problem:

$$\begin{aligned} \text{LB} = \min_{\boldsymbol{z} \in \mathcal{F}} \ & \boldsymbol{q}^{\mathrm{T}} \boldsymbol{z} + \mu_{\mathrm{E}} \lambda_{\mathrm{E}} + \mu_{\mathrm{N}} \lambda_{\mathrm{N}} + \eta && \text{(4.16a)} \\ \text{s.t.} \ & \hat{\xi}^k_{\mathrm{E}} \lambda_{\mathrm{E}} + \hat{\xi}^k_{\mathrm{N}} \lambda_{\mathrm{N}} + \eta \geq \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}^k, \ \forall k \in [3], && \text{(4.16b)} \\ & \boldsymbol{A} \boldsymbol{x}^k + \boldsymbol{B}(\hat{\xi}^k) \boldsymbol{z} \geq \boldsymbol{d}, \ \forall k \in [3], && \text{(4.16c)} \\ & \eta \geq \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}^\ell - \lambda_{\mathrm{E}} \xi^\ell_{\mathrm{E}} - \lambda_{\mathrm{N}} \xi^\ell_{\mathrm{N}}, \ \forall \ell \in \mathcal{T}, && \text{(4.16d)} \\ & \boldsymbol{A} \boldsymbol{x}^\ell + \boldsymbol{B}(\tilde{\boldsymbol{\xi}}^\ell) \boldsymbol{z} \geq \boldsymbol{d}, \ \forall \ell \in \mathcal{T}. && \text{(4.16e)} \end{aligned}$$

– Record an optimal solution $\boldsymbol{z}^*$, λ^*_{E}, λ^*_{N}, and η^*.

3: Solve the following problem:

$$\max_{\tilde{\boldsymbol{\xi}} \in \Xi^N \setminus \Xi^3} \left\{ \mathcal{Q}(\boldsymbol{z}^*, \tilde{\boldsymbol{\xi}}) - \lambda^*_{\mathrm{E}} \tilde{\xi}_{\mathrm{E}} - \lambda^*_{\mathrm{N}} \tilde{\xi}_{\mathrm{N}} \right\}$$

– Record an optimal $\tilde{\boldsymbol{\xi}}^*$ and the optimal value $\mathcal{Z}(\boldsymbol{z}^*, \lambda^*_{\mathrm{E}}, \lambda^*_{\mathrm{N}}, \tilde{\boldsymbol{\xi}}^*)$

4: Update UB by

$$\text{UB} = \min\{\text{UB}, \ \boldsymbol{q}^{\mathrm{T}} \boldsymbol{z}^* + \mu_{\mathrm{E}} \lambda^*_{\mathrm{E}} + \mu_{\mathrm{N}} \lambda^*_{\mathrm{N}} + \mathcal{Z}(\boldsymbol{z}^*, \lambda^*_{\mathrm{E}}, \lambda^*_{\mathrm{N}}, \tilde{\boldsymbol{\xi}}^*)\}$$

5: **if** $(\text{UB} - \text{LB})/\text{UB} \leq \epsilon$ **then**

6: Stop and return $\boldsymbol{z}^*$ as an optimal solution.

7: **else**

8: Update $\xi^{t+1}_{\mathrm{E}} = \tilde{\xi}^*_{\mathrm{E}}$, $\xi^{t+1}_{\mathrm{N}} = \tilde{\xi}^*_{\mathrm{N}}$, $\mathcal{T} = \mathcal{T} \cup \{t+1\}$ and $t = t + 1$.

9: Go to step 2 and solve the updated master problem.

10: **end if**

Algorithm 4.3: Modified CCG for 2-stage DRO with polyhedral support set

$10,000$ [V], and $\overline{I}^{\mathrm{eff}}_e = 2\frac{\overline{s}_e}{\min(\underline{v}_i, \underline{v}_j)}$[Amp], $\forall e_{ij} \in \mathcal{E}^\tau$.

For each system, we generate instances which vary the mean values of $(\mu_{\mathrm{E}}, \mu_{\mathrm{N}})$ and the worst magnitude of the geo-electric field, R, (all units are in [V/km]). We consider *ten* different instances that vary the mean values $(\mu_{\mathrm{E}}, \mu_{\mathrm{N}})$ in proportion to R (Table 4.3).

For each instance, we construct *four* different polyhedral support sets: (1) triangle (Ξ^3_{I}) and (2) pentagon (Ξ^5_{I}) that inner-approximates the convex support set and (3) triangle (Ξ^3_{O}) and (4) hexagon (Ξ^6_{O}) that outer-approximates the convex support set as depicted in Figure 4.5. The objective function values of prob. (4.7) with these sets are non-increasing in the order: Ξ^3_{O}, Ξ^6_{O}, Ξ, Ξ^5_{I}, and Ξ^3_{I}.

Instance #	$(\mu_{\text{E}}, \mu_{\text{N}})$ [V/km]	Instance #	$(\mu_{\text{E}}, \mu_{\text{N}})$ [V/km]
1	(0, R/5)	6	(R/5, 2R/5)
2	(0, 2R/5)	7	(R/5, 3R/5)
3	(0, 3R/5)	8	(-R/5, R/5)
4	(0, 4R/5)	9	(-R/5, 2R/5)
5	(R/5, R/5)	10	(-R/5, 3R/5)

Table 4.3: Ten different $(\mu_{\text{E}}, \mu_{\text{N}})$ depending on R values.

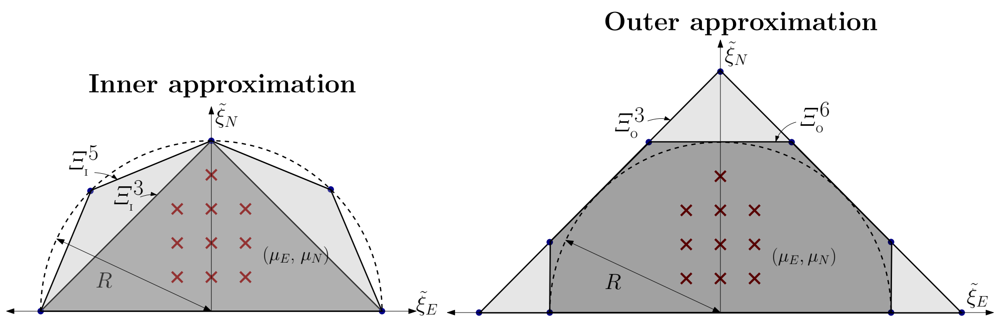


Figure 4.5: Polyhedral support sets that inner- (left) and outer- (right) approximate the nonlinear uncertainty set Ξ.

4.4.2 The epri-21 system

Figure 4.6 shows a simplified diagram of the epri-21 system geo-located near Atlanta, GA. This system has 19 buses, 7 generators, 15 transmission lines, 16 transformers, and 8 substations. In the diagram, the blue lines are 500kV and the green lines are 345kV (see [155] for details).

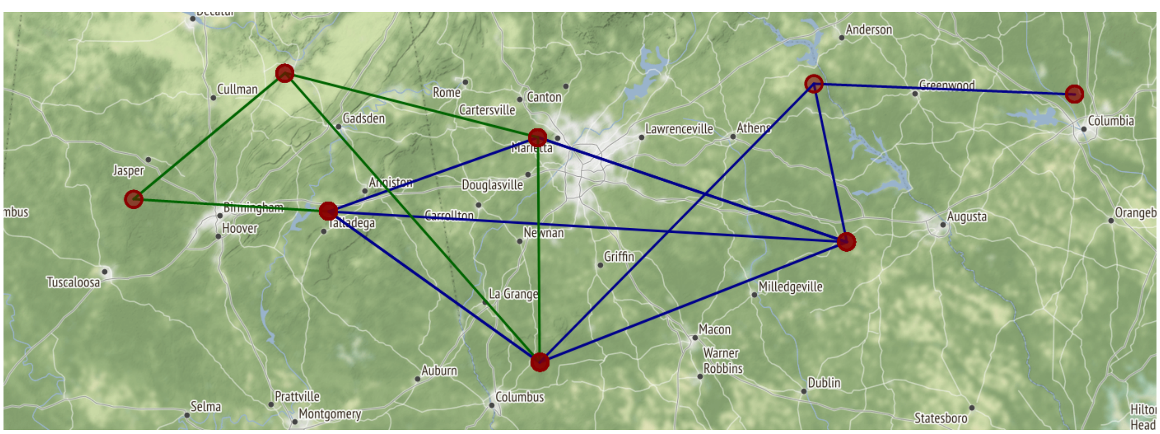


Figure 4.6: The epri-21 system

4.4.2.1 Quality of uncertainty set approximation

Since support set Ξ_{I}^3 has the lowest objective function value for the 4 polyhedral sets, we normalize all the objective function values with respect to that of Ξ_{I}^3. Figure 4.7 shows the normalized objective function values of the DRO model for the 4 polyhedral sets on the 10 instances that have $R = 5$ (left) and $R = 15$ (right). With support sets Ξ_{I}^3 and Ξ_{O}^3, the optimality gaps are less than 0.02% and 0.49% for $R = 5$ and $R = 15$, respectively. With support sets Ξ_{I}^5 and Ξ_{O}^6, the optimality gaps are reduced to 0.01% and 0.03% when $R = 5$ and $R = 15$, respectively. Based on these optimality gaps for the epri-21 system, we observe that the triangle support set is a very good approximation of the convex support set.

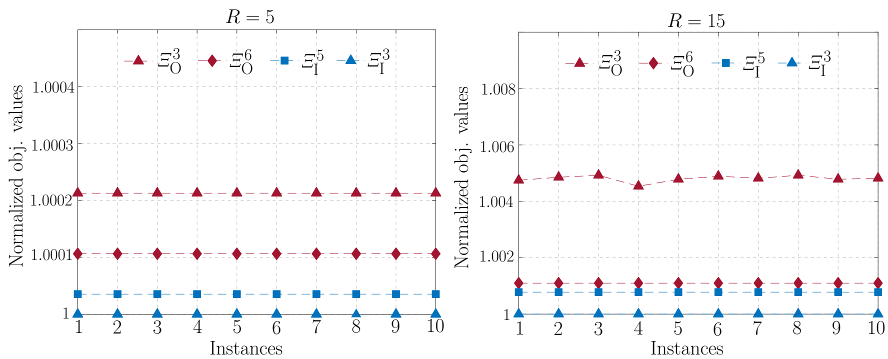


Figure 4.7: Normalized objective function values for $R = 5, 15$.

4.4.2.2 Computational performances

For $R \in \{5, 7.5, 10, 12.5, 15\}$, we construct 10 instances with mean values described in Table 4.3. Figure 4.8 compares the exact monolithic reformulation (Section 4.3.3) with the CCG algorithm (Algorithm 4.2) under the triangle support set Ξ^3. All the points located below the 45 degree blue line are instances where the monolithic reformulation computationally outperforms the CCG algorithm. For larger uncertainty sets ($R = 15$), the monolithic reformulation is on an average *2.1* times faster than CCG. The tightness of the triangle-set approximation (see Section 4.4.2.1) and the computational efficacy of the exact reformulation provides evidence that the monolithic approach is very effective at solving the DRO.

Moreover, under the pentagon support set Ξ^5, we compare the computational time of the modified CCG algorithm (Section 4.3.4) with that of CCG algorithm (Algorithm 4.2). Figure 4.9 shows that the modified CCG algorithm lead to the speed-up over the CCG algorithm for most instances.

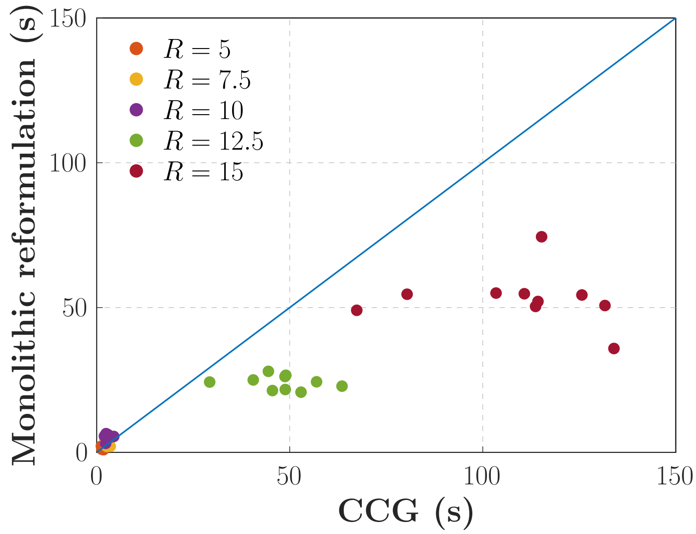


Figure 4.8: Computational times (sec.) of monolithic reformulation and CCG algorithm under the triangle support set (Ξ^3).

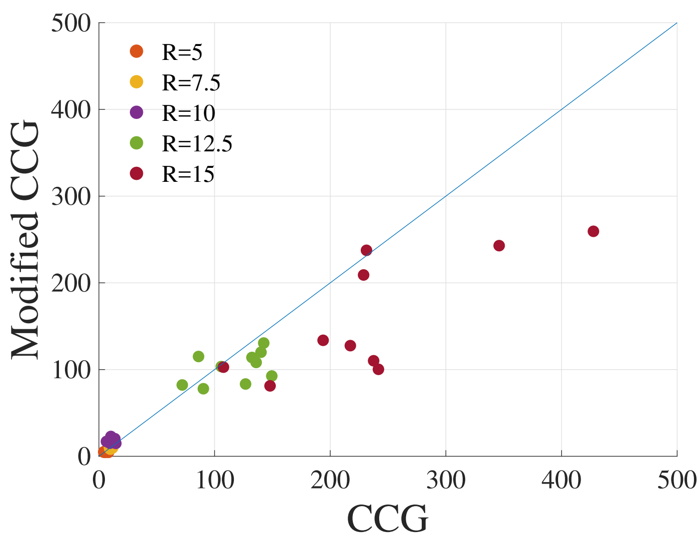


Figure 4.9: Computational times (sec.) of modified CCG and CCG algorithm under the pentagon support set (Ξ^5).

4.4.3 Planning mitigation solutions for uncertain GMD

In this section, we consider how the solutions change when the mean and magnitude of the GMD are varied. Here, the mean values are $(\mu_{\text{E}}, \mu_{\text{N}}) = (5, 4)$ and $R \in \{0, 10, 20, 30, 40\}$. $R = 0$ indicates a deterministic model with $(\tilde{\xi}_{\text{E}}, \tilde{\xi}_{\text{N}}) = (\mu_{\text{E}}, \mu_{\text{N}}) = (5, 4)$.

In Figure 4.10, the optimal objective values (top) and mitigation actions (bottom) change as the size of the uncertainty set (R) increases. As expected, as R increases, more generators, lines, and transformers are turned off to mitigate the effect of GMD and address

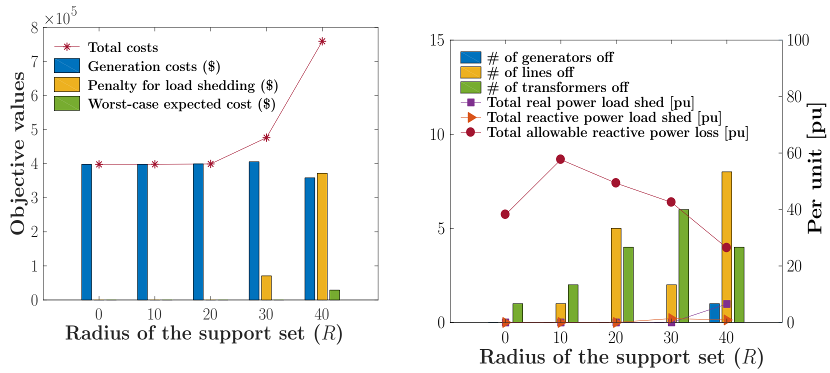


Figure 4.10: Optimal objectives and solutions for various R values.

a larger number of worst-case scenarios. The negative impacts of GMD can be mitigated without shedding loads (only using line and transformer switching) for $R \leq 20$. However, when $R \geq 10$, the total allowable reactive power losses ((4.5)'s solution, bottom figure) decreases, indicating that the topological control actions can mitigate these losses, thus potentially reducing the need for expensive blocking devices used in the network. When $R \geq 30$, the topological control actions are not sufficient to handle the uncertainty in the GMD and some real and reactive power loads are shed. This is reflected in the increase in the total cost of the objective value at $R = 40$.

4.4.4 The uiuc-150 system

Figures 4.11 shows one-line diagram of the uiuc-150 system, which is built from the public load/generation data of the Tennessee region and a statistical analysis of real power systems. This system has 150 buses, 27 generators, 157 transmission lines, 61 transformers, and 98 sub-stations, and in the diagram, the blue lines are 500kV and the green lines are 230kV (See [156] for more details).

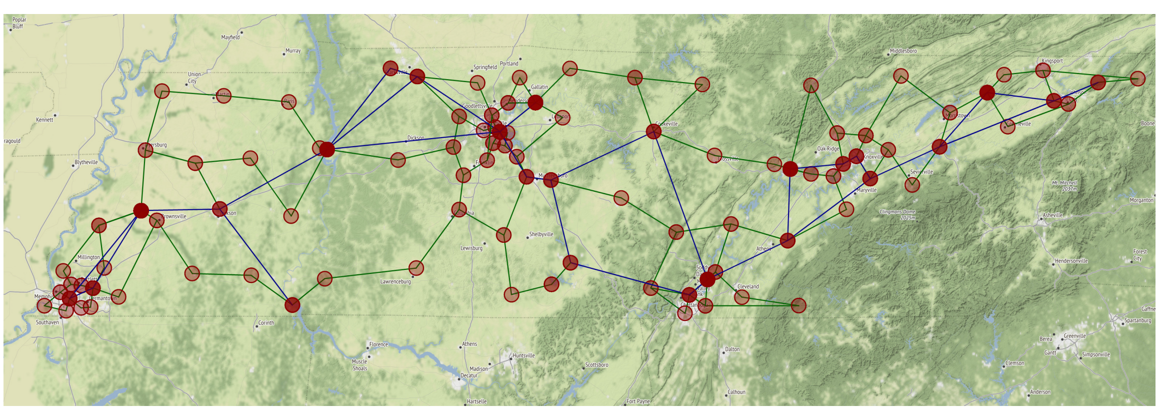


Figure 4.11: The uiuc-150 system

4.4.4.1 Computational performances

For $R = 5$, we construct 7 instances with mean values described in Table 4.4. We compare the exact monolithic reformulation (Section 4.3.3) with the CCG algorithm (Algorithm 4.2) under the triangle support set Ξ^3.

Table 4.4: Comparison of computation time [sec.] under triangle support set

$(\mu_{\text{E}}, \mu_{\text{N}})$	CCG	Mono
$(0, 1)$	Time-out	2983.82
$(0, 2)$	Time-out	3138.12
$(0, 3)$	3678.33	2732.51
$(0, 4)$	3013.87	3099.52
$(1, 1)$	2146.11	2417.11
$(1, 2)$	8534.70	2775.91
$(1, 3)$	4621.94	2752.12

In most instances, the monolithic reformulation outperforms the CCG algorithm, however, we observe that it is computationally challenging mainly due to the long-tail of the optimality gaps. Most instances reached 0.01% of optimality gaps within 1 hour, but closing this gap usually takes much time. We will discuss how we can improve this computationally challenges in Section 4.5.

4.5 Concluding Remarks

In this chapter, we developed a novel two-stage DRO model which uses control of transmission lines, generators, and transformers to mitigate potential negative impacts of uncertain GMD. This model minimizes the expected total cost of mitigation for the worst-case distribution in a convex support set of the GMD's uncertainty. Given this convex support set and mean values for the uncertain GMD, our DRO model is solvable using the CCG algorithm. However, there are no guarantees of finite time convergence. Instead, we approximated the support set with a polytope with N extreme points that allows the CCG to terminate with a finite number of iterations ($O(N)$). We further reformulated the two-stage DRO model into a monolithic MISOCP for the special case when the support set contains three extreme points. We numerically showed the run-time efficacy of this reformulation. Finally, we provided a detailed case study on epri-21 system which analyzed the effects of modeling uncertain GMD.

There are a number of interesting future directions for this work. For example, given the tightness of the triangle support set approximation and the computational efficacy of the exact reformulation, this approach could be used to warm-start the CCG algorithm and speed up the convergence in cases with N extreme points. Second, the approximation could be tightened further by considering different choices of the extreme points for the triangle. Finally, it will be important to scale the DRO to cases with 100's or even 1000's of nodes.

CHAPTER 5

Conclusions

In this dissertation, we tailor the ambiguity sets for three different applications, and derive efficient solution approaches for the three two-stage DRO models, respectively. In Chapter 2, we study a Wasserstein-based DRO model for an appointment scheduling problem under uncertain no-show behaviors of the patients. Based on the various reformulation techniques, we develop efficient solution approaches and show that the solutions obtained by the proposed approach have good out-of-sample performances. In Chapter 3, we consider a decision-dependent moment-based DRO model for a nurse staffing problem under uncertain nurse absenteeism. This works shows the importance of incorporating uncertain nurse absenteeism into staffing decision. Moreover, we observe that the nurse absenteeism play important roles in designing pool structure in a hospital. Finally, in Chapter 4, we propose a DRO model with the *Markov* ambiguity set which decides how to operate the transmission grids under uncertain GMD event. This work shows that controlling transmission lines/transformer/generator can effectively mitigate the future damage due to GMD.

Possible future research directions include:

1. Extending the proposed models in a setting of the sequential decision making under uncertainty, where decisions are made sequentially as the uncertain parameters unfold over time,

2. Developing solution approaches that take advantages of the available data that consists of the past problem instances and their corresponding solutions, which can be done by incorporating either machine learning or randomization policies, and

3. Applying the proposed models and their corresponding solution approaches for a wider range of applications in energy systems, healthcare operation, transportation, and statistical learning.

APPENDIX A

Appendix for Chapter 2

Proof of Lemma 1

We review (a part of) Theorem 2 in [90] that is directly related to our discussion.

Theorem 14 (Adapted from Theorem 2 in [90])**.** *Let* $\mathbb{P} \in \mathcal{P}(\mathbb{R}^n)$ *and* $p > 0$*, and assume that there exist* $\alpha > p$ *and* $\gamma > 0$ *such that* $\int_{\mathbb{R}^n} \exp\{\gamma \|\boldsymbol{u}\|_p^\alpha\} d\mathbb{P}(\boldsymbol{u}) < \infty$*. Then for all* $N \geq 1$ *and* $\epsilon \in (0, \infty)$*,*

$$\mathbb{P}^N \left\{ d_p(\mathbb{P}, \widehat{\mathbb{P}}^N) \geq \epsilon^{1/p} \right\} \leq a(N, \epsilon)\mathbf{1}_{\{\epsilon \leq 1\}} + b(N, \epsilon),$$

where

$$a(N, \epsilon) = C \begin{cases} \exp\{-cN\epsilon^2\} & \text{if } p > n/2 \\ \exp\{-cN(\epsilon/\log(2 + 1/\epsilon))^2\} & \text{if } p = n/2 \\ \exp\{-cN\epsilon^{n/p}\} & \text{if } p \in [1, n/2) \end{cases}$$

$$b(N, \epsilon) = C \exp\{-cN\epsilon^{\alpha/p}\}\mathbf{1}_{\{\epsilon > 1\}}.$$

The positive constants C *and* c *depend only on* p*,* n*,* α*, and* γ*.*

Proof of Lemma 1. First, for any $\alpha > p$ and $\gamma > 0$, $\mathbb{E}_{\mathbb{P}_{\boldsymbol{u}}}[\exp\{\gamma \|\boldsymbol{u}\|_p^\alpha\}] < \infty$ because the support set $\mathcal{U}$ is bounded by Assumption 1. It follows from Theorem 14 that

$$\mathbb{P}_{\boldsymbol{u}}^N \left\{ d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}_{\boldsymbol{u}}^N) \geq \epsilon^{1/p} \right\} \leq a(N, \epsilon)\mathbf{1}_{\{\epsilon \leq 1\}} + b(N, \epsilon), \text{ for all } N \geq 1 \text{ and } \epsilon > 0.$$

Second, we discuss the following two cases based on the value of ϵ.

(1) If $0 < \epsilon \leq 1$, then $0 < \epsilon\left[\log(2 + 1/\epsilon)\right]^2 \geq [\log(3)]^2$. It follows that $(\epsilon/\log(2 + 1/\epsilon))^2 \geq \epsilon^3/[\log(3)]^2$ and $\exp\{-cN(\epsilon/\log(2 + 1/\epsilon))^2\} \leq \exp\{-cN\epsilon^3/[\log(3)]^2\}$.

Then,

$$a(N,\epsilon) \le C \begin{cases} \exp\{-cN\epsilon^2\} & \text{if } p > n/2 \\ \exp\{-cN\epsilon^3/[\log(3)]^2\} & \text{if } p = n/2 \\ \exp\{-cN\epsilon^{n/p}\} & \text{if } p \in [1, n/2) \end{cases}$$
$$\le \; C\exp\left\{-\frac{c}{[\log(3)]^2} N\epsilon^{\max\{3,n/p\}}\right\},$$

where the second inequality is because $\log(3) > 1$ and ϵ^x decreases as x increases with $\epsilon \in (0,1]$.

(2) If $\epsilon > 1$, then we set $\alpha = \max\{3p, n\} > p$. It follows that $b(N,\epsilon) \le C\exp\{-cN\epsilon^{\max\{3,n/p\}}\}$.

Summarizing the above two cases and letting $c_1 = C$ and $c_2 = c/[\log(3)]^2$, we have

$$\mathbb{P}_{\boldsymbol{u}}^N\left\{d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}_{\boldsymbol{u}}^N) \ge \epsilon^{1/p}\right\} \le c_1\exp\left\{-c_2 N\epsilon^{\max\{3,n/p\}}\right\}$$

for all $N \ge 1$ and $\epsilon > 0$. Equating the right-hand side of the above inequality to β and solving for $\epsilon^{1/p}$ yields $\mathbb{P}_{\boldsymbol{u}}^N\{d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}_{\boldsymbol{u}}^N) \ge \epsilon_N(\beta)\} \le \beta$, where $\epsilon_N(\beta)$ is defined in the statement of Lemma 1. This completes the proof. □

Proof of Theorem 1

We first prove the following three lemmas.

Lemma 7. *The function $f(\boldsymbol{s},\boldsymbol{u})$ is bounded on $\mathcal{S} \times \mathcal{U}$. Additionally, $f(\boldsymbol{s},\boldsymbol{u})$ is jointly convex in $\boldsymbol{s}$ and $\boldsymbol{u}$.*

Proof. First, by the dual formulation of liner program (2.2), we represent

$$f(\boldsymbol{s},\boldsymbol{u}) = \max_{\boldsymbol{y}\in\mathcal{Y}} (\boldsymbol{u}-\boldsymbol{s})^\top \boldsymbol{y}$$

, where $\mathcal{Y}$ is defined in (2.5). It follows that $f(\boldsymbol{s},\boldsymbol{u})$ is bounded on $\mathcal{S} \times \mathcal{U}$ because (i) $\mathcal{Y}$ is nonempty and bounded by Lemma 2, (ii) $\mathcal{S}$ is bounded by definition, and (iii) $\mathcal{U}$ is bounded by Assumption 1.

Second, $f(\boldsymbol{s},\boldsymbol{u})$ is jointly convex in $\boldsymbol{s}$ and $\boldsymbol{u}$ because it is represented as the maximum of linear functions of $\boldsymbol{s}$ and $\boldsymbol{u}$. □

Lemma 8. *Let $p \geq 1$. Then, the function $f(\boldsymbol{s}, \boldsymbol{u})$ is continuous on $\mathcal{S} \times \mathcal{U}$. Additionally, for any fixed $\boldsymbol{s} \in \mathcal{S}$, $f(\boldsymbol{s}, \boldsymbol{u})$ is Lipschitz continuous in $\boldsymbol{u}$ with Lipschitz constant $L := \max_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|_q < \infty$, where q is such that $\frac{1}{p} + \frac{1}{q} = 1$.*

Proof. First, as $f(\boldsymbol{s}, \boldsymbol{u}) \equiv \max_{\boldsymbol{y} \in \mathcal{Y}} (\boldsymbol{u} - \boldsymbol{s})^\top \boldsymbol{y}$ and $\mathcal{Y}$ is polyhedral, nonempty, and bounded (see (2.5) and Lemma 2), $f(\boldsymbol{s}, \boldsymbol{u})$ can be represented as the maximum of a *finite* number of linear functions of $\boldsymbol{s}$ and $\boldsymbol{u}$. It follows that $f(\boldsymbol{s}, \boldsymbol{u})$ is continuous on $\mathcal{S} \times \mathcal{U}$.

Second, pick any $\boldsymbol{s} \in \mathcal{S}$ and $\boldsymbol{u}^1$, $\boldsymbol{u}^2 \in \mathcal{U}$. Then,

$$
\begin{aligned}
f(\boldsymbol{s}, \boldsymbol{u}^1) - f(\boldsymbol{s}, \boldsymbol{u}^2) &= \max_{\boldsymbol{y} \in \mathcal{Y}} (\boldsymbol{u}^1 - \boldsymbol{s})^\top \boldsymbol{y} - \max_{\boldsymbol{y} \in \mathcal{Y}} (\boldsymbol{u}^2 - \boldsymbol{s})^\top \boldsymbol{y} \\
&\leq \max_{\boldsymbol{y} \in \mathcal{Y}} \left\{ (\boldsymbol{u}^1 - \boldsymbol{s})^\top \boldsymbol{y} - (\boldsymbol{u}^2 - \boldsymbol{s})^\top \boldsymbol{y} \right\} \\
&= \max_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{y}^\top (\boldsymbol{u}^1 - \boldsymbol{u}^2) \\
&\leq \max_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{y}\|_q \left\| \boldsymbol{u}^1 - \boldsymbol{u}^2 \right\|_p \\
&= L \left\| \boldsymbol{u}^1 - \boldsymbol{u}^2 \right\|_p,
\end{aligned}
$$

where the second inequality follows from the Hölder's inequality. We note that $L < \infty$ because $\mathcal{Y}$ is bounded and $\|\boldsymbol{y}\|_q$ is continuous in $\boldsymbol{y}$. Similarly, we can show that $f(\boldsymbol{s}, \boldsymbol{u}^2) - f(\boldsymbol{s}, \boldsymbol{u}^1) \leq L \|\boldsymbol{u}^1 - \boldsymbol{u}^2\|_p$. Hence, $|f(\boldsymbol{s}, \boldsymbol{u}^1) - f(\boldsymbol{s}, \boldsymbol{u}^2)| \leq L \|\boldsymbol{u}^1 - \boldsymbol{u}^2\|_p$ for any $\boldsymbol{s} \in \mathcal{S}$ and $\boldsymbol{u}^1$, $\boldsymbol{u}^2 \in \mathcal{U}$, which completes the proof. □

Lemma 9 (Adapted from Lemma 3.7 in [11])**.** *Let $p \geq 1$. Consider a sequence of confidence levels $\{\beta_N\}_{N \in \mathbb{N}}$ such that $\sum_{N=1}^{\infty} \beta_N < \infty$ and $\lim_{N \to \infty} \epsilon_N(\beta_N) = 0$. Additionally, let $\{\widehat{\mathbb{Q}}_{\boldsymbol{u}}^N\}_{N \in \mathbb{N}}$ represent a sequence of probability distributions with each $\widehat{\mathbb{Q}}_{\boldsymbol{u}}^N \in \mathcal{D}_p(\widehat{\mathbb{P}}_{\boldsymbol{u}}^N, \epsilon_N(\beta_N))$. Then, $\lim_{N \to \infty} d_1(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) = 0$ $\mathbb{P}_{\boldsymbol{u}}^\infty$-almost surely.*

Proof. First, as $\widehat{\mathbb{Q}}_{\boldsymbol{u}}^N \in \mathcal{D}_p(\widehat{\mathbb{P}}_{\boldsymbol{u}}^N, \epsilon_N(\beta_N))$, by the triangular inequality we have

$$
d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) \leq d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}_{\boldsymbol{u}}^N) + d_p(\widehat{\mathbb{P}}_{\boldsymbol{u}}^N, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) \leq d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}_{\boldsymbol{u}}^N) + \epsilon_N(\beta_N).
$$

By Lemma 1, we have $\mathbb{P}_{\boldsymbol{u}}^N \left\{ d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{P}}_{\boldsymbol{u}}^N) \leq \epsilon_N(\beta_N) \right\} \geq 1 - \beta_N$ and so

$$
\mathbb{P}_{\boldsymbol{u}}^N \left\{ d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) \leq 2\epsilon_N(\beta_N) \right\} \geq 1 - \beta_N.
$$

But as $\sum_{N=1}^{\infty} \beta_N < \infty$, the Borel-Cantelli Lemma implies that

$$
\mathbb{P}_{\boldsymbol{u}}^\infty \{ d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) \leq 2\epsilon_N(\beta_N) \text{ eventually} \} = 1.
$$

In addition, as $\lim_{N\to\infty} \epsilon_N(\beta_N) = 0$, we have $\lim_{N\to\infty} d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) = 0$ $\mathbb{P}_{\boldsymbol{u}}^\infty$-almost surely.

Second, by the Jensen's inequality we have

$$\begin{aligned} d_p(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) &= \left(\inf_{\Pi \in \mathcal{P}(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N)} \mathbb{E}_\Pi \left[\|\boldsymbol{u}_{\mathbb{P}} - \boldsymbol{u}_{\mathbb{Q}}\|_p^p \right] \right)^{1/p} \\ &\geq \left\{ \left(\inf_{\Pi \in \mathcal{P}(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N)} \mathbb{E}_\Pi \left[\|\boldsymbol{u}_{\mathbb{P}} - \boldsymbol{u}_{\mathbb{Q}}\|_p^1 \right] \right)^p \right\}^{1/p} \\ &= d_1(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}^N). \end{aligned}$$

It follows that $\lim_{N\to\infty} d_1(\mathbb{P}_{\boldsymbol{u}}, \widehat{\mathbb{Q}}_{\boldsymbol{u}}^N) = 0$ $\mathbb{P}_{\boldsymbol{u}}^\infty$-almost surely. □

Theorem 1 follows from Lemmas 7–9.

Proof of Theorem 1. By Assumption 1 and Lemmas 7–9, all conditions of Theorem 3.6 in [11] are satisfied. Therefore, the conclusions of Theorem 1 hold valid. □

Proof of Theorem 2

Proof. By Assumption 1 and Lemma 1, all conditions of Theorem 3.5 in [11] are satisfied. Therefore, the conclusion of Theorem 2 holds valid. □

Proof of Lemma 2

Proof. First, $\mathbf{0} \in \mathcal{Y}$ because $\boldsymbol{c} \geq \mathbf{0}$, $\boldsymbol{d} \geq \mathbf{0}$, and $C \geq 0$. It follows that $\mathcal{Y}$ is nonempty. Next, $\mathcal{Y}$ is closed as it is polyhedral. Finally, $-d_n \leq y_n \leq C$ implies that $y_{n-1} \leq y_n + c_n \leq C + c_n$ and so $-d_{n-1} \leq y_{n-1} \leq C + c_n$. Similarly, we have $-d_i \leq y_i \leq C + \sum_{j=i+1}^n c_j$ for all $i \in [n-2]$. This implies that $\mathcal{Y}$ is bounded and completes the proof. □

Proof of Proposition 1

Proof. First, by the definition of Wasserstein distance, $\mathbb{Q}_{\boldsymbol{u}} \in \mathcal{D}_p(\widehat{\mathbb{P}}_{\boldsymbol{u}}^N, \epsilon)$ implies that there exists a joint distribution Π of $(\boldsymbol{u}, \widehat{\boldsymbol{u}})$ with marginals $\mathbb{Q}_{\boldsymbol{u}}$ and $\widehat{\mathbb{P}}_{\boldsymbol{u}}^N$, such that $\mathbb{E}_\Pi \left[\|\boldsymbol{u} - \widehat{\boldsymbol{u}}\|_p^p \right] \leq \epsilon^p$. As $\widehat{\mathbb{P}}_{\boldsymbol{u}}^N = \frac{1}{N} \sum_{j=1}^N \delta_{\widehat{\boldsymbol{u}}^j}$, there exist conditional distributions $\{\mathbb{Q}_{\boldsymbol{u}}^j\}_{j\in[N]}$ such that $\Pi = \frac{1}{N} \sum_{j=1}^N \mathbb{Q}_{\boldsymbol{u}}^j$, where each $\mathbb{Q}_{\boldsymbol{u}}^j$ represents the distribution of $\boldsymbol{u}$

conditional on that $\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{u}}^j$. It follows that problem (2.6) can be recast as

$$\begin{aligned} \widehat{Z}(\boldsymbol{s}) \ := \ \sup \quad & \frac{1}{N}\sum_{j=1}^{N}\int_{\mathcal{U}} f(\boldsymbol{s},\, \boldsymbol{u})\, \mathbb{Q}^j_{\boldsymbol{u}}(d\boldsymbol{u}) \\ \text{s.t.} \quad & \frac{1}{N}\sum_{j=1}^{N}\int_{\mathcal{U}} \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p\, \mathbb{Q}^j_{\boldsymbol{u}}(d\boldsymbol{u}) \le \epsilon^p \\ & \mathbb{Q}^j_{\boldsymbol{u}} \in \mathcal{P}(\mathcal{U}) \quad \forall\, j \in [N]. \end{aligned} \tag{A.1}$$

Second, by a standard duality argument and letting $\rho \ge 0$ be the Lagrangian dual multiplier, we write the dual of (A.1) as:

$$\widehat{Z}(\boldsymbol{s}) \ = \inf_{\rho\ge 0} \sup_{\mathbb{Q}^j_{\boldsymbol{u}}\in\mathcal{P}(\mathcal{U})} \epsilon^p\rho + \frac{1}{N}\sum_{j=1}^{N}\int_{\mathcal{U}} \left(f(\boldsymbol{s},\, \boldsymbol{u}) - \rho\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p\right)\, \mathbb{Q}^j_{\boldsymbol{u}}(d\boldsymbol{u}) \tag{A.2a}$$

$$= \inf_{\rho\ge 0} \epsilon^p\rho + \frac{1}{N}\sum_{j=1}^{N}\sup_{\boldsymbol{u}\in\mathcal{U}} \left\{f(\boldsymbol{s},\, \boldsymbol{u}) - \rho\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p\right\}, \tag{A.2b}$$

where equality (A.2a) follows from the strong duality result for the moment problems (see Proposition 3.4 in [157], Theorem 1 in [25], and Lemma 7 in [158]) and equality (A.2b) follows from the fact that $\mathcal{P}(\mathcal{U})$ contains all the Dirac distributions supported on $\mathcal{U}$. Finally, reformulation (2.7) is obtained by introducing an auxiliary variable θ_j to represent each supremum in (A.2b). □

Proof of Theorem 3

We first establish three technical lemmas. Consider the following quadratic program:

$$\begin{aligned} \sup \quad & \boldsymbol{x}^\top \boldsymbol{Q}_0 \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{e}_1^\top \boldsymbol{x} = 1,\ \boldsymbol{x} \in \mathcal{K} \\ & \boldsymbol{x}^\top \boldsymbol{Q}_i \boldsymbol{x} = 0 \quad \forall\, i \in [n], \end{aligned} \tag{A.3}$$

where $\boldsymbol{Q}_i \in \mathbb{S}^k\ \forall\, i = 0, \ldots, n$. Its copositive programming relaxation (see [91, 159]) reads:

$$\begin{aligned} \sup \quad & \boldsymbol{Q}_0 \bullet \boldsymbol{X} \\ \text{s.t.} \quad & \boldsymbol{e}_1\boldsymbol{e}_1^\top \bullet \boldsymbol{X} = 1,\ \boldsymbol{X} \in \mathcal{CP}(\mathcal{K}) \\ & \boldsymbol{Q}_i \bullet \boldsymbol{X} = 0 \quad \forall\, i \in [n], \end{aligned} \tag{A.4}$$

where $\mathcal{K}$ is a nonempty, closed and convex cone. Defining $\mathcal{L} := \left\{\boldsymbol{x} \in \mathcal{K} : \boldsymbol{e}_1^\top \boldsymbol{x} = 1\right\}$, we review the following result.

Lemma 10 ([159], Corollary 8.4, Theorem 8.3)**.** *Suppose that $\mathcal{L}$ is nonempty and bounded. Then, (A.4) is equivalent to (A.3), i.e., i) the optimal value of (A.4) is equal to that of (A.3); ii) if $\boldsymbol{X}^\star$ is an optimal solution for (A.4), then $\boldsymbol{X}^\star \mathrm{e}_1$ is in the convex hull of optimal solutions for (A.3).*

By the conic programming duality, the dual of (A.4) is the following linear program over the cone of copositive matrices with respect to $\mathcal{K}$:

$$\begin{aligned} \inf \quad & \beta_j \\ \text{s.t.} \quad & \beta_j \, \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \sum_{i=1}^{n} \psi_i \boldsymbol{Q}_i - \boldsymbol{Q}_0 \in \mathcal{COP}(\mathcal{K}) \\ & \beta_j \in \mathbb{R}, \, \boldsymbol{\psi} \in \mathbb{R}^n. \end{aligned} \tag{A.5}$$

We now showcase that strong duality holds between (A.4) and (A.5) in the lemma below.

Lemma 11. *Suppose that $\mathcal{L}$ is noempty and bounded. Then, strong duality holds between (A.4) and (A.5), i.e., the optimal value of* (A.4) *equals that of* (A.5).

Proof. We prove the statement by showing that the dual problem (A.5) admits a Slater point. To this end, we seek for a scalar β_j such that the following relation holds:

$$\big(t, \bar{\boldsymbol{x}}^\top\big)\Big(\beta_j \boldsymbol{e}_1 \boldsymbol{e}_1^\top - \boldsymbol{Q}_0\Big)\big(t, \bar{\boldsymbol{x}}^\top\big)^\top = \beta_j t^2 - \big(t, \bar{\boldsymbol{x}}^\top\big)\boldsymbol{Q}_0\big(t, \bar{\boldsymbol{x}}^\top\big)^\top > 0 \tag{A.6}$$

for all non-zero vector $(t, \bar{\boldsymbol{x}}) \in \mathcal{K}$.

We first show that $t = 0 \implies \bar{\boldsymbol{x}} = 0$. We prove the statement by using the contradiction argument. Suppose that $t = 0$ but $\bar{\boldsymbol{x}} \neq 0$. Choose $(1, \bar{\boldsymbol{x}}') \in \mathcal{K}$. Then for any non-negative scalar $\lambda \geq 0$, we have $\boldsymbol{z}(\lambda) := (1, \bar{\boldsymbol{x}}') + \lambda(0, \bar{\boldsymbol{x}}) \in \mathcal{K}$ because $\mathcal{K}$ is a closed and convex cone. As λ can be arbitrarily large while $\bar{\boldsymbol{x}} \neq 0$, we conclude that $\mathcal{L}$ is unbounded, contradicting the statement.

Therefore, it suffices to consider the case of $t > 0$. We then can divide the expression in (A.6) by t^2, which requires us to show the following equivalent relation:

$$\beta_j - \big(1, (\bar{\boldsymbol{x}}/t)^\top\big)\boldsymbol{Q}_0\big(1, (\bar{\boldsymbol{x}}/t)^\top\big)^\top > 0$$

for all $(t, \bar{\boldsymbol{x}}) \in \mathcal{K}$ and $t > 0$. Then, the boundedness of $\mathcal{L}$ implies that there exists a constant $\beta_j^\star$ such that $\beta_j^\star > \big(1, (\bar{\boldsymbol{x}}/t)^\top\big)\boldsymbol{Q}_0\big(1, (\bar{\boldsymbol{x}}/t)^\top\big)^\top$ for all $\big(1, (\bar{\boldsymbol{x}}/t)^\top\big) \in \mathcal{L}$. The claim thus follows since the point $\beta_j = \beta_j^\star$ constitutes a Slater point for the problem (A.5). □

Proof of Theorem 3. Using the result from Proposition 1, for any given $\boldsymbol{s} \in \mathcal{S}$, we can compute the worst-case expectation value by solving a standard robust optimization problem shown in (2.7), where each of the semi-infinite constraints entails a non-convex program. Then, for any fixed $(\boldsymbol{s}, \rho, \theta_j) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$, we consider the j-th constraint separately:

$$\sup_{\boldsymbol{u} \in \mathcal{U}} \left(f(\boldsymbol{s},\, \boldsymbol{u}) - \rho \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p \right) \leq \theta_j. \tag{A.7}$$

First if $p = 1$, then the maximization problem on the left-hand side of (A.7) can be reformulated as

$$\begin{array}{ll} \sup & (\boldsymbol{u}^+ - \boldsymbol{u}^- + \widehat{\boldsymbol{u}}^j - \boldsymbol{s})^\top \boldsymbol{y} - \rho \boldsymbol{e}^\top (\boldsymbol{u}^+ + \boldsymbol{u}^-) \\ \text{s.t.} & (\boldsymbol{u}^+,\, \boldsymbol{u}^-,\, \boldsymbol{y}) \in \mathcal{F}_1^j. \end{array} \tag{A.8}$$

Letting $\boldsymbol{x} := [t, (\boldsymbol{u}^+)^\top, (\boldsymbol{u}^-)^\top, \boldsymbol{y}^\top]^\top \in \mathbb{R}^{3n+1}$, we rewrite (A.8) as

$$\begin{array}{ll} \sup & \boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) \bullet \boldsymbol{x}\boldsymbol{x}^\top \\ \text{s.t.} & \boldsymbol{e}_1\boldsymbol{e}_1^\top \bullet \boldsymbol{x}\boldsymbol{x}^\top = 1 \\ & \boldsymbol{x} \in \mathcal{K}_1^j. \end{array} \tag{A.9}$$

By Lemma 10, (A.9) can be reformulated as a linear program over the cone of completely positive matrices:

$$\begin{array}{ll} \sup & \boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) \bullet \boldsymbol{X} \\ \text{s.t.} & \boldsymbol{e}_1\boldsymbol{e}_1^\top \bullet \boldsymbol{X} = 1 \\ & \boldsymbol{X} \in \mathcal{CP}(\mathcal{K}_1^j). \end{array} \tag{A.10}$$

If Assumption 1 holds, then the set $\{\boldsymbol{x} \in \mathcal{K}_1^j : \boldsymbol{e}_1^\top \boldsymbol{x} = 1\}$ is nonempty and bounded. Therefore by Lemma 11, the optimal value of (A.10) is equal to that of its dual problem, shown as:

$$\begin{array}{ll} \inf & \beta_j \\ \text{s.t.} & \beta_j \in \mathbb{R},\; \beta_j\, \boldsymbol{e}_1\boldsymbol{e}_1^\top - \boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}_1^j). \end{array} \tag{A.11}$$

Then, the constraint (A.7) is satisfied if and only if there exists θ_j such that

$$\begin{array}{l} \beta_j \leq \theta_j \\ \beta_j\, \boldsymbol{e}_1\boldsymbol{e}_1^\top - \boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}_1^j). \end{array} \tag{A.12}$$

Using the same argument for all N constraints yields the finite constraint system

$$\begin{array}{l} \beta_j \leq \theta_j \quad \forall\, j \in [N] \\ \beta_j\, \boldsymbol{e}_1\boldsymbol{e}_1^\top - \boldsymbol{H}_j^1(\rho,\, \boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}_1^j) \quad \forall\, j \in [N]. \end{array} \tag{A.13}$$

Replacing the semi-finite constraints in (2.7) with the constraint system in (A.13), removing the variables θ_j, and making $\boldsymbol{s}$ as the decision variables, we end up with a copositive programming reformulation (2.10) for (W-DRAS).

Second if $p = 2$, then the maximization problem on the left-hand side of (A.7) can be written as

$$\begin{aligned} \sup \quad & (\boldsymbol{u} - \boldsymbol{s})^\top \boldsymbol{y} - \rho \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_2^2 \\ \text{s.t.} \quad & (\boldsymbol{u},\, \boldsymbol{y}) \in \mathcal{F}_2. \end{aligned} \tag{A.14}$$

Letting $\boldsymbol{x} := (t,\, \boldsymbol{u}^\top,\, \boldsymbol{y}^\top)^\top$, we rewrite (A.14) as

$$\begin{aligned} \sup \quad & \boldsymbol{H}_j^2(\rho,\, \boldsymbol{s}) \bullet \boldsymbol{x}\boldsymbol{x}^\top \\ \text{s.t.} \quad & \boldsymbol{e}_1\boldsymbol{e}_1^\top \bullet \boldsymbol{x}\boldsymbol{x}^\top = 1 \\ & \boldsymbol{x} \in \mathcal{K}_2^j. \end{aligned} \tag{A.15}$$

Using a similar argument, we can reformulate (W-DRAS) to the copositive program in (2.11) for the case of $p = 2$. □

Proof of Proposition 2

Proof. Given $\rho \geq 0$ and $\boldsymbol{s} \in \mathcal{S}$, the problem for computing $\omega_j(\rho, \boldsymbol{s})$ can be rewritten as follows:

$$\begin{aligned} \omega_j(\rho, \boldsymbol{s}) \; &= \sup_{\boldsymbol{u} \in \mathcal{U}} \left\{ \sup_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \sum_{i=1}^n (u_i - s_i) y_i \right\} - \rho \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p \right\} && \text{(A.16a)} \\ &= \sup_{\boldsymbol{u} \in \mathcal{U}} \sup_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \sum_{i=1}^n (u_i - s_i) y_i - \rho \sum_{i=1}^n |u_i - \widehat{u}_i^j|^p \right\} && \text{(A.16b)} \\ &= \sup_{\boldsymbol{y} \in \mathcal{Y}} \sup_{\boldsymbol{u} \in \mathcal{U}} \left\{ \sum_{i=1}^n \Big[(u_i - s_i) y_i - \rho |u_i - \hat{u}_i^j|^p \Big] \right\} && \text{(A.16c)} \\ &= \sup_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \sum_{i=1}^n \left[-s_i y_i + \sup_{u_i^{\mathrm{L}} \leq u_i \leq u_i^{\mathrm{U}}} \left\{ y_i u_i - \rho |u_i - \hat{u}_i^j|^p \right\} \right] \right\}. && \text{(A.16d)} \end{aligned}$$

In the maximization problem (A.16d), the objective function is convex in variables $\boldsymbol{y}$ because $p \geq 1$. Hence, it suffices to consider the extreme points of polytope $\mathcal{Y}$ to obtain $\omega_j(\rho, \boldsymbol{s})$. In what follows, we apply a similar exposition to the proof of Proposition 2 in [42]. In particular, we show that each extreme point of $\mathcal{Y}$ corresponds to a partition of the set $\{1, \ldots, n+1\}$ into intervals in the form of $[k, \ell]_{\mathbb{Z}}$ for some $k, \ell \in [n+1]$. To this end, we remark that for any extreme point $\boldsymbol{y}$ of $\mathcal{Y}$, either $y_n = -d_n$ or $y_n = C$ should hold (see [160, 161]). Moreover, for $i = 2, \ldots, n$, we have $(y_{i-1} + d_{i-1})(y_i - y_{i-1} + c_i) = 0$, which indicates that either $y_{i-1} = -d_{i-1}$ or $y_{i-1} = y_i + c_i$ for all $i = 2, \ldots, n$. In the latter case, the value of y_{i-1} is uniquely determined by y_i. Recursively applying this fact, we obtain the following findings. For any $i \in [n]$, let $\ell \in [i, n+1]_{\mathbb{Z}}$ represent the smallest index such that $y_\ell = -d_\ell$ (for notation convenience, we let $c_{n+1} := 0$, $d_{n+1} := -C$, and $y_{n+1} := C$, so that $y_{n+1} = -d_{n+1}$), then $y_i = y_\ell + \sum_{q=i+1}^{\ell} c_q$, i.e., $y_i = \pi_{i\ell}$, which is defined in (2.13d).

This gives rise to a one-to-one correspondence between an extreme point $\boldsymbol{y}$ and a partition of $\{1, \ldots, n+1\}$ into intervals in the form of $[k, \ell]_{\mathbb{Z}}$, and $y_i = \pi_{i\ell}$ if $i \in [k, \ell]_{\mathbb{Z}}$ for some $k, \ell \in [n+1]$. Therefore, formulation (A.16d) is equivalent to finding an optimal partition of the set $\{1, \ldots, n+1\}$. To this end, for any $k \leq \ell$, we define a binary variable $t_{k\ell}$ such that $t_{k\ell} = 1$ if any only if $[k, \ell]_{\mathbb{Z}}$ is a component of the partition of $\{1, \ldots, n+1\}$. The set $\big\{t_{k\ell} \in \{0, 1\}, \forall\, k \in [n+1], \forall\, \ell \in [k, n+1]_{\mathbb{Z}}\big\}$ represents a partition of $\{1, \ldots, n+1\}$ if and only if

$$\sum_{k=1}^{i} \sum_{\ell=i}^{n+1} t_{k\ell} = 1 \quad \forall\, i \in [n+1].$$

Recall that $z_{(n+1)(n+1)j} = 0$ and $z_{i\ell j} = -s_i \pi_{i\ell} + \sup_{u_i^{\mathrm{L}} \leq u_i \leq u_i^{\mathrm{U}}} \{\pi_{i\ell} u_i - \rho |u_i - \hat{u}_i^j|^p\}$. Then, for all $k \in [n+1]$ and $\ell \in [k, n+1]_{\mathbb{Z}}$, we have when $t_{k\ell} = 1$

$$\sum_{i=k}^{\ell} \Big[-s_i y_i + \sup_{u_i^{\mathrm{L}} \leq u_i \leq u_i^{\mathrm{U}}} \big\{y_i u_i - \rho |u_i - \hat{u}_i^j|^p\big\}\Big] = \sum_{i=k}^{\ell} z_{i\ell j}.$$

Hence, we can obtain $\omega_j(\rho, \boldsymbol{s})$ by solving the following binary integer program:

$$\begin{aligned} \omega_j(\rho, \boldsymbol{s}) \;=\; \max_t \;\; & \sum_{k=1}^{n+1} \sum_{\ell=k}^{n+1} \left(\sum_{i=k}^{\ell} z_{i\ell j} \right) t_{k\ell} \\ \text{s.t.} \;\; & \sum_{k=1}^{i} \sum_{\ell=i}^{n+1} t_{k\ell} = 1 \;\; \forall\, i \in [n+1] \\ & t_{k\ell} \in \{0, 1\} \;\; \forall\, k \in [n+1], \; \forall\, \ell \in [k, n+1]_{\mathbb{Z}}. \end{aligned}$$

Furthermore, we note that the constraint matrix of this formulation is totally unimodular (see, e.g., [162]). Hence, without loss of optimality we relax the binary restrictions to $t_{k\ell} \geq 0$ for all $k \in [n+1]$ and $\ell \in [k, n+1]_{\mathbb{Z}}$. In addition, as $z_{(n+1)(n+1)j} = 0$, we can drop decision variable $t_{(n+1)(n+1)j}$ to obtain

$$\omega_j(\rho, \boldsymbol{s}) \;=\; \max_t \;\; \sum_{k=1}^{n} \sum_{\ell=k}^{n+1} \left(\sum_{i=k}^{\min\{\ell, n\}} z_{i\ell j} \right) t_{k\ell}$$

$$\text{s.t.} \quad \sum_{k=1}^{i} \sum_{\ell=i}^{n+1} t_{k\ell} = 1 \ \ \forall\, i \in [n] \tag{A.17a}$$

$$\sum_{k=1}^{n} t_{k(n+1)} \leq 1 \tag{A.17b}$$

$$t_{k\ell} \geq 0 \ \ \forall\, k \in [n],\ \forall\, \ell \in [k, n+1]_{\mathbb{Z}}.$$

But letting $i = n$ in constraint (A.17a) yields $1 = \sum_{k=1}^{n} \sum_{\ell=n}^{n+1} t_{k\ell} = \sum_{k=1}^{n} t_{kn} + \sum_{k=1}^{n} t_{k(n+1)} \geq \sum_{k=1}^{n} t_{k(n+1)}$, which implies constraint (A.17b). Dropping the redundant constraint (A.17b) leads to the linear programming reformulation (2.13a)–(2.13c) of (W-DRAS). □

Proof of Theorem 4 and Extension to the Rational $p > 1$

Proof. First, if $p = 1$ then

$$\begin{aligned} z_{i\ell j} \ &= \ -s_i \pi_{i\ell} + \sup_{u_i^{\text{L}} \leq u_i \leq u_i^{\text{U}}} \{\pi_{i\ell} u_i - \rho |u_i - \widehat{u}_i^j|\} \\ &= \ \max\{(u_i^{\text{L}} - s_i)\pi_{i\ell} - \rho|u_i^{\text{L}} - \widehat{u}_i^j|,\ (\widehat{u}_i^j - s_i)\pi_{i\ell},\ (u_i^{\text{U}} - s_i)\pi_{i\ell} - \rho|u_i^{\text{U}} - \widehat{u}_i^j|\} \end{aligned}$$

for all $j \in [N]$, $i \in [n]$, and $\ell \in [i, n+1]_{\mathbb{Z}}$. Taking the dual of the linear program (2.13a)–(2.13c) in Proposition 2 yields

$$\begin{aligned} \omega_j(\rho, \boldsymbol{s}) \ = \ \min_{\boldsymbol{\gamma}, \boldsymbol{z}} \quad & \sum_{i=1}^{n} \gamma_{ij} \\ \text{s.t.} \quad & \sum_{k=i}^{\min\{\ell,n\}} \gamma_{kj} \geq \sum_{k=i}^{\min\{\ell,n\}} z_{k\ell j} && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}} \\ & z_{i\ell j} \geq (u_i^{\text{L}} - s_i)\pi_{i\ell} - |u_i^{\text{L}} - \widehat{u}_i^j|\rho && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}} \\ & z_{i\ell j} \geq (\widehat{u}_i^j - s_i)\pi_{i\ell} && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}} \\ & z_{i\ell j} \geq (u_i^{\text{U}} - s_i)\pi_{i\ell} - |u_i^{\text{U}} - \widehat{u}_i^j|\rho && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}}, \end{aligned}$$

where dual variables γ_{ij} are associated with primal constraints (2.13b). The proof is completed by substituting $\omega_j(\rho, \boldsymbol{s}) = \sup_{\boldsymbol{u} \in \mathcal{U}} \{f(\boldsymbol{s},\, \boldsymbol{u}) - \rho \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_1^1\}$ in formulation (2.12).

Second, if $p = 2$ then

$$\begin{aligned} z_{i\ell j} \ &= \ -s_i \pi_{i\ell} + \sup_{u_i^{\text{L}} \leq u_i \leq u_i^{\text{U}}} \left\{\pi_{i\ell} u_i - \rho(u_i - \widehat{u}_i^j)^2\right\} \\ &= \ -s_i \pi_{i\ell} + \inf_{\beta_{i\ell j}^{\text{L}}, \beta_{i\ell j}^{\text{U}} \geq 0} \ \sup_{u_i} \left\{\pi_{i\ell} u_i - \rho(u_i - \widehat{u}_i^j)^2 + \beta_{i\ell j}^{\text{L}}(u_i - u_i^{\text{L}}) + \beta_{i\ell j}^{\text{U}}(u_i^{\text{U}} - u_i)\right\} \end{aligned}$$

$$
\begin{aligned}
&= -s_i\pi_{i\ell} + \inf_{\beta^{\mathrm{L}}_{i\ell j},\beta^{\mathrm{U}}_{i\ell j}\geq 0}\ \sup_{u_i}\Big\{u^{\mathrm{U}}_i\beta^{\mathrm{U}}_{i\ell j} - u^{\mathrm{L}}_i\beta^{\mathrm{L}}_{i\ell j} + (\pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j})\widehat{u}^j_i \\
&\quad + (\pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j})(u_i - \widehat{u}^j_i) - \rho(u_i - \widehat{u}^j_i)^2\Big\} \\
&= -s_i\pi_{i\ell} + \inf_{\beta^{\mathrm{L}}_{i\ell j},\beta^{\mathrm{U}}_{i\ell j}\geq 0}\left\{u^{\mathrm{U}}_i\beta^{\mathrm{U}}_{i\ell j} - u^{\mathrm{L}}_i\beta^{\mathrm{L}}_{i\ell j} + (\pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j})\widehat{u}^j_i + \frac{(\pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j})^2}{4\rho}\right\}
\end{aligned}
$$

for all $j \in [N]$, $i \in [n]$, and $\ell \in [i, n+1]_{\mathbb{Z}}$. Taking the dual of the linear program (2.13a)–(2.13c) in Proposition 2 yields

$$
\begin{aligned}
\omega_j(\rho, \boldsymbol{s}) \;=\; \min_{\boldsymbol{\gamma},\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{r}} \quad & \sum_{i=1}^{n}\gamma_{ij} \\
\text{s.t.} \quad & \sum_{k=i}^{\min\{\ell,n\}}\gamma_{kj} \geq \sum_{k=i}^{\min\{\ell,n\}} z_{k\ell j} && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}} \\
& z_{i\ell j} + \pi_{i\ell}s_i - (u^{\mathrm{U}}_i - \widehat{u}^j_i)\beta^{\mathrm{U}}_{i\ell j} \\
& -(\widehat{u}^j_i - u^{\mathrm{L}}_i)\beta^{\mathrm{L}}_{i\ell j} - r_{i\ell j} \geq \pi_{i\ell}\widehat{u}^j_i && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}} \\
& r_{i\ell j} \geq \frac{\left(\pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j}\right)^2}{4\rho} && \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}}.
\end{aligned}
$$

We represent the last constraint in the following second-order conic form:

$$
\left\| \begin{bmatrix} \pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j} \\ r_{i\ell j} - \rho \end{bmatrix} \right\|_2 \leq r_{i\ell j} + \rho \ \ \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}}.
$$

The proof is completed by substituting $\omega_j(\rho, \boldsymbol{s}) = \sup_{\boldsymbol{u}\in\mathcal{U}} \{f(\boldsymbol{s}, \boldsymbol{u}) - \rho\|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_2^2\}$ in formulation (2.12). □

We extend Theorem 4 and derive a second-order cone reformulation of (W-DRAS) for any rational $p > 1$. To this end, we define $q := p/(p-1)$, $q_1, q_2 \in \mathbb{N}$ such that $q = q_1/q_2$ and $q_1 > q_2$, $M := \lceil \log_2 q_1 \rceil$, and $q_3 := 2^M - q_1$. We summarize the reformulation in the following theorem and note that it involves $\mathcal{O}(N^2 n q_1)$ variables and $\mathcal{O}(N^2 n q_1)$ constraints.

Theorem 15. *Suppose that Assumption 2 holds and $p > 1$ is rational. Then,* (W-DRAS) *yields the same optimal value and the same set of optimal solutions as the following second-*

order cone program:

$$
\begin{array}{lll}
\min & \epsilon^p \rho + \dfrac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\gamma_{ij} & \\
\text{s.t.} & \sum_{k=i}^{\min\{\ell,n\}} \gamma_{kj} \geq \sum_{k=i}^{\min\{\ell,n\}} z_{k\ell j} & \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N] \\
& z_{i\ell j} + \pi_{i\ell} s_i - u_i^{U}\beta^{U}_{i\ell j} + u_i^{L}\beta^{L}_{i\ell j} & \\
& -\widehat{u}_i^{j}(\alpha^{U}_{i\ell j} - {}^{L}_{i\ell j}) - (p^{1-q} - p^{-q}) r_{i\ell j} \geq 0 & \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N] \\
& \varphi^{i\ell j}_{1k} = r_{i\ell j} & \forall\, k \in [q_2],\ \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N] \\
& \varphi^{i\ell j}_{1k} = \rho & \forall\, k \in [q_2+1, q_1]_{\mathbb{Z}},\ \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}}, \\
& & \forall\, j \in [N] \\
& \varphi^{i\ell j}_{1k} = \alpha^{U}_{i\ell j} + \alpha^{L}_{i\ell j} & \forall\, k \in [q_1+1, 2^M]_{\mathbb{Z}},\ \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}}, \\
& & \forall\, j \in [N] \\
& \left\| \begin{bmatrix} 2(\alpha^{U}_{i\ell j} + \alpha^{L}_{i\ell j}) \\ \varphi^{i\ell j}_{M1} - \varphi^{i\ell j}_{M2} \end{bmatrix} \right\|_2 \leq \varphi^{i\ell j}_{M1} + \varphi^{i\ell j}_{M2} & \forall\, i \in [n],\ \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N] \\
& \left\| \begin{bmatrix} 2\varphi^{i\ell j}_{(m+1)k} \\ \varphi^{i\ell j}_{m(2k-1)} - \varphi^{i\ell j}_{m(2k)} \end{bmatrix} \right\|_2 \leq \varphi^{i\ell j}_{m(2k-1)} + \varphi^{i\ell j}_{m(2k)} & \forall\, m \in [M-1],\ \forall\, k \in [2^{M-m}],\ \forall\, i \in [n], \\
& & \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N] \\
& \rho \geq 0,\ \boldsymbol{s} \in \mathcal{S},\ \varphi^{i\ell j}_{mk} \geq 0 & \forall\, m \in [M],\ \forall\, k \in [2^{M-m+1}],\ \forall\, i \in [n], \\
& & \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\ \forall\, j \in [N].
\end{array}
$$

Proof. As $p > 1$, we have

$$
\begin{aligned}
z_{i\ell j} &= -s_i\pi_{i\ell} + \sup_{u_i^{\mathrm{L}} \leq u_i \leq u_i^{\mathrm{U}}} \left\{ \pi_{i\ell} u_i - \rho |u_i - \widehat{u}_i^{j}|^p \right\} \\
&= -s_i\pi_{i\ell} + \inf_{\substack{\beta^{\mathrm{L}}_{i\ell j}, \beta^{\mathrm{U}}_{i\ell j} \\ \alpha^{\mathrm{L}}_{i\ell j}, \alpha^{\mathrm{U}}_{i\ell j} \geq 0}} \sup_{u_i, v_i} \Big\{ \pi_{i\ell} u_i - \rho v_i^p + \beta^{\mathrm{U}}_{i\ell j}(u_i^{\mathrm{U}} - u_i) + \beta^{\mathrm{L}}_{i\ell j}(u_i - u_i^{\mathrm{L}}) \\
&\qquad + \alpha^{\mathrm{U}}_{i\ell j}(v_i - u_i + \widehat{u}_i^{j}) + \alpha^{\mathrm{L}}_{i\ell j}(v_i + u_i - \widehat{u}_i^{j}) \Big\} \\
&= -s_i\pi_{i\ell} + \inf_{\substack{\beta^{\mathrm{L}}_{i\ell j}, \beta^{\mathrm{U}}_{i\ell j} \\ \alpha^{\mathrm{L}}_{i\ell j}, \alpha^{\mathrm{U}}_{i\ell j} \geq 0}} \sup_{u_i, v_i} \Big\{ u_i^{\mathrm{U}}\beta^{\mathrm{U}}_{i\ell j} - u_i^{\mathrm{L}}\beta^{\mathrm{L}}_{i\ell j} + \widehat{u}_i^{j}(\alpha^{\mathrm{U}}_{i\ell j} - \alpha^{\mathrm{L}}_{i\ell j}) - \rho v_i^p + (\alpha^{\mathrm{U}}_{i\ell j} + \alpha^{\mathrm{L}}_{i\ell j}) v_i \\
&\qquad + (\pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \beta^{\mathrm{U}}_{i\ell j} - \alpha^{\mathrm{U}}_{i\ell j} + \alpha^{\mathrm{L}}_{i\ell j}) u_i \Big\} \\
&= -s_i\pi_{i\ell} + \inf_{\substack{\beta^{\mathrm{L}}_{i\ell j}, \beta^{\mathrm{U}}_{i\ell j} \\ \alpha^{\mathrm{L}}_{i\ell j}, \alpha^{\mathrm{U}}_{i\ell j} \geq 0}} \Big\{ u_i^{\mathrm{U}}\beta^{\mathrm{U}}_{i\ell j} - u_i^{\mathrm{L}}\beta^{\mathrm{L}}_{i\ell j} + \widehat{u}_i^{j}(\alpha^{\mathrm{U}}_{i\ell j} - \alpha^{\mathrm{L}}_{i\ell j}) + (p^{1-q} - p^{-q})\rho^{1-q}(\alpha^{\mathrm{U}}_{i\ell j} + \alpha^{\mathrm{L}}_{i\ell j})^q \Big\} \\
&\text{s.t. } \pi_{i\ell} + \beta^{\mathrm{L}}_{i\ell j} - \alpha^{\mathrm{U}}_{i\ell j} - \alpha^{\mathrm{U}}_{i\ell j} + \alpha^{\mathrm{L}}_{i\ell j} = 0
\end{aligned}
$$

for all $j \in [N]$, $i \in [n]$, and $\ell \in [i, n+1]_{\mathbb{Z}}$. Taking the dual of the linear program (2.13a)–

(2.13c) in Proposition 2 yields

$$
\begin{aligned}
\omega_j(\rho, \boldsymbol{s}) \;=\; \min \quad & \sum_{i=1}^{n} \gamma_{ij} \\
\text{s.t.} \quad & \sum_{k=i}^{\min\{\ell,n\}} \gamma_{kj} \geq \sum_{k=i}^{\min\{\ell,n\}} z_{k\ell j} && \forall\, i \in [n],\, \forall\, \ell \in [i, n+1]_{\mathbb{Z}} \\
& z_{i\ell j} + \pi_{i\ell} s_i - u_i^{\text{U}} \beta_{i\ell j}^{\text{U}} + u_i^{\text{L}} \beta_{i\ell j}^{\text{L}} \\
& -\widehat{u}_i^j(\alpha_{i\ell j}^{\text{U}} - {}_{i\ell j}^{\text{L}}) - (p^{1-q} - p^{-q}) r_{i\ell j} \geq 0 && \forall\, i \in [n],\, \forall\, \ell \in [i, n+1]_{\mathbb{Z}},\, \forall\, j \in [N] \\
& r_{i\ell j} \geq \rho^{1-q}(\alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}})^q && \forall\, i \in [n],\, \forall\, \ell \in [i, n+1]_{\mathbb{Z}},
\end{aligned}
$$

where dual variables γ_{ij} are associated with primal constraints (2.13b). In addition, we rewrite the last constraint in the above formulation as

$$
\begin{aligned}
& \alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}} \leq r_{i\ell j}^{1/q} \rho^{(q-1)/q} \\
\Leftrightarrow\; & \alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}} \leq \Bigg[\underbrace{r_{i\ell j} \cdots r_{i\ell j}}_{q_2} \underbrace{\rho \cdots \rho}_{q_1 - q_2} \underbrace{(\alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}}) \cdots (\alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}})}_{q_3} \Bigg]^{\frac{1}{2^M}} \\
\Leftrightarrow\; & \alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}} \leq \left(\varphi_{11}^{i\ell j} \cdots \varphi_{1(2^M)}^{i\ell j} \right)^{\frac{1}{2^M}},
\end{aligned}
$$

where $\varphi_{mk}^{i\ell j}$ are defined by $\varphi_{1k}^{i\ell j} := r_{i\ell j}$ for all $k \in [q_2]$, $\varphi_{1k}^{i\ell j} := \rho$ for all $k \in [q_2 + 1, q_1]_{\mathbb{Z}}$, and $\varphi_{1k}^{i\ell j} := \alpha_{i\ell j}^{\text{U}} + \alpha_{i\ell j}^{\text{L}}$ for all $k \in [q_1 + 1, 2^M]_{\mathbb{Z}}$. Based on a seminal work on second-order conic representability (see Example 11 in Section 3.3 of [163]), the last inequality holds if and only if there exist $\{\varphi_{mk}^{i\ell j} \geq 0 : m \in [2, M]_{\mathbb{Z}}, k \in [2^{M+1-m}]\}$ such that $\varphi_{(m+1)k}^{i\ell j} \leq \sqrt{\varphi_{m(2k-1)}^{i\ell j} \varphi_{m(2k)}^{i\ell j}}$ for all $m \in [M-1]$ and $k \in [2^{M-m}]$. This requirement can be represented in the following second-order conic form:

$$
\left\| \begin{bmatrix} 2\varphi_{(m+1)k}^{i\ell j} \\ \varphi_{m(2k-1)}^{i\ell j} - \varphi_{m(2k)}^{i\ell j} \end{bmatrix} \right\|_2 \leq \varphi_{m(2k-1)}^{i\ell j} + \varphi_{m(2k)}^{i\ell j}.
$$

The proof is completed by substituting $\omega_j(\rho, \boldsymbol{s}) = \sup_{\boldsymbol{u} \in \mathcal{U}} \left\{ f(\boldsymbol{s}, \boldsymbol{u}) - \rho \|\boldsymbol{u} - \widehat{\boldsymbol{u}}^j\|_p^p \right\}$ in formulation (2.12). $\square$

Proof of Theorem 5

Proof. First, for fixed $\bar{\boldsymbol{s}} \in \mathcal{S}$ and $\epsilon \geq 0$, the dual problem of formulation (2.14) is

$$
\begin{aligned}
& \sup_{\mathbb{Q}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}},\epsilon)} && \mathbb{E}_{\mathbb{Q}}[f(\bar{\boldsymbol{s}},\boldsymbol{u})] \\
= & \max_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},\boldsymbol{w}} && \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1} \pi_{i\ell}\left[(u^{\mathrm{L}}_i-\bar{s}_i)q_{i\ell j}+(u^{\mathrm{U}}_i-\bar{s}_i)r_{i\ell j}+(\widehat{u}^j_i-\bar{s}_i)w_{i\ell j}\right] \\
& \text{s.t.} && \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1} \left[(\widehat{u}^j_i-u^{\mathrm{L}}_i)q_{i\ell j}+(u^{\mathrm{U}}_i-\widehat{u}^j_i)r_{i\ell j}\right] \leq \epsilon \\
& && \sum_{\ell=i}^{n+1}\sum_{k=1}^{i} p_{k\ell j} = \tfrac{1}{N} \quad \forall\, i\in[n],\ \forall\, j\in[N] \\
& && \sum_{k=1}^{i} p_{k\ell j} = q_{i\ell j}+w_{i\ell j}+r_{i\ell j} \quad \forall\, i\in[n],\ \forall\,\ell\in[i,\,n+1]_{\mathbb{Z}},\ \forall\, j\in[N] \\
& && p_{i\ell j}\geq 0,\ q_{i\ell j}\geq 0,\ r_{i\ell j}\geq 0,\ w_{i\ell j}\geq 0 \quad \forall\, i\in[n],\ \forall\,\ell\in[i,\,n+1]_{\mathbb{Z}},\ \forall\, j\in[N],
\end{aligned}
$$

where dual variables $p_{k\ell j}$, $q_{i\ell j}$, $w_{i\ell j}$, and $r_{i\ell j}$ are associated with the constraints in (2.14). The strong duality holds because the dual feasible region is non-empty (e.g., we can set $p_{1(n+1)j} = 1/N$, $\forall j \in [N]$, all other $p_{k\ell j}$ to be zero, all $q_{i\ell j}$ and $r_{i\ell j}$ to be zero, and all $w_{i\ell j}$ to be $\sum_{k=1}^{i} p_{k\ell j}$). Replacing $\{p_{i\ell j}, q_{i\ell j}, r_{i\ell j}, w_{i\ell j}\}$ with $\frac{1}{N}\{p_{i\ell j}, q_{i\ell j}, r_{i\ell j}, w_{i\ell j}\}$ and substituting variables $w_{i\ell j}$ with $\sum_{k=1}^{i} p_{k\ell j} - q_{i\ell j} - r_{i\ell j}$ yields formulation (2.16).

Second, to see the existence of $\mathbb{P}^j_{\boldsymbol{t}}$ for all $j \in [N]$, we note that the constraint matrix defining set $\mathcal{T}$ is totally unimodular. Hence, $\text{conv}(\mathcal{T}) = \{\boldsymbol{t} \in [0,1]^{(n+1)(n+2)/2} : \sum_{k=1}^{i}\sum_{\ell=i}^{n+1} t_{k\ell} = 1,\ \forall i \in [n+1]\}$ and so $\boldsymbol{p}^\star_j := [p^\star_{k\ell j} : k \in [n+1], \ell \in [k, n+1]_{\mathbb{Z}}] \in \text{conv}(\mathcal{T})$. It follows that there exists a finite number of points $\boldsymbol{t}^1, \ldots, \boldsymbol{t}^I \in \mathcal{T}$ and weights $\lambda^1, \ldots, \lambda^I \in [0,1]$ such that

$$\boldsymbol{p}^\star_j \;=\; \sum_{i=1}^{I} \lambda^i \boldsymbol{t}^i,\ \ \sum_{i=1}^{I} \lambda^i = 1.$$

Hence, the distribution $\mathbb{P}^j_{\boldsymbol{t}}$ constructed by setting $\mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{t}^i\} = \lambda^i$ for all $i \in [I]$ fulfills the claim.

Third, to prove that $\mathbb{Q}^\star_{\boldsymbol{u}} \in \mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)$, we note that $u_{i\ell j} \in [u^{\mathrm{L}}_i, u^{\mathrm{U}}_i]$. Indeed, on the one hand, if $\sum_{k=1}^{i} p^\star_{k\ell j} = 0$ then $q^\star_{i\ell j} = r^\star_{i\ell j} = 0$ by formulation (2.16). Thus, by the extended arithmetics $0/0 = 0$ we have $u^\star_{i\ell j} = \widehat{u}^j_i \in [u^{\mathrm{L}}_i, u^{\mathrm{U}}_i]$. On the other hand, if $\sum_{k=1}^{i} p^\star_{k\ell j} > 0$ then

$$u_{i\ell j} = \left(\frac{\sum_{k=1}^{i} p^\star_{k\ell j} - q^\star_{i\ell j} - r^\star_{i\ell j}}{\sum_{k=1}^{i} p^\star_{k\ell j}}\right)\widehat{u}^j_i + \left(\frac{q^\star_{i\ell j}}{\sum_{k=1}^{i} p^\star_{k\ell j}}\right)u^{\mathrm{L}}_i + \left(\frac{r^\star_{i\ell j}}{\sum_{k=1}^{i} p^\star_{k\ell j}}\right)u^{\mathrm{U}}_i.$$

It follows that $u_{i\ell j}$ is a convex combination of $\widehat{u}_i^j$, u_i^{L}, and u_i^{U} and so $u^\star_{i\ell j} \in [u_i^{\text{L}}, u_i^{\text{U}}]$. In addition, we define a joint probability distribution $\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}}$ of $(\hat{\boldsymbol{u}}, \boldsymbol{u}, \boldsymbol{t})$ through $\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}} = \frac{1}{N}\sum_{j=1}^N \sum_{\boldsymbol{\tau}\in\mathcal{T}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\}\delta_{\hat{\boldsymbol{u}}^j,\boldsymbol{u}^j(\boldsymbol{\tau}),\boldsymbol{\tau}}$. Note that the projection of $\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}}$ on $\boldsymbol{u}$ is $\mathbb{Q}^\star_{\boldsymbol{u}}$. It follows that

$$
\begin{aligned}
d_1(\mathbb{Q}^\star_{\boldsymbol{u}}, \widehat{\mathbb{P}}^N_{\boldsymbol{u}}) &= \inf_{\Pi\in\mathcal{P}(\mathbb{Q}^\star_{\boldsymbol{u}},\widehat{\mathbb{P}}^N_{\boldsymbol{u}})} \mathbb{E}_\Pi\Big[\|\boldsymbol{u} - \widehat{\boldsymbol{u}}\|_1\Big] \\
&\le \mathbb{E}_{\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}}}\big[\|\boldsymbol{u} - \widehat{\boldsymbol{u}}\|_1\big] && \text{(A.18a)} \\
&= \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}}}\big[|u_i - \widehat{u}_i|\big] \\
&= \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \sum_{\boldsymbol{\tau}\in\mathcal{T}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\} \left| \sum_{\ell=i}^{n+1} \left(\sum_{k=1}^i \tau_{k\ell}\right) u_{i\ell j} - \widehat{u}_i^j \right| && \text{(A.18b)} \\
&= \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \sum_{\ell=i}^{n+1} \sum_{k=1}^i \sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}: \\ \tau_{k\ell}=1}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\} \left| u_{i\ell j} - \widehat{u}_i^j \right| && \text{(A.18c)} \\
&= \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \sum_{\ell=i}^{n+1} \left(\sum_{k=1}^i p^\star_{k\ell j}\right) \left| u_{i\ell j} - \widehat{u}_i^j \right| && \text{(A.18d)} \\
&= \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \sum_{\ell=i}^{n+1} \left(\sum_{k=1}^i p^\star_{k\ell j}\right) \left| \frac{q^\star_{i\ell j}(u_i^{\text{L}} - \widehat{u}_i^j)}{\sum_{k=1}^i p^\star_{k\ell j}} + \frac{r^\star_{i\ell j}(u_i^{\text{U}} - \widehat{u}_i^j)}{\sum_{k=1}^i p^\star_{k\ell j}} \right| \\
&\le \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \sum_{\ell=i}^{n+1} \left[q^\star_{i\ell j} \left| u_i^{\text{L}} - \widehat{u}_i^j \right| + r^\star_{i\ell j} \left| u_i^{\text{U}} - \widehat{u}_i^j \right| \right] \\
&= \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^n \sum_{\ell=i}^{n+1} \left[q^\star_{i\ell j} \left(\widehat{u}_i^j - u_i^{\text{L}}\right) + r^\star_{i\ell j} \left(u_i^{\text{U}} - \widehat{u}_i^j\right) \right] \;\le\; \epsilon, && \text{(A.18e)}
\end{aligned}
$$

where inequality (A.18a) is because the projection of $\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}}$ on $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}$ are $\mathbb{Q}^\star_{\boldsymbol{u}}$ and $\widehat{\mathbb{P}}^N_{\boldsymbol{u}}$, respectively, equality (A.18b) follows from the definition of $\mathbb{Q}_{\widehat{\boldsymbol{u}},\boldsymbol{u},\boldsymbol{t}}$, equality (A.18c) is because, for each $\boldsymbol{\tau} \in \mathcal{T}$ and $i \in [n]$, there is one and only one pair of indices $(k, \ell) \in [i] \times [i, n+1]_{\mathbb{Z}}$ such that $\tau_{k\ell} = 1$, equality (A.18d) is because $\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}: \\ \tau_{k\ell}=1}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\} = \mathbb{P}^j_{\boldsymbol{t}}\{t_{k\ell} = 1\} = p^\star_{k\ell j}$, and the inequality in (A.18e) follows from the first constraint in formulation (2.16).

Finally, we have

$$
\begin{aligned}
&\sup_{\mathbb{Q}_{\boldsymbol{u}}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}},\epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}}, \boldsymbol{u})] && \text{(A.19a)} \\
&\ge \mathbb{E}_{\mathbb{Q}^\star_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}}, \boldsymbol{u})] && \text{(A.19b)}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbb{Q}^\star_{\hat{u}}}\left[\max_{\boldsymbol{y}\in\mathcal{Y}}\left\{\sum_{i=1}^{n} y_i(u_i - \bar{s}_i)\right\}\right] \\
&= \mathbb{E}_{\mathbb{Q}^\star_{\hat{u}}}\left[\max_{\boldsymbol{t}\in\mathcal{T}}\left\{\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\sum_{k=1}^{i} t_{k\ell}\pi_{i\ell}(u_i - \bar{s}_i)\right\}\right] && \text{(A.19c)} \\
&\geq \mathbb{E}_{\mathbb{Q}^\star_{\hat{u},\boldsymbol{u},\boldsymbol{t}}}\left[\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\sum_{k=1}^{i} t_{k\ell}\pi_{i\ell}(u_i - \bar{s}_i)\right] && \text{(A.19d)} \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{\boldsymbol{\tau}\in\mathcal{T}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t}=\boldsymbol{\tau}\}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\sum_{k=1}^{i} t_{k\ell}\pi_{i\ell}(u_{i\ell j} - \bar{s}_i) \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\sum_{k=1}^{i}\sum_{\boldsymbol{\tau}\in\mathcal{T}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t}=\boldsymbol{\tau}\} t_{k\ell}\pi_{i\ell}(u_{i\ell j} - \bar{s}_i) \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\sum_{k=1}^{i}\sum_{\substack{\boldsymbol{\tau}\in\mathcal{T}:\\ \tau_{k\ell}=1}} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t}=\boldsymbol{\tau}\}\pi_{i\ell}(u_{i\ell j} - \bar{s}_i) \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\left(\sum_{k=1}^{i} p^\star_{k\ell j}\right)\pi_{i\ell}(u_{i\ell j} - \bar{s}_i) && \text{(A.19e)} \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{\ell=i}^{n+1}\left(\sum_{k=1}^{i} p^\star_{k\ell j}\right)\pi_{i\ell}\left(\widehat{u}^j_i - \bar{s}_i + \frac{q^\star_{i\ell j}(u^{\text{L}}_i - \widehat{u}^j_i)}{\sum_{k=1}^{i} p^\star_{k\ell j}} + \frac{r^\star_{i\ell j}(u^{\text{U}}_i - \widehat{u}^j_i)}{\sum_{k=1}^{i} p^\star_{k\ell j}}\right) \\
&= \sup_{\mathbb{Q}_{\boldsymbol{u}}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}},\epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}},\,\boldsymbol{u})], && \text{(A.19f)}
\end{aligned}
$$

where inequality (A.19b) is because $\mathbb{Q}_{\boldsymbol{u}} \in \mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}}, \epsilon)$, equality (A.19c) is because $\mathcal{T}$ consists of all extreme points of polytope $\mathcal{Y}$, inequality (A.19d) is because we replace the maximization over variables $\boldsymbol{t}$ with a distribution of $\boldsymbol{t}$, equality (A.19e) is because $\sum_{\boldsymbol{\tau}\in\mathcal{T}:\, \tau_{k\ell}=1} \mathbb{P}^j_{\boldsymbol{t}}\{\boldsymbol{t} = \boldsymbol{\tau}\} = \mathbb{P}^j_{\boldsymbol{t}}\{t_{k\ell} = 1\} = p^\star_{k\ell j}$, and equality (A.19f) is because the optimal value of formulation (2.16) equals $\sup_{\mathbb{Q}_{\boldsymbol{u}}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{u}},\epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{u}}}[f(\bar{\boldsymbol{s}}, \boldsymbol{u})]$. This completes the proof. □

Proof of Theorem 6

Proof. If $p = 1$, we rewrite the maximization problem embedded in (2.20) for any $j \in [N]$ as

$$
\begin{aligned}
\sup \quad & (\boldsymbol{\mu} - \boldsymbol{s})^\top \boldsymbol{y} - \rho\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^j\|_1^1 - \rho\|\boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}}^j\|_1^1 \\
\text{s.t.} \quad & (\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \Xi,\ \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{\lambda}),
\end{aligned} \tag{A.20}
$$

where we use the dual formulation (2.19) of $g(\boldsymbol{s}, \boldsymbol{\xi})$ and the fact that $\boldsymbol{\xi} = [\boldsymbol{\mu}^\top, \boldsymbol{\lambda}^\top]^\top$.

Then, (A.20) can be reformulated to

$$\begin{array}{ll} \sup & (\boldsymbol{\mu}^+ - \boldsymbol{\mu}^- + \widehat{\boldsymbol{\mu}}^j - \boldsymbol{s})^\top \boldsymbol{y} - \rho \boldsymbol{e}^\top(\boldsymbol{\mu}^+ + \boldsymbol{\mu}^-) - \rho \boldsymbol{e}^\top(\boldsymbol{\lambda}^+ + \boldsymbol{\lambda}^-) \\ \text{s.t.} & (\boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-, \boldsymbol{y}) \in \mathcal{F}^j_{\text{NS},1} \\ & \boldsymbol{\lambda}^+ \in \{0,1\}^n,\ \boldsymbol{\lambda}^- \in \{0,1\}^n. \end{array} \tag{A.21}$$

Note that $\lambda_i^+ \in \{0,1\} \Leftrightarrow (\lambda_i^+)^2 = \lambda_i^+$. Therefore, these binary constraints can be enforced by using $2n$ quadratic equality constraints. Letting $\boldsymbol{x} := [t, (\boldsymbol{\mu}^+)^\top, (\boldsymbol{\mu}^-)^\top, (\boldsymbol{\lambda}^+)^\top, (\boldsymbol{\lambda}^-)^\top, (\boldsymbol{y})^\top]^\top$, we rewrite (A.21) as

$$\begin{array}{ll} \sup & \boldsymbol{G}^1_j(\rho,\, \boldsymbol{s}) \bullet \boldsymbol{x}\boldsymbol{x}^\top \\ \text{s.t.} & \boldsymbol{e_1}\boldsymbol{e_1}^\top \bullet \boldsymbol{x}\boldsymbol{x}^\top = 1 \\ & \boldsymbol{M}_i \bullet \boldsymbol{x}\boldsymbol{x}^\top = 0 \quad \forall\, i \in [n] \\ & \boldsymbol{J}_i \bullet \boldsymbol{x}\boldsymbol{x}^\top = 0 \quad \forall\, i \in [n] \\ & \boldsymbol{x} \in \mathcal{K}^j_{\text{NS},1}, \end{array} \tag{A.22}$$

which, by Lemma 10, can be reformulated as the following copositive program

$$\begin{array}{ll} \sup & \boldsymbol{G}^1_j(\rho,\, \boldsymbol{s}) \bullet \boldsymbol{X} \\ \text{s.t.} & \boldsymbol{e}_1\boldsymbol{e}_1^\top \bullet \boldsymbol{X} = 1 \\ & \boldsymbol{M}_i \bullet \boldsymbol{X} = 0 \quad \forall\, i \in [n] \\ & \boldsymbol{J}_i \bullet \boldsymbol{X} = 0 \quad \forall\, i \in [n] \\ & \boldsymbol{X} \in \mathcal{CP}(\mathcal{K}^j_{\text{NS},1}). \end{array} \tag{A.23}$$

It can be easily verified that $\{\boldsymbol{x} \in \mathcal{K}^j_{\text{NS},1} : \boldsymbol{e}^\top \boldsymbol{x} = 1\}$ is nonempty and bounded. Therefore, by Lemma 11, the optimal value of (A.23) is equal to that of its dual problem:

$$\begin{array}{ll} \inf & \beta_j \\ \text{s.t.} & \beta_j \in \mathbb{R},\ \psi_{ij} \in \mathbb{R},\ \phi_{ij} \in \mathbb{R}\, \forall\, i \in [n] \\ & \beta_j\, \boldsymbol{e}_1\boldsymbol{e}_1^\top + \sum\limits_{i=1}^n \psi_{ij}\boldsymbol{M}_i + \sum\limits_{i=1}^n \phi_{ij}\boldsymbol{J}_i - \boldsymbol{G}^1_j(\rho,\, \boldsymbol{s}) \in \mathcal{COP}(\mathcal{K}^j_{\text{NS},1}). \end{array} \tag{A.24}$$

Using the same argument for all N maximization problems yields the copositive programming reformulation in (2.21) for (W-NS).

Second, if $p = 2$, then the maximization problem embedded in (2.20) for any $j \in [N]$ as

$$\begin{array}{ll} \sup & (\boldsymbol{\mu} - \boldsymbol{s})^\top \boldsymbol{y} - \rho\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^j\|_2^2 - \rho\|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^j\|_2^2 \\ \text{s.t.} & (\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \Xi,\ \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{\lambda}), \end{array} \tag{A.25}$$

which can be reformulated to

$$\begin{array}{ll} \sup & \boldsymbol{G}_j^2(\rho,\, \boldsymbol{s}) \bullet \boldsymbol{x}\boldsymbol{x}^\top \\ \text{s.t.} & \boldsymbol{e_1}\boldsymbol{e_1}^\top \bullet \boldsymbol{x}\boldsymbol{x}^\top = 1,\ \boldsymbol{x} \in \mathcal{K}^j_{\text{NS},2} \\ & \boldsymbol{N}_i \bullet \boldsymbol{x}\boldsymbol{x}^\top = 0 \quad \forall\, i \in [n], \end{array} \tag{A.26}$$

where $\boldsymbol{x} := (t, \boldsymbol{\mu}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{y}^\top)^\top$. Using a similar argument, we can reformulate (W-NS) to the copositive program in (2.22) for the case of $p = 2$. □

Proof of Proposition 4

Proof. First, we rewrite $\omega'_j(\rho, \boldsymbol{s})$ as follows:

$$\begin{aligned} \omega'_j(\rho, \boldsymbol{s}) &= \sup_{(\boldsymbol{\mu},\boldsymbol{\lambda})\in\Xi} \left\{ \sup_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \left\{ \sum_{i=1}^n (\mu_i - s_i) y_i \right\} - \rho \|(\boldsymbol{\mu}^\top, \boldsymbol{\lambda}^\top)^\top - \widehat{\boldsymbol{\xi}}^j\|_p^p \right\} \\ &= \sup_{(\boldsymbol{\mu},\boldsymbol{\lambda})\in\Xi}\ \sup_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \left\{ \sum_{i=1}^n \Big(y_i(\mu_i - s_i) - \rho|\mu_i - \widehat{\mu}_i^j|^p - \rho|\lambda_i - \widehat{\lambda}_i^j|^p \Big) \right\} &\text{(A.27a)} \\ &= \sup_{\boldsymbol{\lambda}\in\Lambda,\ \boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \left\{ \sum_{i=1}^n \sup_{u_i^{\mathrm{L}}\lambda_i \le \mu_i \le u_i^{\mathrm{U}}\lambda_i} \left\{ \sum_{i=1}^n y_i(\mu_i - s_i) - \rho|\mu_i - \widehat{\mu}_i^j|^p - \rho|\lambda_i - \widehat{\lambda}_i^j|^p \right\} \right\} \\ & &\text{(A.27b)} \\ &= \sup_{\boldsymbol{\lambda}\in\Lambda,\ \boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \left\{ \sum_{i=1}^n f_{ij}(\lambda_i, y_i) \right\}, &\text{(A.27c)} \end{aligned}$$

where equality (A.27b) is because, for fixed $\boldsymbol{\lambda}$ and $\boldsymbol{y}$, the objective function in (A.27a) is separable in the index i and in each μ_i.

Second, for any fixed $\boldsymbol{\lambda} \in \Lambda$, $\sum_{i=1}^n f_{ij}(\lambda_i, y_i)$ is convex in $\boldsymbol{y}$. It follows that there exists an optimal $\boldsymbol{y}^*$ to problem (A.27c) such that $\boldsymbol{y}^*$ lies in an extreme point of the polytope $\mathcal{Y}(\boldsymbol{\lambda})$. Hence, without loss of optimality, variables $\boldsymbol{y}$ satisfy the first two conditions in (OC), i.e., $y_n = C$ or $y_n = -d_0$, and $y_i = -d_0$ or $y_i = y_{i+1} + \lambda_{i+1}$ for all $i \in [n-1]$. As a result, $y_n \in \{-d_0\} \cup \{C\} \equiv \mathcal{Y}_n$. As $y_{n-1} = -d_0$ or $y_{n-1} = y_n + \lambda_n$, we have $y_{n-1} \in \{-d_0, -d_0 + 1\} \cup \{C, C + 1\}$ because $\lambda_n \in \{0, 1\}$. Backward recursion of this analysis yields that $y_i \in \mathcal{Y}_i$ for all $i \in [n]$, which is the final condition in (OC). This completes the proof. □

Proof of Proposition 5

Proof. By construction, there exists a one-to-one mapping between a $(\boldsymbol{\lambda}, \boldsymbol{y}) \in \Lambda \times \mathcal{Y}(\boldsymbol{\lambda})$ and an S–E path of the network $(\mathcal{G}, \mathcal{E})$. In addition, the length of the S–E path generating $(\boldsymbol{\lambda}, \boldsymbol{y})$ coincides with $\sum_{i=1}^{n} f_{ij}(\lambda_i, y_i)$ by definition of $g_{k\ell j}$. Therefore, the optimal value of problem (A.27c), and so $\omega'_j(\rho, \boldsymbol{s})$, equals the length of the longest path in the network $(\mathcal{G}, \mathcal{E})$, i.e., the optimal value of problem (2.25a)–(2.25c). $\square$

Proof of Theorem 7

Proof. First, when $p = 1$, we have

$$f_{ij}(\lambda_i, y_i) = \sup_{u_i^{\mathrm{L}}\lambda_i \le \mu_i \le u_i^{\mathrm{U}}\lambda_i} \left\{ y_i(\mu_i - s_i) - \rho|\mu_i - \widehat{\mu}_i^j| - \rho|\lambda_i - \widehat{\lambda}_i^j| \right\}, \ \forall i \in [n], j \in [N].$$

As $\lambda_i \in \{0, 1\}$,

$$f_{ij}(\lambda_i, y_i) = \begin{cases} -y_i s_i - \rho\widehat{\mu}_i^j - \rho\widehat{\lambda}_i^j & \text{if } \lambda_i = 0 \\[2ex] \max\Big\{ y_i(u_i^{\mathrm{L}} - s_i) - \rho(\widehat{\mu}_i^j - u_i^{\mathrm{L}}) - \rho(1 - \widehat{\lambda}_i^j), & \\ \qquad y_i(\widehat{\mu}_i^j - s_i) - \rho(1 - \widehat{\lambda}_i^j), & \text{if } \lambda_i = 1. \\ \qquad y_i(u_i^{\mathrm{U}} - s_i) - \rho(u_i^{\mathrm{U}} - \widehat{\mu}_i^j) - \rho(1 - \widehat{\lambda}_i^j) \Big\} & \end{cases}$$

Second, taking the dual of the linear program (2.25a)–(2.25c) yields

$$\begin{aligned} \omega'_j(\rho,\, \boldsymbol{s}) \;=\; \min_{\boldsymbol{\alpha}^j} \;& \alpha_{\mathrm{S}}^j - \alpha_{\mathrm{E}}^j \\ \text{s.t.} \;& \alpha_k^j - \alpha_\ell^j \ge g_{k\ell j} \;\; \forall\, (k, \ell) \in \mathcal{E}, \end{aligned}$$

where dual variables α_k^j, $k \in \mathcal{G}$ are associated with primal constraints (2.25b). By definition, $g_{k\ell j} = 0$ if $(k, \ell) \in \mathcal{E}_{\mathrm{E}}$. In addition, for all $i \in [n]$ and $j \in [N]$, if $(k, \ell) \in \mathcal{E}_i^0$ then $g_{k\ell j} = f_{ij}(0, y_i)$; and if $(k, \ell) \in \mathcal{E}_i^1$ then $g_{k\ell j} = f_{ij}(1, y_i)$. It follows that

$$\begin{aligned} \omega'_j(\rho,\, \boldsymbol{s}) \;=\; \min_{\boldsymbol{\alpha}^j} \;& \alpha_{\mathrm{S}}^j - \alpha_{\mathrm{E}}^j \\ \text{s.t.} \;& \alpha_k^j - \alpha_\ell^j \ge 0 \;\; \forall (k, \ell) \in \mathcal{E}_{\mathrm{E}} \\ & \alpha_k^j - \alpha_\ell^j \ge -y_i s_i - (\widehat{\mu}_i^j + \widehat{\lambda}_i^j)\rho \;\; \forall\, i \in [n],\; \forall\, (k, \ell) \in \mathcal{E}_i^0 \end{aligned}$$

$$\left.\begin{array}{l}\alpha_k^j - \alpha_\ell^j \geq -y_i s_i - (1 - \widehat{\lambda}_i^j + \widehat{\mu}_i^j - u_i^{\text{L}})\rho + u_i^{\text{L}} y_i \\ \alpha_k^j - \alpha_\ell^j \geq -y_i s_i - (1 - \widehat{\lambda}_i^j)\rho + \widehat{\mu}_i^j y_i \\ \alpha_k^j - \alpha_\ell^j \geq -y_i s_i - (1 - \widehat{\lambda}_i^j + u_i^{\text{U}} - \widehat{\mu}_i^j)\rho + u_i^{\text{U}} y_i\end{array}\right\} \forall\, i \in [n],\ \forall\,(k,\ell) \in \mathcal{E}_i^1.$$

The claimed reformulation follows from substituting $\omega_j'(\rho, \boldsymbol{s})$ back into formulation (2.20) with the above linear program representation.

Third, when $p = 2$, we have

$$f_{ij}(\lambda_i, y_i) = \sup_{u_i^{\text{L}}\lambda_i \leq \mu_i \leq u_i^{\text{U}}\lambda_i} \left\{ y_i(\mu_i - s_i) - \rho(\mu_i - \widehat{\mu}_i^j)^2 - \rho(\lambda_i - \widehat{\lambda}_i^j)^2 \right\},\ \forall i \in [n],\ j \in [N].$$

Hence, if $\lambda_i = 0$ then $f_{ij}(0, y_i) = -y_i s_i - \rho\big(\widehat{\mu}_i^j\big)^2 - \rho\widehat{\lambda}_i^j$; and if $\lambda_i = 1$ then

$$\begin{aligned} & f_{ij}(1, y_i) \\ = & \sup_{u_i^{\text{L}} \leq \mu_i \leq u_i^{\text{U}}} \left\{ y_i(\mu_i - s_i) - \rho(\mu_i - \widehat{\mu}_i^j)^2 - \rho(1 - \widehat{\lambda}_i^j)^2 \right\} \\ = & \inf_{\beta^{\text{L}} \geq 0,\ \beta^{\text{U}} \geq 0} \sup_{\mu_i \in \mathbb{R}} \left\{ y_i(\mu_i - s_i) - \rho(\mu_i - \widehat{\mu}_i^j)^2 - \rho(1 - \widehat{\lambda}_i^j) + \beta^{\text{L}}(\mu_i - u_i^{\text{L}}) + \beta^{\text{U}}(u_i^{\text{U}} - \mu_i) \right\} \quad \text{(A.28)} \\ = & \inf_{\beta^{\text{L}} \geq 0,\ \beta^{\text{U}} \geq 0} \left\{ \frac{(y_i + \beta^{\text{L}} - \beta^{\text{U}})^2}{4\rho} + y_i(\widehat{\mu}_i^j - s_i) + \beta^{\text{L}}(\widehat{\mu}_i^j - u_i^{\text{L}}) + \beta^{\text{U}}(u_i^{\text{U}} - \widehat{\mu}_i^j) - \rho(1 - \widehat{\lambda}_i^j) \right\} \\ = & \inf_{\beta^{\text{L}} \geq 0,\ \beta^{\text{U}} \geq 0,\ \varphi} \left\{ \varphi + y_i(\widehat{\mu}_i^j - s_i) + \beta^{\text{L}}(\widehat{\mu}_i^j - u_i^{\text{L}}) + \beta^{\text{U}}(u_i^{\text{U}} - \widehat{\mu}_i^j) - \rho(1 - \widehat{\lambda}_i^j) \right\} \\ & \quad \text{s.t.} \quad \left\| \begin{bmatrix} \beta^{\text{L}} - \beta^{\text{U}} + y_i \\ \varphi - \rho \end{bmatrix} \right\|_2 \leq \varphi + \rho, \end{aligned}$$

where the strong (Lagrangian) duality holds in equality (A.28) because the primal formulation is strictly feasible. It follows that

$$\begin{aligned} \omega_j'(\rho,\, \boldsymbol{s}) \ = \ \min_{\boldsymbol{\alpha}^j, \boldsymbol{\beta}^{\text{L}}, \boldsymbol{\beta}^{\text{U}}, \boldsymbol{\varphi}} \ & \alpha_{\text{S}}^j - \alpha_{\text{E}}^j \\ \text{s.t.}\ & \alpha_k^j - \alpha_\ell^j \geq 0 \ \ \forall (k,\ell) \in \mathcal{E}_{\text{E}} \\ & \alpha_k^j - \alpha_\ell^j \geq -y_i s_i - \left(\big(\widehat{\mu}_i^j\big)^2 + \widehat{\lambda}_i^j \right) \rho \ \ \forall\, i \in [n],\ \forall\,(k,\ell) \in \mathcal{E}_i^0 \\ & \alpha_k^j - \alpha_\ell^j \geq -y_i s_i - (1 - \widehat{\lambda}_i^j)\rho + \varphi_{k\ell j} + (\widehat{\mu}_i^j - u_i^{\text{L}})\beta_{k\ell j}^{\text{L}} + (u_i^{\text{U}} - \widehat{\mu}_i^j)\beta_{k\ell j}^{\text{U}} + y_i \widehat{\mu}_i^j \\ & \forall\, i \in [n],\ \forall\,(k,\ell) \in \mathcal{E}_i^1 \\ & \left\| \begin{bmatrix} \beta_{k\ell j}^{\text{L}} - \beta_{k\ell j}^{\text{U}} + y_i \\ \varphi_{k\ell j} - \rho \end{bmatrix} \right\|_2 \leq r_{k\ell j} + \rho \ \ \forall\, i \in [n],\ \forall\,(k,\ell) \in \mathcal{E}_i^1. \end{aligned}$$

The claimed reformulation follows from substituting $\omega'_j(\rho, \boldsymbol{s})$ back into formulation (2.20) with the above second-order conic program representation. The proof is completed. $\square$

Proof of Theorem 8

Proof. First, for fixed $\bar{\boldsymbol{s}} \in \mathcal{S}$ and $\epsilon \geq 0$, the dual problem of formulation (2.26) is

$$\max_{\boldsymbol{o},\boldsymbol{p},\boldsymbol{q},\boldsymbol{w},\boldsymbol{r}} \ \sum_{j=1}^{N}\sum_{i=1}^{n} y_i \left[\sum_{(k,\ell)\in\mathcal{E}_i^1} \big((u_i^L - \bar{s}_i)q_{k\ell j} + (\widehat{\mu}_i^j - \bar{s}_i)w_{k\ell j} + (u_i^U - \bar{s}_i)r_{i\ell j}\big) - \sum_{(k,\ell)\in\mathcal{E}_i^0} \bar{s}_i p_{k\ell j} \right]$$

$$\text{s.t.} \ \sum_{j=1}^{N}\sum_{i=1}^{n}\Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0}(\widehat{\mu}_i^j + \widehat{\lambda}_i^j)p_{k\ell j} + \sum_{(k,\ell)\in\mathcal{E}_i^1}\Bigg(\big(1 - \widehat{\lambda}_i^j + \widehat{\mu}_i^j - u_i^{\mathrm{L}}\big)q_{k\ell j}$$
$$+ \big(1 - \widehat{\lambda}_i^j\big)w_{k\ell j} + \big(1 - \widehat{\lambda}_i^j + u_i^U - \widehat{\mu}_i^j\big)r_{k\ell j}\Bigg)\Bigg] \leq \epsilon$$

$$\sum_{\ell:(k,\ell)\in\mathcal{E}^0} p_{k\ell j} + \sum_{\ell:(k,\ell)\in\mathcal{E}^1}(q_{k\ell j} + w_{k\ell j} + r_{k\ell j}) - \sum_{\ell:(\ell,k)\in\mathcal{E}^0} p_{\ell k j}$$
$$- \sum_{\ell:(\ell,k)\in\mathcal{E}^1}(q_{\ell k j} + w_{\ell k j} + r_{\ell k j}) = \begin{cases} \frac{1}{N} & \text{if } k = \mathrm{S} \\ 0 & \text{if } k \neq \mathrm{S} \end{cases} \quad \forall\, k \in \mathcal{E} \setminus \big(\mathcal{L}(n) \cup \mathrm{E}\big),\ \forall\, j \in [N]$$

$$o_{k\mathrm{E}j} - \sum_{\ell:(\ell,k)\in\mathcal{E}^0} p_{\ell k j} - \sum_{\ell:(\ell,k)\in\mathcal{E}^1}(q_{\ell k j} + w_{\ell k j} + r_{\ell k j}) = 0 \ \ \forall\, k \in \mathcal{L}(n),\ \forall\, j \in [N]$$

$$\sum_{k:(k,\mathrm{E})\in\mathcal{E}_{\mathrm{E}}} o_{k\mathrm{E}j} = \frac{1}{N} \ \ \forall\, j \in [N]$$

$$o_{k\ell j},\ p_{k\ell j},\ q_{k\ell j},\ w_{k\ell j},\ r_{k\ell j} \geq 0 \ \ \forall\,(k,\ell) \in \mathcal{E},\ \forall\, j \in [N],$$

where dual variables $o_{k\ell j}$, $p_{k\ell j}$, $q_{k\ell j}$, $w_{k\ell j}$, and $r_{k\ell j}$ are associated with the constraints in (2.26). The strong duality holds because the primal feasible region is non-empty. Indeed, if needed, we can always increase the values of α_k^j for some $j \in [N]$ and $k \in \mathcal{G}$ to satisfy the constraints in (2.26) (note that this can be done arbitrarily because there are no directed cycles in the network $(\mathcal{G}, \mathcal{E})$). Replacing $\{o_{k\ell j}, p_{k\ell j}, q_{k\ell j}, w_{k\ell j}, r_{k\ell j}\}$ with $\frac{1}{N}\{o_{k\ell j}, p_{k\ell j}, q_{k\ell j}, w_{k\ell j}, r_{k\ell j}\}$ yields formulation (2.27a)–(2.27f).

Second, to see the existence of $\mathbb{P}_{\boldsymbol{z}}^j$ for all $j \in [N]$, we note that the set $\mathcal{P}$ is defined through constraints (2.25b), which are network balance constraints and so yield a totally unimodular constraint matrix. Hence, conv($\mathcal{P}$) $= \{\boldsymbol{z} \in [0,1]^{|\mathcal{E}|} : \ (2.25\text{b})\}$. As constraints (2.27c)–(2.27f) possess the structure of network balance constraints with respect to $[\boldsymbol{p}_j^\star, \boldsymbol{\eta}_j^\star, \boldsymbol{o}_j^\star] := [p_{k\ell j}^\star : (k,\ell) \in \mathcal{E}^0;\ (q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star) : (k,\ell) \in \mathcal{E}^1;\ o_{k\ell j}^\star : (k,\ell) \in \mathcal{E}_{\mathrm{E}}]$, we have $[\boldsymbol{p}_j^\star, \boldsymbol{\eta}_j^\star, \boldsymbol{o}_j^\star] \in$ conv($\mathcal{P}$). It follows that there exists a finite number of points

$\boldsymbol{z}^1, \dots, \boldsymbol{z}^I \in \mathcal{P}$ and weights $\theta^1, \dots, \theta^I \in [0,1]$ such that

$$[\boldsymbol{p}_j^\star, \boldsymbol{\eta}_j^\star, \boldsymbol{o}_j^\star] \;=\; \sum_{i=1}^{I} \theta^i \boldsymbol{z}^i, \;\; \sum_{i=1}^{I} \theta^i = 1.$$

Hence, the distribution $\mathbb{P}_{\boldsymbol{z}}^j$ constructed by setting $\mathbb{P}_{\boldsymbol{z}}^j\{\boldsymbol{z} = \boldsymbol{z}^i\} = \theta^i$ for all $i \in [I]$ fulfills the claim.

Third, to prove that $\mathbb{Q}_{\boldsymbol{\xi}}^\star \in \mathcal{D}_1(\widehat{\mathbb{P}}_{\boldsymbol{\xi}}^N, \epsilon)$, we note that $\mu_{k\ell j} \in [u_i^{\text{L}}, u_i^{\text{U}}]$ whenever $(k,\ell) \in \mathcal{E}_i^1$. Indeed, on the one hand, if $q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star = 0$ then $q_{k\ell j}^\star = r_{k\ell j}^\star = 0$ because $q_{k\ell j}^\star, w_{k\ell j}^\star, r_{k\ell j}^\star \geq 0$. Thus, by the extended arithmetics $0/0 = 0$ we have $\mu_{k\ell j}^\star = \widehat{\mu}_i^j \in [u_i^{\text{L}}, u_i^{\text{U}}]$. On the other hand, if $q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star > 0$ then

$$\mu_{k\ell j} = \left(\frac{w_{k\ell j}^\star}{q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star}\right)\widehat{\mu}_i^j + \left(\frac{q_{k\ell j}^\star}{q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star}\right)u_i^{\text{L}} + \left(\frac{r_{k\ell j}^\star}{q_{k\ell j}^\star + w_{k\ell j}^\star + r_{k\ell j}^\star}\right)u_i^{\text{U}}.$$

It follows that $\mu_{k\ell j}$ is a convex combination of $\widehat{\mu}_i^j$, u_i^{L}, and u_i^{U} and so $\mu_{k\ell j}^\star \in [u_i^{\text{L}}, u_i^{\text{U}}]$. In addition, we define a joint probability distribution $\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}$ of $(\widehat{\boldsymbol{\xi}}, \boldsymbol{\xi}, \boldsymbol{z})$ through $\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}} = \frac{1}{N}\sum_{j=1}^N \sum_{\boldsymbol{\zeta}\in\mathcal{P}} \mathbb{P}_{\boldsymbol{z}}^j\{\boldsymbol{z} = \boldsymbol{\zeta}\}\delta_{\widehat{\boldsymbol{\xi}}^j, \boldsymbol{\xi}^j(\boldsymbol{\zeta}), \boldsymbol{\zeta}}$. Note that the projection of $\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}$ on $\boldsymbol{\xi}$ is $\mathbb{Q}_{\boldsymbol{\xi}}^\star$. It follows that

$$\begin{aligned}
& d_1(\mathbb{Q}_{\boldsymbol{\xi}}^\star, \widehat{\mathbb{P}}_{\boldsymbol{\xi}}^N) \\
=\; & \inf_{\Pi \in \mathcal{P}(\mathbb{Q}_{\boldsymbol{\xi}}^\star, \widehat{\mathbb{P}}_{\boldsymbol{\xi}}^N)} \mathbb{E}_{\Pi}\big[\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\|_1\big] \\
\leq\; & \mathbb{E}_{\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}}\big[\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\|_1\big] && \text{(A.29a)} \\
=\; & \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}}\big[|\mu_i - \widehat{\mu}_i| + |\lambda_i - \hat{\lambda}_i|\big] \\
=\; & \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \sum_{\boldsymbol{\zeta}\in\mathcal{P}} \mathbb{P}_{\boldsymbol{z}}^j\{\boldsymbol{z} = \boldsymbol{\zeta}\}\left[\sum_{(k,\ell)\in\mathcal{E}_i^0} \zeta_{k\ell}(\widehat{\mu}_i^j + \widehat{\lambda}_i^j) + \sum_{(k,\ell)\in\mathcal{E}_i^1} \zeta_{k\ell}\big(|\mu_{k\ell j} - \widehat{\mu}_i^j| + 1 - \widehat{\lambda}_i^j\big)\right] && \text{(A.29b)} \\
=\; & \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} \sum_{\zeta\in\mathcal{P}:\zeta_{k\ell}=1} \mathbb{P}_{\boldsymbol{z}}^j\{\boldsymbol{z} = \boldsymbol{\zeta}\}(\widehat{\mu}_i^j + \widehat{\lambda}_i^j) \\
& \quad + \sum_{(k,\ell)\in\mathcal{E}_i^1} \sum_{\zeta\in\mathcal{P}:\zeta_{k\ell}=1} \mathbb{P}_{\boldsymbol{z}}^j\{\boldsymbol{z} = \boldsymbol{\zeta}\}\big(|\mu_{k\ell j} - \widehat{\mu}_i^j| + 1 - \widehat{\lambda}_i^j\big)\Bigg] \\
=\; & \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} p_{k\ell j}^\star(\widehat{\mu}_i^j + \widehat{\lambda}_i^j)
\end{aligned}$$

$$+ \sum_{(k,\ell)\in\mathcal{E}_i^1} (q^\star_{k\ell j} + w^\star_{k\ell j} + r^\star_{k\ell j}) \left(\frac{|q^\star_{k\ell j}(u_i^L - \widehat{\mu}_i^j) + r^\star_{k\ell j}(u_i^U - \widehat{\mu}_i^j)|}{q^\star_{k\ell j} + w^\star_{k\ell j} + r^\star_{k\ell j}} + 1 - \widehat{\lambda}_i^j \right) \Bigg] \tag{A.29c}$$

$$\leq \sum_{i=1}^n \left(\frac{1}{N}\right) \sum_{j=1}^N \Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} p^\star_{k\ell j}(\widehat{\mu}_i^j + \widehat{\lambda}_i^j)$$

$$+ \sum_{(k,\ell)\in\mathcal{E}_i^1} (q^\star_{k\ell j} + w^\star_{k\ell j} + r^\star_{k\ell j}) \left(\frac{q^\star_{k\ell j}(\widehat{\mu}_i^j - u_i^L) + r^\star_{k\ell j}(u_i^U - \widehat{\mu}_i^j)}{q^\star_{k\ell j} + w^\star_{k\ell j} + r^\star_{k\ell j}} + 1 - \widehat{\lambda}_i^j \right) \Bigg]$$

$$= \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^N \Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} p^\star_{k\ell j}(\widehat{\mu}_i^j + \widehat{\lambda}_i^j) + \sum_{(k,\ell)\in\mathcal{E}_i^1} \Big(q^\star_{k\ell j}(\widehat{\mu}_i^j - u_i^L + 1 - \widehat{\lambda}_i^j)$$

$$+ w^\star_{k\ell j}(1 - \widehat{\lambda}_i^j) + r^\star_{k\ell j}(u_i^U - \widehat{\mu}_i^j + 1 - \widehat{\lambda}_i^j) \Big) \Bigg] \leq \epsilon, \tag{A.29d}$$

where inequality (A.29a) is because the projections of $\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}$ on $\boldsymbol{\xi}$ and $\widehat{\boldsymbol{\xi}}$ are $\mathbb{Q}^\star_{\boldsymbol{\xi}}$ and $\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}}$, respectively, equality (A.29b) follows from the definition of $\mathbb{Q}_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}$, equality (A.29c) follows from the facts that, for all $j \in [N]$, $\sum_{\zeta\in\mathcal{P}:\zeta_{k\ell}=1} \mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z} = \boldsymbol{\zeta}\} = p^\star_{k\ell j}$ whenever $(k,\ell) \in \mathcal{E}^0$ and $\sum_{\zeta\in\mathcal{P}:\zeta_{k\ell}=1} \mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z} = \boldsymbol{\zeta}\} = q^\star_{k\ell j} + w^\star_{k\ell j} + r^\star_{k\ell j}$ whenever $(k,\ell) \in \mathcal{E}^1$, and the inequality in (A.29d) follows from constraint (2.27b).

Finally, we have

$$\sup_{\mathbb{Q}_{\boldsymbol{\xi}}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}}, \boldsymbol{\xi})]$$

$$\geq \mathbb{E}_{\mathbb{Q}^\star_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}}, \boldsymbol{\xi})] \tag{A.30a}$$

$$= \mathbb{E}_{\mathbb{Q}^\star_{\boldsymbol{\xi}}} \left[\max_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{\lambda})} \left\{ \sum_{i=1}^n y_i(\mu_i - \bar{s}_i) \right\} \right]$$

$$\geq \mathbb{E}_{\mathbb{Q}^\star_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}} \left[\sum_{i=1}^n y_i(\mu_i - \bar{s}_i) \right] \tag{A.30b}$$

$$= \frac{1}{N} \sum_{j=1}^N \sum_{\boldsymbol{\zeta}\in\mathcal{P}} \mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z} = \boldsymbol{\zeta}\} \sum_{i=1}^n \Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} \zeta_{k\ell}\, y_i\, (0 - \bar{s}_i) + \sum_{(k,\ell)\in\mathcal{E}_i^1} \zeta_{k\ell}\, y_i\, (\mu_{k\ell j} - \bar{s}_i) \Bigg] \tag{A.30c}$$

$$= \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^n \Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} \sum_{\boldsymbol{\zeta}\in\mathcal{P}:\zeta_{k\ell}=1} \mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z} = \boldsymbol{\zeta}\}(-y_i\bar{s}_i) + \sum_{(k,\ell)\in\mathcal{E}_i^1} \sum_{\boldsymbol{\zeta}\in\mathcal{P}:\zeta_{k\ell}=1} \mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z} = \boldsymbol{\zeta}\} y_i(\mu_{k\ell j} - \bar{s}_i) \Bigg]$$

$$= \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\Bigg[\sum_{(k,\ell)\in\mathcal{E}_i^0} p^\star_{k\ell j}(-y_i\bar{s}_i)$$
$$+ \sum_{(k,\ell)\in\mathcal{E}_i^1}(q^\star_{k\ell j}+w^\star_{k\ell j}+r^\star_{k\ell j})y_i\left(\widehat{\mu}_i^j + \frac{q^\star_{k\ell j}(u_i^L-\widehat{\mu}_i^j)}{q^\star_{k\ell j}+w^\star_{k\ell j}+r^\star_{k\ell j}} + \frac{r^\star_{k\ell j}(u_i^U-\widehat{\mu}_i^j)}{q^\star_{k\ell j}+w^\star_{k\ell j}+r^\star_{k\ell j}} - \bar{s}_i\right)\Bigg] \tag{A.30d}$$

$$= \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n} y_i\Bigg[-\sum_{(k,\ell)\in\mathcal{E}_i^0} p^\star_{k\ell j}\bar{s}_i + \sum_{(k,\ell)\in\mathcal{E}_i^1}\Big(q^\star_{k\ell j}(u_i^L-\bar{s}_i) + w^\star_{k\ell j}(\widehat{\mu}_i^j-\bar{s}_i) + r^\star_{k\ell j}(u_i^U-\bar{s}_i)\Big)\Bigg]$$
$$= \sup_{\mathbb{Q}_{\boldsymbol{\xi}}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\epsilon)} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}},\boldsymbol{\xi})], \tag{A.30e}$$

where inequality (A.30a) is because $\mathbb{Q}^\star_{\boldsymbol{\xi}} \in \mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\epsilon)$, inequality (A.30b) is because we replace the maximization over variables $\boldsymbol{y}$ with a distribution of $\boldsymbol{y}$ pushed forward by the random path $\boldsymbol{z}$, equality (A.30c) follows from the definition of $\mathbb{Q}^\star_{\widehat{\boldsymbol{\xi}},\boldsymbol{\xi},\boldsymbol{z}}$, equality (A.30d) follows from the facts that, for all $j \in [N]$, $\sum_{\zeta\in\mathcal{P}:\zeta_{k\ell}=1}\mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z}=\boldsymbol{\zeta}\} = p^\star_{k\ell j}$ whenever $(k,\ell)\in\mathcal{E}^0$ and $\sum_{\zeta\in\mathcal{P}:\zeta_{k\ell}=1}\mathbb{P}^j_{\boldsymbol{z}}\{\boldsymbol{z}=\boldsymbol{\zeta}\} = q^\star_{k\ell j}+w^\star_{k\ell j}+r^\star_{k\ell j}$ whenever $(k,\ell)\in\mathcal{E}^1$, and equality (A.30e) is because $\sup_{\mathbb{Q}_{\boldsymbol{\xi}}\in\mathcal{D}_1(\widehat{\mathbb{P}}^N_{\boldsymbol{\xi}},\epsilon)}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\xi}}}[g(\bar{\boldsymbol{s}},\boldsymbol{\xi})]$ equals the optimal value of formulation (2.27a)–(2.27f). This completes the proof. □

APPENDIX B

Appendix for Chapter 3

B.1 Proof of Lemma 4

Proof. Representing variables $e_j \equiv \sum_{i:j\in P_i} z_{ij} + x_j - \tilde{d}_j + \tilde{w}_j$ by constraints (3.1b), we rewrite formulation (3.1a)–(3.1d) as

$$V(\tilde{w}, \tilde{y}, \tilde{d}) = \min_{z,x} \sum_{j=1}^{J} \Big[(c^{x} - c^{e}) x_j - c^{e} \sum_{i:j\in P_i} z_{ij} \Big] + \sum_{j=1}^{J} c^{e} (\tilde{d}_j - \tilde{w}_j)$$

$$\text{s.t. } x_j + \sum_{i:j\in P_i} z_{ij} \geq \tilde{d}_j - \tilde{w}_j, \ \ \forall j \in [J], \tag{B.1a}$$

$$\sum_{j\in P_i} z_{ij} \leq \tilde{y}_i, \ \ \forall i \in [I], \tag{B.1b}$$

$$x_j \in \mathbb{Z}_+, \ \ \forall j \in [J], \ \ z_{ij} \in \mathbb{Z}_+, \ \ \forall i \in [I], \forall j \in P_i.$$

We note that the constraint matrix of the above formulation is totally unimodular (TU), and so the conclusion follows. To see the TU property, we multiply each of the constraints (B.1b) by -1 on both sides and recast the constraint matrix of (B.1a)–(B.1b) in the following form:

$$\begin{bmatrix} \Big(x_j + \sum_{i:j\in P_i} z_{ij} \Big) \\ \Big(- \sum_{j\in P_i} z_{ij} \Big) \end{bmatrix}.$$

It follows that (a) each entry of this matrix is -1, 0, or 1, (b) this matrix has at most two nonzero entries in each column, and (c) the entries sum up to be zero for any column containing two nonzero entries. Hence, the constraint matrix is TU based on Proposition 2.6 in [135]. The conclusion follows because $\tilde{d}_j - \tilde{w}_j$ and $-\tilde{y}_i$ are integers for all $j \in [J]$ and for all $i \in [I]$, respectively. $\square$

B.2 Verifying Assumption 4

We present necessary and sufficient conditions for Assumption 4 in the following proposition.

Proposition 16. *For any given w and y, $\mathcal{D}$ is non-empty if and only if the following three conditions are satisfied:*

1. $f_j(w_j) \in [0, w_j]$ *for all* $j \in [J]$*;*
2. $g_i(y_i) \in [0, y_i]$ *for all* $i \in [I]$*;*
3. *For all $j \in [J]$, the optimal value of the following linear program is non-positive:*

$$\min_{p_j \geq 0, \tau \geq 0} \quad \sum_{q=1}^{Q} (\tau_q^+ + \tau_q^-) \tag{B.2a}$$

$$\text{s.t.} \quad \sum_{k=d_j^L}^{d_j^U} k^q p_{jk} + \tau_q^+ - \tau_q^- = \mu_{jq}, \quad \forall q \in [Q], \tag{B.2b}$$

$$\sum_{k=d_j^L}^{d_j^U} p_{jk} = 1. \tag{B.2c}$$

Proof. (**Necessity**) Suppose that $\mathcal{D} \neq \emptyset$. Then, there exists a $\mathbb{P} \in \mathcal{P}(\Xi)$ such that $\mathbb{E}_{\mathbb{P}}[\tilde{d}_j^q] = \mu_{jq}$ for all $j \in [J]$ and $q \in [Q]$, $\mathbb{E}_{\mathbb{P}}[\tilde{w}_j] = f_j(w_j)$ for all $j \in [J]$, and $\mathbb{E}_{\mathbb{P}}[\tilde{y}_i] = g_i(y_i)$ for all $i \in [I]$. It follows that, for all $j \in [J]$, we have $f_j(w_j) \leq \text{esssup}_{\Xi}\{\tilde{w}_j\} \leq w_j$ and $f_j(w_j) \geq \text{essinf}_{\Xi}\{\tilde{w}_j\} \geq 0$, leading to $f_j(w_j) \in [0, w_j]$. Likewise, it holds that $g_i(y_i) \in [0, y_i]$ for all $i \in [I]$. In addition, for all $j \in [J]$ and $k \in [d_j^{\text{L}}, d_j^{\text{U}}]_{\mathbb{Z}}$, we let $\bar{p}_{jk} = \mathbb{P}\{\tilde{d}_j = k\}$. It follows that, for all $q \in [Q]$, $\sum_{k=d_j^{\text{L}}}^{d_j^{\text{U}}} \bar{p}_{jk} = \sum_{k=d_j^{\text{L}}}^{d_j^{\text{U}}} \mathbb{P}\{\tilde{d}_j = k\} = 1$. Moreover,

$$\sum_{k=d_j^{\text{L}}}^{d_j^{\text{U}}} k^q \bar{p}_{jk} \;=\; \sum_{k=d_j^{\text{L}}}^{d_j^{\text{U}}} k^q \mathbb{P}\{\tilde{d}_j = k\} \;=\; \mathbb{E}_{\mathbb{P}}[\tilde{d}_j^q] \;=\; \mu_{jq}.$$

Hence, together with $\tau_q^+ = \tau_q^- = 0$, $\bar{p}_{jk}$ constitutes a feasible solution to linear program (B.2) with an objective value being zero. As zero is also a lower bound of the objective value, $\bar{p}_{jk}$ is optimal to (B.2) and accordingly the optimal value of this linear program equals zero. This holds for all $j \in [J]$ and proves the necessity of the three conditions.

(**Sufficiency**) Suppose that the three conditions are satisfied. For all $j \in [J]$, as $f_j(w_j) \in [0, w_j] \equiv \text{conv}([0, w_j]_{\mathbb{Z}})$ by condition 1, there exists a $\mathbb{P}_{\tilde{w}_j} \in \mathcal{P}([0, w_j]_{\mathbb{Z}})$ such that $f_j(w_j) =$

$\mathbb{E}_{\mathbb{P}_{\tilde{w}_j}}[\tilde{w}_j]$. Likewise, for all $i \in [I]$, there exists a $\mathbb{P}_{\tilde{y}_i} \in \mathcal{P}([0, y_i]_{\mathbb{Z}})$ such that $g_i(y_i) = \mathbb{E}_{\mathbb{P}_{\tilde{y}_i}}[\tilde{y}_i]$. In addition, as the optimal value of (B.2) is non-positive and $\tau_q^+, \tau_q^- \geq 0$ for all $q \in [Q]$, the optimal value of (B.2) equals zero. It follows that, for all $j \in [J]$, there exist p_{jk} such that $\sum_{k=d_j^{\mathrm{L}}}^{d_j^{\mathrm{U}}} k^q p_{jk} = \mu_{jq}$ for all $q \in [Q]$ and $\sum_{k=d_j^{\mathrm{L}}}^{d_j^{\mathrm{U}}} p_{jk} = 1$. Defining $\mathbb{P}_{\tilde{d}_j} \in \mathcal{P}([d_j^{\mathrm{L}}, d_j^{\mathrm{U}}]_{\mathbb{Z}})$ such that $\mathbb{P}_{\tilde{d}_j}\{\tilde{d}_j = k\} = p_{jk}$ for all $k \in [d_j^{\mathrm{L}}, d_j^{\mathrm{U}}]_{\mathbb{Z}}$, we have $\mathbb{E}_{\mathbb{P}_{\tilde{d}_j}}[\tilde{d}_j^q] = \mu_{jq}$. Therefore, the probability distribution

$$\mathbb{P} := \Pi_{j=1}^J \mathbb{P}_{\tilde{w}_j} \times \Pi_{i=1}^I \mathbb{P}_{\tilde{y}_i} \times \Pi_{j=1}^J \mathbb{P}_{\tilde{d}_j}$$

satisfies constraints (3.2a)–(3.2c) and hence $\mathbb{P} \in \mathcal{D}$. It follows that $\mathcal{D} \neq \emptyset$ and the proof is completed. □

B.3 Proof of Proposition 6

Proof. First, denoting $\tilde{\xi} := (\tilde{w}, \tilde{y}, \tilde{d})$, we present $\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[V(\tilde{w}, \tilde{y}, \tilde{d})]$ as the following optimization problem:

$$\max_{p \geq 0} \sum_{\tilde{\xi} \in \Xi} p_{\tilde{\xi}} V(\tilde{\xi})$$

$$\text{s.t.} \sum_{\tilde{\xi} \in \Xi} p_{\tilde{\xi}} \tilde{w}_j = f_j(w_j), \ \ \forall j \in [J], \tag{B.3a}$$

$$\sum_{\tilde{\xi} \in \Xi} p_{\tilde{\xi}} \tilde{y}_i = g_i(y_i), \ \ \forall i \in [I], \tag{B.3b}$$

$$\sum_{\tilde{\xi} \in \Xi} p_{\tilde{\xi}} \tilde{d}_j^q = \mu_{jq}, \ \ \forall j \in [J], \ \forall q \in [Q], \tag{B.3c}$$

$$\sum_{\tilde{\xi} \in \Xi} p_{\tilde{\xi}} = 1, \tag{B.3d}$$

where decision variables $p_{\tilde{\xi}}$ represent the probability of the random variables being realized as $\tilde{\xi}$, and constraints (B.3a)–(B.3d) describe the ambiguity set $\mathcal{D}$ defined in (3.2a)–(3.2c). The dual of this formulation is

$$\min_{\gamma,\lambda,\rho,\theta} \sum_{j=1}^J \sum_{q=1}^Q \mu_{jq}\rho_{jq} + \sum_{j=1}^J f_j(w_j)\gamma_j + \sum_{i=1}^I g_i(y_i)\lambda_i + \theta \tag{B.4a}$$

$$\text{s.t.} \ \theta + \sum_{j=1}^J \sum_{q=1}^Q \rho_{jq}\tilde{d}_j^q + \sum_{j=1}^J \gamma_j \tilde{w}_j + \sum_{i=1}^I \lambda_i \tilde{y}_i \geq V(\tilde{\xi}), \ \ \forall \tilde{\xi} \in \Xi, \tag{B.4b}$$

where dual variables γ_j, λ_i, ρ_{jq}, and θ are associated with primal constraints (B.3a)–(B.3d), respectively, and dual constraints (B.4b) are associated with primal variables $p_{\tilde{\xi}}$. By Assumption 4, strong duality holds between the primal and dual formulations because they are both linear programs. As the objective function aims to minimize the value of θ, we observe by constraints (B.4b) that $\theta = \sup_{\tilde{\xi}\in\Omega}\{V(\tilde{\xi}) - \sum_{j=1}^{J}\sum_{q=1}^{Q}\rho_{jq}\tilde{d}_j^q - \sum_{j=1}^{J}\gamma_j\tilde{w}_j - \sum_{i=1}^{I}\lambda_i\tilde{y}_i\}$. Hence, $\sup_{\mathbb{P}\in\mathcal{D}}\mathbb{E}_{\mathbb{P}}[V(\tilde{\xi})]$ equals the optimal value of the following min-max optimization problem:

$$\begin{aligned}&\min_{\gamma,\lambda,\rho}\max_{\tilde{\xi}\in\Xi}\left\{V(\tilde{\xi}) - \sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\rho_{jq}\tilde{d}_j^q + \gamma_j\tilde{w}_j\right] - \sum_{i=1}^{I}\lambda_i\tilde{y}_i\right\}\\ &+\sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\mu_{jq}\rho_{jq} + f_j(w_j)\gamma_j\right] + \sum_{i=1}^{I}g_i(y_i)\lambda_i. \qquad \text{(B.5a)}\end{aligned}$$

Second, in view of the dual formulation (3.4a)–(3.4c) of $V(\tilde{\xi})$, we rewrite the maximum term in (B.5a) as

$$\begin{aligned}&\max_{(\tilde{w},\tilde{y},\tilde{d})\in\Xi}\max_{(\alpha,\beta)\in\Lambda}\left\{\sum_{j=1}^{J}(\tilde{d}_j-\tilde{w}_j)\alpha_j + \sum_{i=1}^{I}\tilde{y}_i\beta_i - \sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\rho_{jq}\tilde{d}_j^q + \gamma_j\tilde{w}_j\right] - \sum_{i=1}^{I}\lambda_i\tilde{y}_i\right\}\\ =\ &\max_{(\alpha,\beta)\in\Lambda}\max_{(\tilde{w},\tilde{y},\tilde{d})\in\Xi}\left\{\sum_{j=1}^{J}(\tilde{d}_j-\tilde{w}_j)\alpha_j + \sum_{i=1}^{I}\tilde{y}_i\beta_i - \sum_{j=1}^{J}\left[\sum_{q=1}^{Q}\rho_{jq}\tilde{d}_j^q + \gamma_j\tilde{w}_j\right] - \sum_{i=1}^{I}\lambda_i\tilde{y}_i\right\}\\ =\ &\max_{(\alpha,\beta)\in\Lambda}\Bigg\{\sum_{j=1}^{J}\max_{\tilde{w}_j\in[0,w_j]_{\mathbb{Z}}}\Big\{(-\alpha_j-\gamma_j)\tilde{w}_j\Big\} + \sum_{i=1}^{I}\max_{\tilde{y}_i\in[0,y_i]_{\mathbb{Z}}}\Big\{(\beta_i-\lambda_i)\tilde{y}_i\Big\} +\\ &\qquad\sum_{j=1}^{J}\max_{\tilde{d}_j\in[d_j^{\mathrm{L}},d_j^{\mathrm{U}}]_{\mathbb{Z}}}\Big\{\alpha_j\tilde{d}_j - \sum_{q=1}^{Q}\rho_{jq}\tilde{d}_j^q\Big\}\Bigg\}.\end{aligned}$$

Finally, as $(-\alpha_j-\gamma_j)\tilde{w}_j$ is linear in $\tilde{w}_j$, we have

$$\max_{\tilde{w}_j\in[0,w_j]_{\mathbb{Z}}}\Big\{(-\alpha_j-\gamma_j)\tilde{w}_j\Big\} = \max\Big\{0,(-\alpha_j-\gamma_j)w_j\Big\} = \Big[(-\alpha_j-\gamma_j)w_j\Big]_+.$$

Similarly, we have $\max_{\tilde{y}_i\in[0,y_i]_{\mathbb{Z}}}\{(\beta_i-\lambda_i)\tilde{y}_i\} = [(\beta_i-\lambda_i)y_i]_+$. This completes the proof. □

B.4 Proof of Lemma 5

Proof. As $\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta)$ is to maximize a convex function over a polyhedron, we only need to analyze the extreme directions and extreme points of Λ.

First, the extreme directions of Λ are $(\alpha,\beta) = (0, -e_i)$ for all $i \in [I]$, where e_i represents the i^{th} standard basis vector. As $y_i \geq 0$, moving along any of these extreme directions (i.e., decreasing the value of any β_i) does not increase the value of $F(\alpha,\beta)$. Hence, we can omit these extreme directions in the attempt of maximizing $F(\alpha,\beta)$ and accordingly $\bar{\beta}_i = \min\{-\bar{\alpha}_j : j \in P_i\} = -\max\{\bar{\alpha}_j : j \in P_i\}$ without loss of optimality. This proves property (b) in the claim. In addition, there exists an extreme point of Λ that is optimal to $\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta)$.

Second, we prove, by contradiction, that each extreme point of Λ satisfies property (a) in the claim. Suppose that there exists an extreme point $(\bar{\alpha},\bar{\beta})$ such that property (a) fails, i.e., $\bar{\alpha}_{j^*} \in (c^{\text{e}}, c^{\text{x}})$ for some $j^* \in [J]$. Consider the set $\mathcal{I}(j^*) := \{i \in [I] : -\bar{\beta}_i = \bar{\alpha}_{j^*}\}$. We discuss the following two cases. In each case, we shall construct two points in Λ such that their midpoint is $(\bar{\alpha},\bar{\beta})$, which provides a desired contradiction.

1. If $\mathcal{I}(j^*) = \emptyset$, then $-\bar{\beta}_i > \bar{\alpha}_{j^*}$ for all i such that $j^* \in P_i$. Defining $\epsilon := (1/2)\min\{-\bar{\beta}_i - \bar{\alpha}_{j^*}, \bar{\alpha}_{j^*} - c^{\text{e}}, c^{\text{x}} - \bar{\alpha}_{j^*}\} > 0$, we construct two points $(\bar{\alpha}^+,\bar{\beta})$ and $(\bar{\alpha}^-,\bar{\beta})$ such that $\bar{\alpha}^+_{j^*} = \bar{\alpha}_{j^*} + \epsilon$, $\bar{\alpha}^-_{j^*} = \bar{\alpha}_{j^*} - \epsilon$, and $\bar{\alpha}^+_j = \bar{\alpha}^-_j = \bar{\alpha}_j$ for all $j \neq j^*$. Then, it is clear that $(\bar{\alpha}^+,\bar{\beta}), (\bar{\alpha}^-,\bar{\beta}) \in \Lambda$. But $(\bar{\alpha},\bar{\beta}) = (1/2)(\bar{\alpha}^+,\bar{\beta}) + (1/2)(\bar{\alpha}^-,\bar{\beta})$, which contradicts the fact that $(\bar{\alpha},\bar{\beta})$ is an extreme point of Λ.

2. If $\mathcal{I}(j^*) \neq \emptyset$, then we define $\mathcal{J}(j^*) := \bigcup_{i\in\mathcal{I}(j^*)}\{j \in P_i : \bar{\alpha}_j = -\bar{\beta}_i\}$. It follows that $\bar{\alpha}_j = \bar{\alpha}_{j^*}$ for all $j \in \mathcal{J}(j^*)$. Hence, for each $i \in \mathcal{I}(j^*)$, $\bar{\alpha}_j = \bar{\alpha}_{j^*}$ for all $j \in P_i \cap \mathcal{J}(j^*)$ and $\bar{\alpha}_j < \bar{\alpha}_{j^*}$ for all $j \in P_i \setminus \mathcal{J}(j^*)$. We define $\epsilon := (1/2)\min\Big\{\min\{\bar{\alpha}_{j^*} - \bar{\alpha}_j : i \in \mathcal{I}(j^*), j \in P_i \setminus \mathcal{J}(j^*)\},\ \min\{-\bar{\beta}_i - \bar{\alpha}_{j^*} : i \notin \mathcal{I}(j^*),\ -\bar{\beta}_i > \bar{\alpha}_{j^*}\},\ \bar{\alpha}_{j^*} - c^{\text{e}}, c^{\text{x}} - \bar{\alpha}_{j^*}\Big\}$. Then $\epsilon > 0$ because it is the minimum of a finite number of positive reals.[1] We construct two points $(\bar{\alpha}^+,\bar{\beta}^+)$ and $(\bar{\alpha}^-,\bar{\beta}^-)$ such that

$$\bar{\alpha}^+_j = \begin{cases} \bar{\alpha}_{j^*} + \epsilon & \forall j \in \mathcal{J}(j^*) \\ \bar{\alpha}_j & \text{otherwise} \end{cases}, \quad \bar{\alpha}^-_j = \begin{cases} \bar{\alpha}_{j^*} - \epsilon & \forall j \in \mathcal{J}(j^*) \\ \bar{\alpha}_j & \text{otherwise} \end{cases},$$

$$\bar{\beta}^+_i = \begin{cases} -(\bar{\alpha}_{j^*} + \epsilon) & \forall i \in \mathcal{I}(j^*) \\ \bar{\beta}_i & \text{otherwise} \end{cases}, \quad \bar{\beta}^-_i = \begin{cases} -(\bar{\alpha}_{j^*} - \epsilon) & \forall i \in \mathcal{I}(j^*) \\ \bar{\beta}_i & \text{otherwise} \end{cases}.$$

[1] Here we adopt the convention that $\min\{a : a \in A\} = \infty$ if $A = \emptyset$. For example, if there does not exist an $i \notin \mathcal{I}(j^*)$ such that $-\bar{\beta}_i > \bar{\alpha}_{j^*}$, then $\min\{-\bar{\beta}_i - \bar{\alpha}_{j^*} : i \notin \mathcal{I}(j^*),\ -\bar{\beta}_i > \bar{\alpha}_{j^*}\} = \infty$.

It is clear that $(\bar{\alpha}, \bar{\beta}) = (1/2)(\bar{\alpha}^+, \bar{\beta}^+) + (1/2)(\bar{\alpha}^-, \bar{\beta}^-)$. To finish the proof, it remains to show that $(\bar{\alpha}^+, \bar{\beta}^+), (\bar{\alpha}^-, \bar{\beta}^-) \in \Lambda$. To see this, we check constraints (3.4b) and (3.4c). For constraints (3.4c), we have $\bar{\alpha}_j^+ \in (c^{\text{e}}, c^{\text{x}})$ for all $j \in \mathcal{J}(j^*)$ by the definition of ϵ. Additionally, for all $j \notin \mathcal{J}(j^*)$, we have $\bar{\alpha}_j^+ = \bar{\alpha}_j^- = \bar{\alpha}_j \in [c^{\text{e}}, c^{\text{x}}]$. Hence, constraints (3.4c) are indeed satisfied and it remains to check constraints (3.4b). For each $i \in \mathcal{I}(j^*)$, $-\bar{\beta}_i^+ = \bar{\alpha}_{j^*} + \epsilon = \bar{\alpha}_j^+$ for all $j \in P_i \cap \mathcal{J}(j^*)$ and $-\bar{\beta}_i^+ = \bar{\alpha}_{j^*} + \epsilon \geq \bar{\alpha}_{j^*} \geq \bar{\alpha}_j = \bar{\alpha}_j^+$ for all $j \in P_i \setminus \mathcal{J}(j^*)$, where the first inequality is because $\epsilon > 0$, and the second inequality follows from the definition of $\mathcal{J}(j^*)$. Meanwhile, $-\bar{\beta}_i^- = \bar{\alpha}_{j^*} - \epsilon = \bar{\alpha}_j^-$ for all $j \in P_i \cap \mathcal{J}(j^*)$, and $-\bar{\beta}_i^- = \bar{\alpha}_{j^*} - \epsilon \geq \bar{\alpha}_j = \bar{\alpha}_j^-$ for all $j \in P_i \setminus \mathcal{J}(j^*)$, where the inequality follows from the definition of ϵ and the last equality is because $j \notin \mathcal{J}(j^*)$. It follows that constraints (3.4b) are indeed satisfied for all $i \in \mathcal{I}(j^*)$. For each $i \notin \mathcal{I}(j^*)$, $\bar{\beta}_i^+ = \bar{\beta}_i^- = \bar{\beta}_i$ and $-\bar{\beta}_i \neq \bar{\alpha}_{j^*}$. We discuss the following two sub-cases to complete the proof.

(a) If $-\bar{\beta}_i > \bar{\alpha}_{j^*}$, then $-\bar{\beta}_i^+ = -\bar{\beta}_i \geq \bar{\alpha}_{j^*} + \epsilon \geq \bar{\alpha}_j^+$, where the first inequality follows from the definition of ϵ. In addition, by construction $-\bar{\beta}_i^- = -\bar{\beta}_i > \bar{\alpha}_{j^*} \geq \bar{\alpha}_j^-$ for all $j \in P_i$.

(b) If $-\bar{\beta}_i < \bar{\alpha}_{j^*}$, then $j \notin \mathcal{J}(j^*)$ for all $j \in P_i$ because otherwise $-\bar{\beta}_i \geq \bar{\alpha}_j = \bar{\alpha}_{j^*}$. It follows that $\bar{\alpha}_j^+ = \bar{\alpha}_j^- = \bar{\alpha}_j$ and so $-\bar{\beta}_i^+ = -\bar{\beta}_i \geq \bar{\alpha}_j = \bar{\alpha}_j^+$ and $-\bar{\beta}_i^- = -\bar{\beta}_i \geq \bar{\alpha}_j = \bar{\alpha}_j^-$. □

B.5 Proof of Theorem 9

Proof. First, pick any $(\alpha, \beta) \in \Lambda$ that satisfies the optimality conditions (a)–(b) stated in Lemma 5. We shall show that there exists a feasible solution (t, s, r, p) to formulation (3.7a)–(3.7d) that attains the same objective function value as $F(\alpha, \beta)$. To this end, for all $j \in [J]$, we let $t_j = 1$ if $\alpha_j = c^{\text{x}}$ and $t_j = 0$ if $\alpha_j = c^{\text{e}}$. In addition, for all $i \in [I]$, if $\alpha_j = c^{\text{e}}$ for all $j \in P_i$ then we let $s_{ij} = 0$ for all $j \in P_i$; and otherwise, we pick the largest $j^* \in P_i$ such that $\alpha_{j^*} = c^{\text{x}}$, and let $s_{ij^*} = 1$ and all other $s_{ij} = 0$. Also, we define r and p as in (3.7c) and (3.7d), respectively. By construction (t, s, r, p) satisfies constraints (3.7b)–(3.7d). It follows that the objective function value of (t, s, r, p) equals

$$
\begin{aligned}
& \sum_{j=1}^{J} \left(c_j^{\text{t}} t_j + c_j^{\text{r}} r_j\right) + \sum_{i=1}^{I} \Big(c_i^{\text{p}} p_i + \sum_{j \in P_i} c_i^{\text{s}} s_{ij}\Big) \\
= & \sum_{j:\alpha_j = c^{\text{x}}} \left\{ \Big[(-c^{\text{x}} - \gamma_j) w_j\Big]_+ + \sup_{\tilde{d}_j \in [d_j^{\text{L}}, d_j^{\text{U}}]_{\mathbb{Z}}} \Big\{ c^{\text{x}} \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{j:\alpha_j = c^{\mathrm{e}}} \left\{ \Big[(-c^{\mathrm{e}} - \gamma_j) w_j\Big]_+ + \sup_{\tilde{d}_j \in [d_j^{\mathrm{L}}, d_j^{\mathrm{U}}]_{\mathbb{Z}}} \Big\{ c^{\mathrm{e}} \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\} \right\} \\
&+ \sum_{i=1}^{I} \left\{ \mathbb{1}_{\{c^{\mathrm{x}}\}}\Big(\max_{j \in P_i}\{\alpha_j\}\Big) \Big[(-c^{\mathrm{x}} - \lambda_i) y_i\Big]_+ + \mathbb{1}_{\{c^{\mathrm{e}}\}}\Big(\max_{j \in P_i}\{\alpha_j\}\Big) \Big[(-c^{\mathrm{e}} - \lambda_i) y_i\Big]_+ \right\} \\
= \; & \sum_{j=1}^{J} \left\{ \Big[(-\alpha_j - \gamma_j) w_j\Big]_+ + \sup_{\tilde{d}_j \in [d_j^{\mathrm{L}}, d_j^{\mathrm{U}}]_{\mathbb{Z}}} \Big\{ \alpha_j \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\} \right\} + \sum_{i=1}^{I} \Big[(\beta_i - \lambda_i) y_i\Big]_+ \\
= \; & F(\alpha, \beta),
\end{aligned}
$$

where the first equality follows from the definition of (t, s, r, p) and the second equality follows from the optimality condition (b).

Second, pick any feasible solution (t, s, r, p) to formulation (3.7a)–(3.7d). We construct an $(\alpha, \beta) \in \Lambda$ such that it satisfies the optimality conditions (a)–(b) and $F(\alpha, \beta)$ equals the objective function value (3.7a) of (t, s, r, p). Specifically, for all $j \in [J]$, we let $\alpha_j = c^{\mathrm{x}} t_j + c^{\mathrm{e}} r_j$ and, for all $i \in [I]$, $\beta_i = -c^{\mathrm{e}} p_i - c^{\mathrm{x}} \sum_{j \in P_i} s_{ij}$. Then, for all $i \in [I]$ and $j \in P_i$,

$$
\begin{aligned}
\beta_i + \alpha_j \; &= \; -c^{\mathrm{e}} p_i - c^{\mathrm{x}} \sum_{j \in P_i} s_{ij} + c^{\mathrm{x}} t_j + c^{\mathrm{e}} r_j \\
&= \; -c^{\mathrm{e}} \Big(1 - \sum_{j \in P_i} s_{ij}\Big) - c^{\mathrm{x}} \sum_{j \in P_i} s_{ij} + c^{\mathrm{x}} t_j + c^{\mathrm{e}}(1 - t_j) \\
&= (c^{\mathrm{e}} - c^{\mathrm{x}}) \Big(\sum_{j \in P_i} s_{ij} - t_j\Big) \; \leq \; 0,
\end{aligned}
$$

where the first equality is due to constraints (3.7c)–(3.7d) and the inequality is due to constraints (3.6d). Next, we have $\alpha_j \in \{c^{\mathrm{x}}, c^{\mathrm{e}}\}$ for all $j \in [J]$ due to constraint (3.7c). Hence, $(\alpha, \beta) \in \Lambda$. Also, for all $i \in [I]$, if $\sum_{\ell \in P_i} s_{i\ell} = 0$ then $p_i = 1$ due to constraints (3.7d) and $t_j = 0$ for all $j \in P_i$ due to constraint (3.6d). It follows that $\alpha_j = c^{\mathrm{e}}$ for all $j \in P_i$ and so $\beta_i = -\max\{\alpha_j : j \in P_i\}$. On the other hand, if $\sum_{\ell \in P_i} s_{i\ell} = 1$ then $p_i = 0$ and there exists an $j^* \in P_i$ with $t_{j^*} = 1$ due to constraints (3.6b). It follows that $\alpha_{j^*} = c^{\mathrm{x}}$ and so $\beta_i = -\max\{\alpha_j : j \in P_i\}$. Hence, (α, β) satisfies the optimality conditions (a)–(b). Finally,

$$
\begin{aligned}
F(\alpha, \beta) = \; & \sum_{j=1}^{J} \left\{ \Big[(-c^{\mathrm{x}} t_j - c^{\mathrm{e}} r_j - \gamma_j) w_j\Big]_+ + \sup_{\tilde{d}_j \in [d_j^{\mathrm{L}}, d_j^{\mathrm{U}}]_{\mathbb{Z}}} \Big\{ \big(c^{\mathrm{x}} t_j + c^{\mathrm{e}} r_j\big) \tilde{d}_j - \sum_{q=1}^{Q} \rho_{jq} \tilde{d}_j^q \Big\} \right\} \\
&+ \sum_{i=1}^{I} \Big[\Big(-c^{\mathrm{e}} p_i - c^{\mathrm{x}} \sum_{j \in P_i} s_{ij} - \lambda_i\Big) y_i\Big]_+
\end{aligned}
$$

$$= \sum_{j=1}^{J} \left(c_j^{\mathrm{t}} t_j + c_j^{\mathrm{r}} r_j\right) + \sum_{i=1}^{I} \Big(c_i^{\mathrm{p}} p_i + \sum_{j \in P_i} c_i^{\mathrm{s}} s_{ij}\Big)$$

by the definition of (α, β) and constraints (3.7c)–(3.7d). This completes the proof. □

B.6 Proof of Proposition 7

Proof. Let $G(\gamma, \lambda, \rho)$ be the objective function of problem (3.5a)–(3.5b), $(\hat{\gamma}, \hat{\lambda}, \hat{\rho})$ be any feasible solution, and S^* be the set of optimal solution to problem $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta)$ for the given $(\hat{\gamma}, \hat{\lambda}, \hat{\rho})$. Suppose that there exists a $j \in [J]$ such that $\hat{\gamma}_j < -c^{\mathrm{x}}$. Then, $-\hat{\gamma}_j - \alpha_j^* > 0$ and $[(-\hat{\gamma}_j - \alpha_j^*)w_j]_+ = (-\hat{\gamma}_j - \alpha_j^*)w_j$ for all $(\alpha^*, \beta^*) \in S^*$ because $\alpha_j^* \le c^{\mathrm{x}}$ by Lemma 5. Additionally, due to Lemma 5, we can replace polyhedron Λ by the (compact) set of its extreme points ex(Λ) without loss of optimality, i.e., $\max_{(\alpha,\beta)\in\Lambda} F(\alpha, \beta) = \max_{(\alpha,\beta)\in\mathrm{ex}(\Lambda)} F(\alpha, \beta)$. It then follows from Theorem 2.87 in [164] that, for all subgradient $\varpi \in \partial G(\hat{\gamma}, \hat{\lambda}, \hat{\rho})$, the entry in ϖ with regard to variable γ_j at $(\hat{\gamma}, \hat{\lambda}, \hat{\rho})$ equals $f_j(w_j) - w_j$, i.e., $\varpi(\gamma_j)|_{(\hat{\gamma},\hat{\lambda},\hat{\rho})} = f_j(w_j) - w_j \le 0$. Noting that $\varpi(\gamma_j)|_{(\hat{\gamma},\hat{\lambda},\hat{\rho})} \le 0$ holds valid whenever $\hat{\gamma}_j < -c^{\mathrm{x}}$, we can increase $\hat{\gamma}_j$ to $-c^{\mathrm{x}}$ without any loss of optimality.

Now suppose that $\hat{\gamma}_j > 0$. Then, we have $-\hat{\gamma}_j - \alpha_j^* < 0$ and $[(-\hat{\gamma}_j - \alpha_j^*)w_j]_+ = 0$ for all $(\alpha^*, \beta^*) \in S^*$ because $\alpha^* \ge 0$ by Lemma 5. It follows from a similar implication as in the previous paragraph that, for all subgradient $\varpi \in \partial G(\hat{\gamma}, \hat{\lambda}, \hat{\rho})$, we have $\varpi(\gamma_j)|_{(\hat{\gamma},\hat{\lambda},\hat{\rho})} = f_j(w_j) \ge 0$. Noting that this holds valid whenever $\hat{\gamma}_j > 0$, we can decrease $\hat{\gamma}_j$ to 0 without any loss of optimality. Therefore, there exists an optimal solution $(\gamma^*, \lambda^*, \rho^*)$ to problem (3.5a)–(3.5b) such that $\gamma_j^* \in [-c^{\mathrm{x}}, 0]$ for all $j \in [J]$.

Following a similar proof, we can show that there exists an optimal solution $(\gamma^*, \lambda^*, \rho^*)$ to problem (3.5a)–(3.5b) such that $\lambda_i^* \in [-c^{\mathrm{x}}, 0]$ for all $i \in [I]$. We omit the details for the sake of saving space. □

B.7 Proof of Theorem 11

Proof. In each iteration of Algorithm 3.1, we solve a relaxation of the (DRNS) reformulation (3.9a)–(3.9d). It follows that, if the algorithm stops in an iteration and returns a solution (u^*, v^*) then (u^*, v^*) satisfies all the constraints (3.9c) because of Step 5. Then, (u^*, v^*) is feasible to formulation (3.9a)–(3.9d) and meanwhile optimal to its relaxation. Hence, (u^*, v^*) is optimal to formulation (3.9a)–(3.9d), i.e., optimal to (DRNS).

It remains to show that Algorithm 3.1 stops in a finite number of iterations. To see this, we notice that the set $\mathcal{H}$ contains a finite number of elements. Indeed, binary variables t and s only have a finite number of possible values. Although r and p are continuous variables, they also only have a finite number of possible values due to constraints (3.7c)–(3.7d). $\square$

B.8 Proof of Lemma 6

Proof. Pick any $i \in [I]$ and $j \in P_i$. We note that $\sum_{\ell \in P_i} s_{i\ell} \in \{0, 1\}$ due to constraints (3.6a) and discuss the following three cases. First, if $\sum_{\ell \in P_i} s_{i\ell} = 0$, i.e., if $s_{i\ell} = 0$ for all $\ell \in P_i$, then $t_j = 0$ by constraints (3.6d). The inequalities (3.10) hold valid because all $s_{i\ell} \geq 0$.

Second, if $\sum_{\ell \in P_i} s_{i\ell} = 1$ and $\sum_{\ell \in P_i : \ell \geq j} s_{ij} = 1$, then inequalities (3.10) hold valid because $t_j \leq 1$.

Third, if $\sum_{\ell \in P_i} s_{i\ell} = 1$ and $\sum_{\ell \in P_i : \ell \geq j} s_{ij} = 0$, then there exists some $k \in P_i$, $k < j$ such that $s_{ik=1}$. Then, $t_j \leq 1 - s_{ik} = 0$ by constraints (3.6c). Inequalities (3.10) follow and the proof is complete. $\square$

B.9 Proof of Theorem 12

Proof. As $I = 1$ and $P_1 = [J]$, we drop the index i and re-write $\overline{\mathcal{H}} = \big\{(t, s, r, p) : \sum_{j=1}^{J} s_j \leq 1,\ \sum_{j=1}^{J} s_j + p = 1,\ s_j \leq t_j,\ t_j \leq \sum_{\ell=j}^{J} s_\ell,\ t_j + r_j = 1,\ s_j, t_j \in \mathbb{R}_+,\ \forall j \in [J]\big\}$. To show that $\text{conv}(\mathcal{H}) = \overline{\mathcal{H}}$, we first note that $\mathcal{H} \subseteq \overline{\mathcal{H}}$. This is because inequalities (3.10) are satisfied by all $(t, s, r, p) \in \mathcal{H}$. It follows that $\mathcal{H} \subseteq \overline{\mathcal{H}}$ and hence $\text{conv}(\mathcal{H}) \subseteq \overline{\mathcal{H}}$.

Second, we prove that $\overline{\mathcal{H}} \subseteq \text{conv}(\mathcal{H})$. To this end, we claim that optimizing any linear objective function over $\overline{\mathcal{H}}$ yields at least an optimal solution that lies in $\mathcal{H}$ (see [135]). If this claim holds valid then all the extreme points of $\overline{\mathcal{H}}$ lie in $\mathcal{H}$ and hence $\overline{\mathcal{H}} \subseteq \text{conv}(\mathcal{H})$ by the Minkowski's theorem on polyhedron. To prove this claim, we consider a linear program

$$\max_{t,s,r,p \geq 0} \quad \sum_{j=1}^{J} (a_j^{\mathrm{s}} s_j + a_j^{\mathrm{t}} t_j + a_j^{\mathrm{r}} r_j + a^{\mathrm{p}} p) \tag{B.7a}$$

$$\text{s.t.} \quad \sum_{j=1}^{J} s_j \leq 1, \tag{B.7b}$$

$$s_j \leq t_j, \quad \forall j \in [J], \tag{B.7c}$$

$$t_j \le \sum_{\ell=j}^{J} s_\ell, \ \ \forall j \in [J], \tag{B.7d}$$

$$t_j + r_j = 1, \ \ \forall j \in [J], \tag{B.7e}$$

$$\sum_{j=1}^{J} s_j + p = 1, \tag{B.7f}$$

where a^{s}_j, a^{t}_j, a^{r}_j, and a^{p} represent arbitrary objective function coefficients. By the last two constraints, we re-write this linear program as

$$\begin{aligned}
&\max_{t,s\ge 0} \ \sum_{j=1}^{J}\big[(a^{\mathrm{s}}_j - a^{\mathrm{p}})s_j + (a^{\mathrm{t}}_j - a^{\mathrm{r}}_j)t_j\big] + a^{\mathrm{p}} + \sum_{j=1}^{J} a^{\mathrm{r}}_j \\
&\quad \text{s.t. } \ \text{(B.7b)–(B.7d)} \\
= \ & \max_{s\ge 0:\ \text{(B.7b)}} \ \sum_{j=1}^{J}(a^{\mathrm{s}}_j - a^{\mathrm{p}})s_j + \max_{t\ge 0:\ \text{(B.7c)–(B.7d)}} \sum_{j=1}^{J}(a^{\mathrm{t}}_j - a^{\mathrm{r}}_j)t_j + a^{\mathrm{p}} + \sum_{j=1}^{J} a^{\mathrm{r}}_j \\
= \ & \max_{s\ge 0:\ \text{(B.7b)}} \ \sum_{j=1}^{J}(a^{\mathrm{s}}_j - a^{\mathrm{p}})s_j + \sum_{j=1}^{J} \max_{s_j \le t_j \le \sum_{\ell=j}^{J} s_\ell} (a^{\mathrm{t}}_j - a^{\mathrm{r}}_j)t_j + a^{\mathrm{p}} + \sum_{j=1}^{J} a^{\mathrm{r}}_j && \text{(B.7g)} \\
= \ & \max_{s\ge 0:\ \text{(B.7b)}} \ \sum_{j=1}^{J}(a^{\mathrm{s}}_j - a^{\mathrm{p}})s_j + \sum_{j:a^{\mathrm{t}}_j \ge a^{\mathrm{r}}_j} (a^{\mathrm{t}}_j - a^{\mathrm{r}}_j)\sum_{\ell=j}^{J} s_\ell + \sum_{j:a^{\mathrm{t}}_j < a^{\mathrm{r}}_j} (a^{\mathrm{t}}_j - a^{\mathrm{r}}_j)s_j + a^{\mathrm{p}} + \sum_{j=1}^{J} a^{\mathrm{r}}_j \\
= \ & \max_{s\ge 0:\ \text{(B.7b)}} \ \sum_{j=1}^{J}\Big(a^{\mathrm{s}}_j - a^{\mathrm{p}} + a^{\mathrm{t}}_j - a^{\mathrm{r}}_j + \sum_{\ell=1}^{j-1}(a^{\mathrm{t}}_j - a^{\mathrm{r}}_j)_+\Big)s_j + a^{\mathrm{p}} + \sum_{j=1}^{J} a^{\mathrm{r}}_j. && \text{(B.7h)}
\end{aligned}$$

As the formulation (B.7h) optimizes a linear function of s over a simplex, there exists an optimal solution s^* to (B.7h), and hence an optimal solution (t^*, s^*, r^*, p^*) to (B.7a)–(B.7f), such that $s^* \in \{0\} \cup \{e_j : j \in [J]\}$ and, for all $t \in [J]$, either $t^*_j = s^*_j$ or $t^*_j = \sum_{\ell=j}^{J} s^*_\ell$ in view of the inner maximization problem in (B.7g). It follows that $(t^*, s^*) \in \mathbb{B}^{2J}$ and hence $(t^*, s^*, r^*, p^*) \in \mathcal{H}$. This proves that $\text{conv}(\mathcal{H}) = \overline{\mathcal{H}}$.

Third, it follows from the above convex hull result and Theorem 9 that $\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta) = \max_{(t,s,r,p)\in\overline{\mathcal{H}}}\big\{\sum_{j=1}^{J}(c^{\mathrm{t}}_j t_j + c^{\mathrm{r}}_j r_j) + c^{\mathrm{p}}_1 p_1\big\}$ (note that $c^{\mathrm{s}}_1 = 0$ by (3.9f)). Then, following (B.7h) we have

$$\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta) \ = \ \max_{s\ge 0:\ \sum_{j=1}^{J} s_{1j} \le 1} \ \sum_{j=1}^{J}\Big(-c^{\mathrm{p}}_1 + c^{\mathrm{t}}_j - c^{\mathrm{r}}_j + \sum_{\ell=1}^{j-1}(c^{\mathrm{t}}_j - c^{\mathrm{r}}_j)_+\Big)s_{1j} + c^{\mathrm{p}}_1 + \sum_{j=1}^{J} c^{\mathrm{r}}_j,$$

which optimizes a linear function over the simplex $\{s \ge 0 :\ \sum_{j=1}^{J} s_{1j} \le 1\}$. Enumerating the extreme points of this simplex, i.e., $\{0\} \cup \{s_{1j} = 1, s_{1\ell} = 0,\ \forall \ell \ne j\}_{j=1}^{J}$, yields the claimed closed-form solution of $\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta)$. This completes the proof. $\square$

B.10 Proof of Proposition 8

Proof. By Theorem 12, constraints (3.9c) are equivalent to

$$\theta \;\geq\; \max\left\{ c^{\mathrm{p}}_1 + \sum_{\ell=1}^{J} c^{\mathrm{r}}_\ell,\; \max_{j\in[J]}\left\{ c^{\mathrm{t}}_j + \sum_{\ell=1}^{j-1} \max\{c^{\mathrm{t}}_\ell, c^{\mathrm{r}}_\ell\} + \sum_{\ell=j+1}^{J} c^{\mathrm{r}}_\ell \right\} \right\},$$

where c^{t}_j, c^{r}_j, and c^{p}_1 are computed by (3.9e)–(3.9h). Defining auxiliary variables

$$\left.\begin{array}{rcl} \zeta_j &:=& \left[(-c^{\mathrm{e}} - \gamma_j)w^{\mathrm{L}}_j - \sum_{k=1}^{w^{\mathrm{U}}_j - w^{\mathrm{L}}_j}(\varphi_{jk} + c^{\mathrm{e}} u_{jk})\right]_+ \\ \eta^{\mathrm{x}}_j &:=& \sup_{\tilde{d}_j \in [d^{\mathrm{L}}_j, d^{\mathrm{U}}_j]_{\mathbb{Z}}} \left\{ c^{\mathrm{x}} \tilde{d}_j - \sum_{q=1}^{Q} \tilde{d}^q_j \rho_{jq} \right\} \\ \eta^{\mathrm{e}}_j &:=& \sup_{\tilde{d}_j \in [d^{\mathrm{L}}_j, d^{\mathrm{U}}_j]_{\mathbb{Z}}} \left\{ c^{\mathrm{e}} \tilde{d}_j - \sum_{q=1}^{Q} \tilde{d}^q_j \rho_{jq} \right\} \end{array}\right\} \;\forall j \in [J],$$

$$\text{and}\;\; \phi_1 \;:=\; \left[(-c^{\mathrm{e}} - \lambda_1)y^{\mathrm{L}}_1 - \sum_{\ell=1}^{y^{\mathrm{U}}_1 - y^{\mathrm{L}}_1} (\nu_{1\ell} + c^{\mathrm{e}} v_{1\ell})\right]_+,$$

we have $c^{\mathrm{t}}_j = \eta^{\mathrm{x}}_j$, $c^{\mathrm{r}}_j = \zeta_j + \eta^{\mathrm{e}}_j$ for all $j \in [J]$, and $c^{\mathrm{p}}_1 = \phi_1$. It follows that constraints (3.9c) are equivalent to

$$\theta \;\geq\; \max\left\{ \phi_1 + \sum_{\ell=1}^{J}(\zeta_\ell + \eta^{\mathrm{e}}_\ell),\; \max_{j\in[J]}\left\{ \eta^{\mathrm{x}}_j + \sum_{\ell=1}^{j-1} \max\{\eta^{\mathrm{x}}_\ell, \zeta_\ell + \eta^{\mathrm{e}}_\ell\} + \sum_{\ell=j+1}^{J} (\zeta_\ell + \eta^{\mathrm{e}}_\ell) \right\} \right\}$$

$$\Leftrightarrow\; \left\{ \begin{array}{l} \theta \;\geq\; \phi_1 + \sum_{\ell=1}^{J}(\zeta_\ell + \eta^{\mathrm{e}}_\ell) \\ \theta \;\geq\; \eta^{\mathrm{x}}_j + \sum_{\ell=1}^{j-1} \max\{\eta^{\mathrm{x}}_\ell, \zeta_\ell + \eta^{\mathrm{e}}_\ell\} + \sum_{\ell=j+1}^{J}(\zeta_\ell + \eta^{\mathrm{e}}_\ell), \;\; \forall j \in [J] \end{array}\right.$$

$$\Leftrightarrow\; \exists\{\chi_j\}_{j=1}^{J} : \left\{ \begin{array}{l} \theta \;\geq\; \phi_1 + \sum_{\ell=1}^{J}(\zeta_\ell + \eta^{\mathrm{e}}_\ell) \\ \theta \;\geq\; \eta^{\mathrm{x}}_j + \sum_{\ell=1}^{j-1} \chi_\ell + \sum_{\ell=j+1}^{J}(\zeta_\ell + \eta^{\mathrm{e}}_\ell), \;\; \forall j \in [J] \\ \chi_j \geq \eta^{\mathrm{x}}_j, \;\; \chi_j \geq \zeta_j + \eta^{\mathrm{e}}_j, \;\; \forall j \in [J] \end{array}\right. \tag{B.8}$$

Replacing constraints (3.9c) with (B.8) in the formulation (3.9a)–(3.9d) and incorporating the definition of the auxiliary variables ζ_j, η^{x}_j, η^{e}_j, and ϕ_1 leads to the claimed reformulation of (DRNS). This completes the proof. $\square$

B.11 Proof of Proposition 9

We start by proving the following technical lemma.

Lemma 12. *Consider sets $A_i \subseteq \mathbb{R}^{k_i}$ for all $i \in [I]$ and let $A := \Pi_{i=1}^I A_i$. Then, conv$(A) = \Pi_{i=1}^I$conv(A_i).*

Proof. First, as $A = \Pi_{i=1}^I A_i$ and $A_i \subseteq \text{conv}(A_i)$, we have $A \subseteq \Pi_{i=1}^I \text{conv}(A_i)$ and hence $\text{conv}(A) \subseteq \Pi_{i=1}^I \text{conv}(A_i)$. Second, to show that $\Pi_{i=1}^I \text{conv}(A_i) \subseteq \text{conv}(A)$, we pick any $a \in \Pi_{i=1}^I \text{conv}(A_i)$ and prove that $a \in \text{conv}(A)$. To this end, we denote $a := [a_1, \ldots, a_I]^\top$, where $a_i \in \text{conv}(A_i)$ for all $i \in [I]$. Then, for all $i \in [I]$, there exist $\{\lambda_i^{n_i}\}_{n_i=1}^{N_i}$ and $\{a_i^{n_i}\}_{n_i=1}^{N_i}$ such that each $\lambda_i^{n_i} \geq 0$, each $a_i^{n_i} \in A_i$, $\sum_{n_i=1}^{N_i} \lambda_i^{n_i} = 1$, and $\sum_{n_i=1}^{N_i} \lambda_i^{n_i} a_i^{n_i} = a_i$ for all $i \in [I]$.

Denote set $\mathcal{N} := \{(n_1, \ldots, n_I) : n_i \in [N_i],\ \forall i \in [I]\}$, vector $a^n := [a_1^{n_1}, \ldots, a_I^{n_I}]^\top$ for all $n := [n_1, \ldots, n_I]^\top \in \mathcal{N}$, and scalar $\lambda^n := \Pi_{i=1}^I \lambda_i^{n_i}$ for all $n \in \mathcal{N}$. Then, $\lambda^n \geq 0$ and $a^n \in A$ for all $n \in \mathcal{N}$. In addition, $\sum_{n \in \mathcal{N}} \lambda^n = \sum_{n \in \mathcal{N}} \Pi_{i=1}^I \lambda_i^{n_i} = (\lambda_1^1 + \cdots + \lambda_1^{N_1})(\lambda_2^1 + \cdots + \lambda_2^{N_2}) \cdots (\lambda_I^1 + \cdots + \lambda_I^{N_I}) = 1$. Furthermore, for all $i \in [I]$, we have

$$\begin{aligned}
\sum_{n \in \mathcal{N}} \lambda^n a_i^{n_i} &= \sum_{m_i=1}^{N_i} \sum_{n \in \mathcal{N}: n_i = m_i} \lambda^n a_i^{n_i} \\
&= \sum_{m_i=1}^{N_i} a_i^{m_i} \sum_{n \in \mathcal{N}: n_i = m_i} \lambda^n \\
&= \sum_{m_i=1}^{N_i} a_i^{m_i} \lambda_i^{m_i} \\
&= a_i,
\end{aligned}$$

where third equality is because, for fixed $i \in [I]$ and $m_i \in [N_i]$, $\sum_{n \in \mathcal{N}: n_i = m_i} \lambda^n = (\lambda_1^1 + \cdots + \lambda_1^{N_1}) \cdots (\lambda_{m_i-1}^1 + \cdots + \lambda_{m_i-1}^{N_{m_i-1}}) \lambda_i^{m_i} (\lambda_{m_i+1}^1 + \cdots + \lambda_{m_i+1}^{N_{m_i+1}}) \cdots (\lambda_I^1 + \cdots + \lambda_I^{N_I}) = \lambda_i^{m_i}$, and the last inequality follows from the definitions of $\{\lambda_i^{m_i}\}_{m_i=1}^{N_i}$ and $\{a_i^{m_i}\}_{m_i=1}^{N_i}$. It follows that $a \equiv [a_1, \ldots, a_I]^\top = \sum_{n \in \mathcal{N}} \lambda^n [a_1^{n_1}, \ldots, a_I^{n_I}]^\top \equiv \sum_{n \in \mathcal{N}} \lambda^n a^n$ and hence $a \in \text{conv}(A)$. This completes the proof. □

We are now ready to present the main proof of this section.

Proof of Proposition 9: First, as $\mathcal{H}$ is separable in index i, we have $\mathcal{H} = \Pi_{i=1}^I \mathcal{H}_i$, where each $\mathcal{H}_i$ is defined as

$$\mathcal{H}_i := \Big\{ (t, s_i, r, p_i) : \sum_{j \in P_i} s_{ij} \leq 1, \tag{B.9a}$$

$$s_{ij} \leq t_j, \ \ \forall j \in P_i, \tag{B.9b}$$

$$t_j + s_{i\ell} \leq 1, \ \ \forall j, \ell \in P_i : j > \ell, \tag{B.9c}$$

$$t_j \le \sum_{\ell \in P_i} s_{i\ell}, \ \ \forall j \in P_i, \tag{B.9d}$$

$$t_j, s_{ij} \in \mathbb{B}, \ \ \forall j \in P_i, \tag{B.9e}$$

$$t_j + r_j = 1, \ \ \forall j \in P_i, \tag{B.9f}$$

$$p_i + \sum_{j \in P_i} s_{ij} = 1 \Big\}. \tag{B.9g}$$

Following a similar proof as that of Theorem 12, we can show that incorporating inequalities (3.10) produces the convex hull of $\mathcal{H}_i$, i.e., $\text{conv}(\mathcal{H}_i) = \{(t,s,r,p) \ge 0 : \text{(B.9a)–(B.9b), (B.9f)–(B.9g)}, \ t_j \le \sum_{\ell \in P_i : \ell \ge j} s_{ij}, \ \forall j \in P_i\}$. Then, it follows from Lemma 12 that $\text{conv}(\mathcal{H}) = \Pi_{i=1}^{I} \text{conv}(\mathcal{H}_i) = \big\{(t,s,r,p) \ge 0 : \text{(3.6a)–(3.6b), (3.7c)–(3.7d), (3.10)}\big\}$, as claimed.

Second, using this convex hull result, we have

$$\begin{aligned}
&\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta) \\
= &\max_{(t,s,r,p)\ge 0} \ \sum_{j=1}^{J} \big(c^{\text{t}}_j t_j + c^{\text{r}}_j r_j\big) + \sum_{i=1}^{I} c^{\text{p}}_i p_i \\
&\quad \text{s.t.} \quad \text{(3.6a)–(3.6b), (3.7c)–(3.7d), (3.10)} \\
= &\sum_{i=1}^{I} \Big\{ c^{\text{p}}_i + \sum_{j \in P_i} c^{\text{r}}_j + \max_{(t,s)\ge 0: \text{(3.6a)–(3.6b), (3.10)}} \sum_{j \in P_i} \big[(c^{\text{t}}_j - c^{\text{r}}_j) t_j - c^{\text{p}}_i s_{ij}\big] \Big\} \\
= &\sum_{i=1}^{I} \Big\{ c^{\text{p}}_i + \sum_{j \in P_i} c^{\text{r}}_j + \max_{s \ge 0: \text{(3.6a)}} \sum_{j \in P_i} \Big[-c^{\text{p}}_i + c^{\text{t}}_j - c^{\text{r}}_j + \sum_{\ell \in P_i : \ell < j} (c^{\text{t}}_j - c^{\text{r}}_j)_+ \Big] s_{ij} \Big\} \\
= &\sum_{i=1}^{I} \Big\{ c^{\text{p}}_i + \sum_{j \in P_i} c^{\text{r}}_j + \max\Big\{0, \ \max_{j \in P_i} \big\{-c^{\text{p}}_i + c^{\text{t}}_j - c^{\text{r}}_j + \sum_{\ell \in P_i : \ell < j} (c^{\text{t}}_j - c^{\text{r}}_j)_+ \big\}\Big\}\Big\} \\
= &\sum_{i \in [I]} \max\left\{ c^{\text{p}}_i + \sum_{\ell \in P_i} c^{\text{r}}_\ell, \ \max_{j \in P_i} \Big\{ c^{\text{t}}_j + \sum_{\ell \in P_i : \ell < j} \max\{c^{\text{t}}_\ell, c^{\text{r}}_\ell\} + \sum_{\ell \in P_i : \ell > j} c^{\text{r}}_\ell \Big\} \right\}.
\end{aligned}$$

Third, constraints (3.9c) in the reformulation of (DRNS) are equivalent to $\theta \ge \max_{i \in [I]}\{\theta_i\}$, where

$$\theta_i \ \ge \ \max\left\{ c^{\text{p}}_i + \sum_{\ell \in P_i} c^{\text{r}}_\ell, \ \max_{j \in P_i} \Big\{ c^{\text{t}}_j + \sum_{\ell \in P_i : \ell < j} \max\{c^{\text{t}}_\ell, c^{\text{r}}_\ell\} + \sum_{\ell \in P_i : \ell > j} c^{\text{r}}_\ell \Big\} \right\}.$$

The claimed reformulation of (DRNS) then follows from a similar proof as that of Proposition 8. □

B.12 Proof of Theorem 13

Proof. First, by construction and Theorem 9, the DP yields the same optimal value as $\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta)$. Second, each trajectory of states $\widehat{t}_1, (\widehat{t}_1, \widehat{t}_2) \ldots, (\widehat{t}_1, \widehat{t}_I)$ in the DP corresponds to a S-T path in the network $(\mathcal{N}, \mathcal{A})$, where the objective function value of the trajectory $V_I(\widehat{t}_1, \widehat{t}_I) + c^{\mathrm{p}}_I(1-\widehat{t}_I)(1-\widehat{t}_1)$ equals the length of the S-T path by definition of the arc lengths $c_{[m,n]}$. Likewise, each S-T path in $(\mathcal{N}, \mathcal{A})$ corresponds to a trajectory of states in the DP and the length of the path equals the objective function value of the trajectory. This proves that the length of the longest path in $(\mathcal{N}, \mathcal{A})$ equals $\max_{(\alpha,\beta)\in\Lambda} F(\alpha,\beta)$ and completes the proof. □

B.13 Proof of Proposition 10

Proof. Taking the dual of the longest-path formulation yields

$$\begin{aligned}
\min_{\pi} \ & \pi_{\mathrm{S}} - \pi_{\mathrm{T}} \\
\text{s.t. } & \pi_{\mathrm{S}} - \pi_{\widehat{t}_1} \geq c^{\mathrm{t}}_1 \widehat{t}_1 + c^{\mathrm{r}}_1(1-\widehat{t}_1), \ \ \forall \widehat{t}_1 \in \mathbb{B}, \\
& \pi_{(\widehat{t}_1, \widehat{t}_{i-1})} - \pi_{(\widehat{t}_1, \widehat{t}_i)} \geq c^{\mathrm{t}}_i + (c^{\mathrm{r}}_i - c^{\mathrm{t}}_i)(1-\widehat{t}_i) + c^{\mathrm{p}}_{i-1}(1-\widehat{t}_{i-1})(1-\widehat{t}_i) \\
& \forall \widehat{t}_{i-1}, \widehat{t}_i \in \mathbb{B}, \ \forall i \in [2, I]_{\mathbb{Z}}, \\
& \pi_{(\widehat{t}_1, \widehat{t}_I)} - \pi_{\mathrm{T}} \geq c^{\mathrm{p}}_I(1-\widehat{t}_I)(1-\widehat{t}_1), \ \ \forall \widehat{t}_1, \widehat{t}_I \in \mathbb{B},
\end{aligned}$$

where dual variables π are associated with the (primal) flow balance constraints and all dual constraints are associated with the primal variables x. The strong duality holds valid because the longest-path formulation is finitely optimal. The claimed reformulation of (DRNS) then follows from a similar proof as that of Proposition 8. □

B.14 Proof of Proposition 11

Proof. We linearize the bilinear terms in constraints (3.17c)–(3.17e). First, for constraints (3.17d)–(3.17e), we define auxiliary variables

$$\zeta_{ij} := \zeta_j a_{ij}, \ \ \eta^{\mathrm{e}}_{ij} := \eta^{\mathrm{e}}_j a_{ij}, \ \text{ and } \ \eta^{\mathrm{x}}_{ij} := \eta^{\mathrm{x}}_j a_{ij}, \ \ \forall i \in [I], \ \forall j \in [J]. \tag{B.10a}$$

We equivalently linearize these bilinear equalities as

$$\zeta_j = \sum_{i=1}^{I+1} \zeta_{ij}, \ \ \eta^{\text{x}}_j = \sum_{i=1}^{I+1} \eta^{\text{x}}_{ij}, \ \ \eta^{\text{e}}_j = \sum_{i=1}^{I+1} \eta^{\text{e}}_{ij}, \ \ \forall j \in [J], \tag{B.10b}$$

$$0 \le \zeta_{ij} \le K a_{ij}, \ \ -K a_{ij} \le \eta^{\text{x}}_{ij} \le K a_{ij}, \ \ -K a_{ij} \le \eta^{\text{e}}_{ij} \le K a_{ij}. \tag{B.10c}$$

To see the equivalence, on the one hand, we notice that constraints (B.10b) follow from (B.10a) and (3.16a). Similarly, constraints (B.10c) follow from (B.10a) and the facts that a_{ij} are binary and $\zeta_j \ge 0$. On the other hand, constraints (B.10b) and (3.16a) imply that $\zeta_{ij} = \zeta_j$ if $a_{ij} = 1$, and constraints (B.10c) imply that $\zeta_{ij} = 0$ if $a_{ij} = 0$. We hence have $\zeta_{ij} = \zeta_j a_{ij}$. Likewise, we establish $\eta^{\text{e}}_{ij} = \eta^{\text{e}}_j a_{ij}$, $\eta^{\text{x}}_{ij} = \eta^{\text{x}}_j a_{ij}$, and hence constraints (B.10a). It follows that constraints (3.17d)–(3.17e) can be recast as $\theta_i \ge \eta^{\text{x}}_{ij} + \sum_{\ell=1}^{j-1} \max\{\zeta_{i\ell} + \eta^{\text{e}}_{i\ell}, \eta^{\text{x}}_{i\ell}\} + \sum_{\ell=j+1}^{J} (\zeta_{i\ell} + \eta^{\text{e}}_{i\ell})$, $\theta_i \ge \phi_i + \sum_{\ell=1}^{J} (\zeta_{i\ell} + \eta^{\text{e}}_{i\ell})$, plus (B.10b)–(B.10c).

Second, we linearize constraint (3.17c) by claiming that

$$\begin{aligned} & \Big(\theta_i + c^{\text{y}} y^{\text{L}}_i + g_i(y^{\text{L}}_i)\lambda_i + \sum_{\ell=1}^{y^{\text{U}}_i - y^{\text{L}}_i} \big(\delta_{i\ell}\nu_{i\ell} + c^{\text{y}} v_{i\ell}\big)\Big) o_i \\ = \ & \theta_i + c^{\text{y}} y^{\text{L}}_i o_i + g_i(y^{\text{L}}_i)\lambda_i + \sum_{\ell=1}^{y^{\text{U}}_i - y^{\text{L}}_i} \big(\delta_{i\ell}\nu_{i\ell} + c^{\text{y}} v_{i\ell}\big), \end{aligned} \tag{B.10d}$$

which holds valid if $\theta_i = 0$, $\lambda_i = 0$, and $v_{i\ell} = 0$ for all $\ell \in [y^{\text{U}}_i - y^{\text{L}}_i]$ whenever $o_i = 0$ (note that variables $\nu_{i\ell}$ also vanish in this case because $\nu_{i\ell} = \lambda_i v_{i\ell}$). To this end, we incorporate constraints

$$-c^{\text{x}} o_i \le \lambda_i \le 0 \tag{B.10e}$$

because $\lambda_i \in [-c^{\text{x}}, 0]$ without loss of optimality by Proposition 7. This guarantees that $o_i = 0$ implies $\lambda_i = 0$. Additionally, we incorporate constraints

$$v_{i1} \le o_i. \tag{B.10f}$$

Then, $o_i = 0$ implies $v_{i1} = 0$ and hence $v_{i\ell} = 0$ for all $\ell \in [y^{\text{U}}_i - y^{\text{L}}_i]$ due to constraints (3.8f). Furthermore, $o_i = 0$ implies that $a_{ij} = 0$ for all $j \in [J]$ by constraints (3.16b). It follows that $\zeta_{ij} = \eta^{\text{e}}_{ij} = \eta^{\text{x}}_{ij} = 0$ for all $j \in [J]$. It remains to ensure that $\phi_i = 0$ whenever

$o_i = 0$. To that end, we replace constraints (3.12a) with

$$\phi_i \geq -c^{\mathrm{e}} y_i^{\mathrm{L}} o_i - y_i^{\mathrm{L}} \lambda_i - \sum_{\ell=1}^{y_i^{\mathrm{U}} - y_i^{\mathrm{L}}} (\nu_{i\ell} + c^{\mathrm{e}} v_{i\ell}). \tag{B.10g}$$

Indeed, if $o_i = 0$ then $\lambda_i = 0$ by constraints (B.10e) and $\nu_{i\ell} = v_{i\ell} = 0$ for all $\ell \in [y_i^{\mathrm{U}} - y_i^{\mathrm{L}}]$ by constraints (B.10f). Therefore, constraint (3.17c) is equivalently linearized through equality (B.10d) and incorporating constraints (B.10e)–(B.10g). This completes the proof. □

B.15 Symmetry Breaking Inequalities for the (OPD) Model

We consider two types of symmetry among integer solutions. First, suppose that there are 2 pools and 4 units. The following two unit assignments lead to symmetric integer solutions: (i) assigning all units to pool 1 and no unit to pool 2 (i.e., $a_{1j} = 1$ and $a_{2j} = 0$ for all $j \in [4]$) and (ii) assigning all units to pool 2 and no unit to pool 1 (i.e., $a_{1j} = 0$ and $a_{2j} = 1$ for all $j \in [4]$). We call this "pool symmetry." To break this symmetry, we designate that all open pools have smaller indices than the closed ones. This designation breaks the pool symmetry because the above case (ii) is now prohibited. Accordingly, we add the following inequalities to the (OPD) formulation:

$$o_i \geq o_{i+1}, \;\; \forall i \in [I-1].$$

Second, the following two unit assignments also lead to symmetric integer solutions: (iii) assigning units 1 and 3 to pool 1 and units 2 and 4 to pool 2 (i.e., $a_{11} = 1 - a_{12} = a_{13} = 1 - a_{14} = 1$ and $1 - a_{21} = a_{22} = 1 - a_{23} = a_{24} = 1$) and (iv) assigning units 1 and 3 to pool 2 and units 2 and 4 to pool 1 (i.e., $1 - a_{11} = a_{12} = 1 - a_{13} = a_{14} = 1$ and $a_{21} = 1 - a_{22} = a_{23} = 1 - a_{24} = 1$). We call this "unit symmetry." To break this symmetry, we rank the pools based on the smallest unit index in each pool. That is, we designate that the smallest unit index in pool i is smaller than that in pool $i+1$ for all $i \in [I-1]$, if both pools are opened. This designation breaks the unit symmetry because the above case (iv) is now prohibited. Accordingly, we add the following inequalities to the (OPD) formulation:

$$\sum_{\ell=1}^{j-1} a_{i\ell} \geq a_{(i+1)j}, \;\; \forall i \in [I-1], \; \forall j \in [J].$$

B.16 Input Parameters of the Instance Reported in Figure 3.5, Section 3.6.4

Table B.1: Input parameters of the representative instance

unit j	unit 1	unit 2	unit 3	unit 4	unit 5	unit 6	unit 7
μ_{j1}	11.42	6.34	17.73	19.15	19.69	15.67	14.84
sd_j	5.05	4.03	6.44	17.06	16.39	16.52	15.92
A_j^{u}	0.97	0.98	0.98	0.61	0.75	0.67	0.67
A_i^{p}	0.99						
costs	$c^{\text{w}} = 100$, $c^{\text{y}} = 130$, $c^{\text{x}} = 400$, $c^{\text{e}} = 50$						

BIBLIOGRAPHY

[1] United States Government Accountability Office, "Critical Infrastructure Protection: Protecting the Electric Grid from Geomagnetic Disturbances," Dec 2018.

[2] Shapiro, A., Dentcheva, D., and Ruszczyński, A., *Lectures on stochastic programming: modeling and theory*, SIAM, 2014.

[3] Birge, J. R. and Louveaux, F., *Introduction to stochastic programming*, Springer Science & Business Media, 2011.

[4] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A., *Robust optimization*, Vol. 28, Princeton University Press, 2009.

[5] Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A., "Adjustable robust solutions of uncertain linear programs," *Math. Program., Ser. A*, Vol. 99, 2004, pp. 351–376.

[6] Scarf, H., "A min-max solution of an inventory problem," *Studies in the mathematical theory of inventory and production*, 1958.

[7] Žáčková, J., "On minimax solutions of stochastic linear programming problems," *Časopis pro pěstování matematiky*, Vol. 91, No. 4, 1966, pp. 423–430.

[8] Dupačová, J., "Optimization under exogenous and endogenous uncertainty," *University of West Bohemia in Pilsen*, 2006.

[9] Shapiro, A. and Kleywegt, A., "Minimax analysis of stochastic problems," *Optimization Methods and Software*, Vol. 17, No. 3, 2002, pp. 523–542.

[10] Delage, E. and Ye, Y., "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, Vol. 58, No. 3, 2010, pp. 595–612.

[11] Esfahani, P. M. and Kuhn, D., "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, Vol. 171, No. 1-2, 2018, pp. 115–166.

[12] Rahimian, H. and Mehrotra, S., "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.

[13] Pflug, G. and Wozabal, D., "Ambiguity in portfolio selection," *Quantitative Finance*, Vol. 7, No. 4, 2007, pp. 435–442.

[14] Blanchet, J. and Murthy, K., "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, Vol. 44, No. 2, 2019, pp. 565–600.

[15] Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G., "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, Vol. 59, No. 2, 2013, pp. 341–357.

[16] Bayraksan, G. and Love, D. K., "Data-driven stochastic programming using phi-divergences," *The Operations Research Revolution*, INFORMS, 2015, pp. 1–19.

[17] Jiang, R. and Guan, Y., "Data-driven chance constrained stochastic program," *Mathematical Programming*, Vol. 158, No. 1-2, 2016, pp. 291–327.

[18] Rahimian, H., Bayraksan, G., and Homem-de Mello, T., "Identifying effective scenarios in distributionally robust stochastic programs with total variation distance," *Mathematical Programming*, Vol. 173, No. 1-2, 2019, pp. 393–430.

[19] Jiang, R. and Guan, Y., "Risk-averse two-stage stochastic program with distributional ambiguity," *Operations Research*, Vol. 66, No. 5, 2018, pp. 1390–1405.

[20] Erdoğan, E. and Iyengar, G., "Ambiguous chance constrained problems and robust optimization," *Mathematical Programming*, Vol. 107, No. 1-2, 2006, pp. 37–61.

[21] Zhao, C. and Guan, Y., "Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics," *Available on Optimization Online*, 2015.

[22] Pichler, A. and Xu, H., "Quantitative stability analysis for minimax distributionally robust risk optimization," *Mathematical Programming*, 2017, pp. 1–31.

[23] Gao, R. and Kleywegt, A. J., "Distributionally robust stochastic optimization with Wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[24] Zhao, C. and Guan, Y., "Data-driven risk-averse stochastic optimization with Wasserstein metric," *Operations Research Letters*, Vol. 46, No. 2, 2018, pp. 262–267.

[25] Hanasusanto, G. A. and Kuhn, D., "Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls," *Operations Research*, Vol. 66, No. 3, 2018, pp. 849–869.

[26] Noyan, N., Rudolf, G., and Lejeune, M., "Distributionally Robust Optimization with Decision-Dependent Ambiguity Set," *Available Online:* http://www.optimization-online.org/DB_FILE/2018/09/6821.pdf, 2018.

[27] Luo, F. and Mehrotra, S., "Distributionally Robust Optimization with Decision Dependent Ambiguity Sets," *Available Online:* https://arxiv.org/abs/1806.09215, 2018.

[28] Hanasusanto, G. A., Roitch, V., Kuhn, D., and Wiesemann, W., "A distributionally robust perspective on uncertainty quantification and chance constrained programming," *Mathematical Programming*, Vol. 151, No. 1, 2015, pp. 35–62.

[29] Dupačová, J., "The minimax approach to stochastic programming and an illustrative application," *Stochastics: An International Journal of Probability and Stochastic Processes*, Vol. 20, No. 1, 1987, pp. 73–88.

[30] Goh, J. and Sim, M., "Distributionally robust optimization and its tractable approximations," *Operations Research*, Vol. 58, No. 4-part-1, 2010, pp. 902–917.

[31] Bertsimas, D., Doan, X. V., Natarajan, K., and Teo, C.-P., "Models for minimax stochastic linear optimization problems with risk aversion," *Mathematics of Operations Research*, Vol. 35, No. 3, 2010, pp. 580–602.

[32] Wiesemann, W., Kuhn, D., and Sim, M., "Distributionally robust convex optimization," *Operations Research*, Vol. 62, No. 6, 2014, pp. 1358–1376.

[33] Royset, J. O. and Wets, R. J.-B., "Variational theory for optimization under stochastic ambiguity," *SIAM Journal on Optimization*, Vol. 27, No. 2, 2017, pp. 1118–1149.

[34] Berg, B. P., Denton, B. T., Erdogan, S. A., Rohleder, T., and Huschka, T., "Optimal booking and scheduling in outpatient procedure centers," *Computers & Operations Research*, Vol. 50, 2014, pp. 24–37.

[35] Begen, M. A. and Queyranne, M., "Appointment scheduling with discrete random durations," *Mathematics of Operations Research*, Vol. 36, No. 2, 2011, pp. 240–257.

[36] Kim, K. and Mehrotra, S., "A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management," *Operations Research*, Vol. 63, No. 6, 2015, pp. 1431–1451.

[37] Green, L. V., Savin, S., and Savva, N., ""Nursevendor Problem": personnel staffing in the presence of endogenous absenteeism," *Management Science*, Vol. 59, No. 10, 2013, pp. 2237–2256.

[38] Denton, B. and Gupta, D., "A sequential bounding approach for optimal appointment scheduling," *IIE Transactions*, Vol. 35, No. 11, 2003, pp. 1003–1016.

[39] Kaandorp, G. C. and Koole, G., "Optimal outpatient appointment scheduling," *Health Care Management Science*, Vol. 10, No. 3, 2007, pp. 217–229.

[40] Chen, R. R. and Robinson, L. W., "Sequencing and scheduling appointments with potential call-in patients," *Production and Operations Management*, Vol. 23, No. 9, 2014, pp. 1522–1538.

[41] Kong, Q., Lee, C.-Y., Teo, C.-P., and Zheng, Z., "Scheduling arrivals to a stochastic service delivery system using copositive cones," *Operations Research*, Vol. 61, No. 3, 2013, pp. 711–726.

[42] Mak, H.-Y., Rong, Y., and Zhang, J., "Appointment Scheduling with Limited Distributional Information," *Management Science*, Vol. 61, No. 2, 2015, pp. 316–334.

[43] Kong, Q., Li, S., Liu, N., Teo, C.-P., and Yan, Z., "Appointment scheduling under time-dependent patient no-show behavior," *Management Science*, Forthcoming, 2019.

[44] Jiang, R., Shen, S., and Zhang, Y., "Integer programming approaches for appointment scheduling with random no-shows and service durations," *Operations Research*, Vol. 65, No. 6, 2017, pp. 1638–1656.

[45] Davis, A., Mehrotra, S., Holl, J., and Daskin, M. S., "Nurse staffing under demand uncertainty to reduce costs and enhance patient safety," *Asia-Pacific Journal of Operational Research*, Vol. 31, No. 01, 2014, pp. 1450005.

[46] Ahmed, S., King, A. J., and Parija, G., "A multi-stage stochastic integer programming approach for capacity expansion under uncertainty," *Journal of Global Optimization*, Vol. 26, No. 1, 2003, pp. 3–24.

[47] Park, H., Baldick, R., and Morton, D. P., "A stochastic transmission planning model with dependent load and wind forecasts," *IEEE Transactions on Power Systems*, Vol. 30, No. 6, 2015, pp. 3003–3011.

[48] Jiang, R., Wang, J., and Guan, Y., "Robust unit commitment with wind power and pumped storage hydro," *IEEE Transactions on Power Systems*, Vol. 27, No. 2, 2011, pp. 800–810.

[49] Jabr, R. A., "Robust transmission network expansion planning with uncertain renewable generation and loads," *IEEE Transactions on Power Systems*, Vol. 28, No. 4, 2013, pp. 4558–4567.

[50] Yao, L., Wang, X., Duan, C., Guo, J., Wu, X., and Zhang, Y., "Data-driven distributionally robust reserve and energy scheduling over Wasserstein balls," *IET Generation, Transmission & Distribution*, Vol. 12, No. 1, 2017, pp. 178–189.

[51] Duan, C., Fang, W., Jiang, L., Yao, L., and Liu, J., "Distributionally robust chance-constrained approximate AC-OPF with Wasserstein metric," *IEEE Transactions on Power Systems*, Vol. 33, No. 5, 2018, pp. 4924–4936.

[52] Zhu, R., Wei, H., and Bai, X., "Wasserstein metric based distributionally robust approximate framework for unit commitment," *IEEE Transactions on Power Systems*, Vol. 34, No. 4, 2019, pp. 2991–3001.

[53] Zhang, Y., Shen, Z.-J. M., and Song, S., "Distributionally Robust Optimization of Two-Stage Lot-Sizing Problems," *Production and Operations Management*, Vol. 25, No. 12, 2016, pp. 2116–2131.

[54] Zhao, C. and Jiang, R., "Distributionally robust contingency-constrained unit commitment," *IEEE Transactions on Power Systems*, Vol. 33, No. 1, 2017, pp. 94–102.

[55] Velloso, A., Pozo, D., and Street, A., "Distributionally Robust Transmission Expansion Planning: a Multi-scale Uncertainty Approach," *arXiv preprint arXiv:1810.05212*, 2018.

[56] Nguyen, V. A., Kuhn, D., and Esfahani, P. M., "Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator," *arXiv preprint arXiv:1805.07194*, 2018.

[57] Lee, C. and Mehrotra, S., "A distributionally-robust approach for finding support vector machines," *Manuscript, available at http://www. optimization-online. org/DB_ HTML/2015/06/4965. html*, 2015.

[58] Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D., "Distributionally robust logistic regression," *Advances in Neural Information Processing Systems*, 2015, pp. 1576–1584.

[59] Bailey, N. T., "A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1952, pp. 185–199.

[60] Fetter, R. B. and Thompson, J. D., "Patients' waiting time and doctors' idle time in the outpatient setting," *Health Services Research*, Vol. 1, No. 1, 1966, pp. 66.

[61] Ho, C.-J. and Lau, H.-S., "Minimizing total cost in scheduling outpatient appointments," *Management Science*, Vol. 38, No. 12, 1992, pp. 1750–1764.

[62] Gurvich, I., Luedtke, J., and Tezcan, T., "Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach," *Management Science*, Vol. 56, No. 7, 2010, pp. 1093–1115.

[63] Shen, S. and Wang, J., "Stochastic modeling and approaches for managing energy footprints in cloud computing service," *Service Science*, Vol. 6, No. 1, 2014, pp. 15–33.

[64] Cayirli, T. and Veral, E., "Opatient scheduling in health care: a review of literature," *Production and Operations Management*, Vol. 12, No. 4, 2003, pp. 519–549.

[65] Kall, P. and Wallace, S. W., *Stochastic Programming*, Springer, 1994.

[66] Shapiro, A., Dentcheva, D., and Ruszczynski, A., *Lectures on stochastic programming: modeling and theory*, Vol. 16, SIAM, 2014.

[67] Erdogan, S. A. and Denton, B., "Dynamic appointment scheduling of a stochastic server with uncertain demand," *INFORMS Journal on Computing*, Vol. 25, No. 1, 2013, pp. 116–132.

[68] Gupta, D. and Denton, B., "Appointment scheduling in health care: Challenges and opportunities," *IIE Transactions*, Vol. 40, No. 9, 2008, pp. 800–819.

[69] Bertsimas, D. and Popescu, I., "Optimal inequalities in probability theory: A convex optimization approach," *SIAM Journal on Optimization*, Vol. 15, No. 3, 2005, pp. 780–804.

[70] Jiang, R., Ryu, M., and Xu, G., "Data-Driven Distributionally Robust Appointment Scheduling over Wasserstein Balls," *arXiv preprint arXiv:1907.03219*, 2019.

[71] Wozabal, D., "A framework for optimization under ambiguity," *Annals of Operations Research*, Vol. 193, No. 1, 2012, pp. 21–47.

[72] Brahimi, M. and Worthington, D., "Queueing models for out-patient appointment systems–A case study," *Journal of the Operational Research Society*, Vol. 42, No. 9, 1991, pp. 733–746.

[73] Ahmadi-Javid, A., Jalali, Z., and Klassen, K. J., "Outpatient appointment systems in healthcare: A review of optimization studies," *European Journal of Operational Research*, Vol. 258, No. 1, 2017, pp. 3–34.

[74] Cardoen, B., Demeulemeester, E., and Beliën, J., "Operating room planning and scheduling: A literature review," *European Journal of Operational Research*, Vol. 201, No. 3, 2010, pp. 921–932.

[75] Ge, D., Wan, G., Wang, Z., and Zhang, J., "A note on appointment scheduling with piecewise linear cost functions," *Mathematics of Operations Research*, Vol. 39, No. 4, 2013, pp. 1244–1251.

[76] Begen, M. A., Levi, R., and Queyranne, M., "A sampling-based approach to appointment scheduling," *Operations Research*, Vol. 60, No. 3, 2012, pp. 675–681.

[77] Hassin, R. and Mendel, S., "Scheduling arrivals to queues: A single-server model with no-shows," *Management Science*, Vol. 54, No. 3, 2008, pp. 565–572.

[78] LaGanga, L. R. and Lawrence, S. R., "Clinic overbooking to improve patient access and increase provider productivity," *Decision Sciences*, Vol. 38, No. 2, 2007, pp. 251–276.

[79] Macario, A., "Truth in scheduling: is it possible to accurately predict how long a surgical case will last?" 2009.

[80] Mittal, S., Schulz, A. S., and Stiller, S., "Robust appointment scheduling," *LIPIcs-Leibniz International Proceedings in Informatics*, Vol. 28, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[81] Rowse, E. L., *Robust optimisation of operating theatre schedules*, Ph.D. thesis, Cardiff University, 2015.

[82] Bertsimas, D., Gupta, V., and Kallus, N., "Robust sample average approximation," *Mathematical Programming*, 2017, pp. 1–66.

[83] Denton, B., Viapiano, J., and Vogl, A., "Optimization of surgery sequencing and scheduling decisions under uncertainty," *Health Care Management Science*, Vol. 10, No. 1, 2007, pp. 13–24.

[84] He, S., Sim, M., and Zhang, M., "Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach," *Management Science*, Forthcoming, 2019.

[85] Mak, H.-Y., Rong, Y., and Zhang, J., "Sequencing appointments for service systems using inventory approximations," *Manufacturing & Service Operations Management*, Vol. 16, No. 2, 2014, pp. 251–262.

[86] LaGanga, L. R. and Lawrence, S. R., "Appointment overbooking in health care clinics to improve patient service and clinic performance," *Production and Operations Management*, Vol. 21, No. 5, 2012, pp. 874–888.

[87] Liu, N. and Ziya, S., "Panel size and overbooking decisions for appointment-based services under patient no-shows," *Production and Operations Management*, Vol. 23, No. 12, 2014, pp. 2209–2223.

[88] Zacharias, C. and Pinedo, M., "Appointment scheduling with no-shows and overbooking," *Production and Operations Management*, Vol. 23, No. 5, 2014, pp. 788–801.

[89] Zacharias, C. and Pinedo, M., "Managing customer arrivals in service systems with multiple identical servers," *Manufacturing & Service Operations Management*, Vol. 19, No. 4, 2017, pp. 639–656.

[90] Fournier, N. and Guillin, A., "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, Vol. 162, No. 3-4, 2015, pp. 707–738.

[91] Burer, S., "On the Copositive Representation of Binary and Continuous Nonconvex Quadratic Programs," *Mathematical Programming Series A*, Vol. 120, No. 2, September 2009, pp. 479–495.

[92] Burer, S., "Copositive Programming," *Handbook of Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications*, edited by M. Anjos and J. Lasserre, International Series in Operational Research and Management Science, Springer, 2011, pp. 201–218.

[93] Burer, S., "A gentle, geometric introduction to copositive optimization," *Mathematical Programming*, Vol. 151, No. 1, 2015, pp. 89–116.

[94] Dür, M., "Copositive programming–a survey," *Recent advances in optimization and its applications in engineering*, Springer, 2010, pp. 3–20.

[95] Bomze, I. M. and De Klerk, E., "Solving standard quadratic optimization problems via linear, semidefinite and copositive programming," *Journal of Global Optimization*, Vol. 24, No. 2, 2002, pp. 163–185.

[96] De Klerk, E. and Pasechnik, D. V., "Approximation of the stability number of a graph via copositive programming," *SIAM Journal on Optimization*, Vol. 12, No. 4, 2002, pp. 875–892.

[97] Lasserre, J. B., "Convexity in semialgebraic geometry and polynomial optimization," *SIAM Journal on Optimization*, Vol. 19, No. 4, 2009, pp. 1995–2014.

[98] Parrilo, P. A., *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, California Institute of Technology, 2000.

[99] Xu, G. and Hanasusanto, G. A., "Improved Decision Rule Approximations for Multi-Stage Robust Optimization via Copositive Programming," *arXiv preprint arXiv:1808.06231*, 2018.

[100] ApS, M., *The MOSEK optimization toolbox for MATLAB manual. Version 8.0.*, 2016.

[101] Toh, K.-C., Todd, M. J., and Tütüncü, R. H., "SDPT3a MATLAB software package for semidefinite programming, version 1.3," *Optimization methods and software*, Vol. 11, No. 1-4, 1999, pp. 545–581.

[102] Grötschel, M., Lovász, L., and Schrijver, A., "The ellipsoid method and its consequences in combinatorial optimization," *Combinatorica*, Vol. 1, No. 2, 1981, pp. 169–197.

[103] Welton, J. M., "Hospital nursing workforce costs, wages, occupational mix, and resource utilization," *Journal of Nursing Administration*, Vol. 41, No. 7/8, 2011, pp. 309–314.

[104] Tevington, P., "Mandatory nurse-patient ratios," *Medsurg Nursing*, Vol. 20, No. 5, 2011, pp. 265.

[105] California Department of Health, "California RN Staffing Ratio Law," 2004.

[106] of Health, V. D., "Safe Patient Care (Nurse to Patient and Midwife to Patient Ratios) Act," 2015.

[107] Bard, J. F., "Nurse scheduling models," *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

[108] Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., and Bakker, P. J., "Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS," *Health Systems*, Vol. 1, No. 2, 2012, pp. 129–175.

[109] Kortbeek, N., Braaksma, A., Burger, C. A., Bakker, P. J., and Boucherie, R. J., "Flexible nurse staffing based on hourly bed census predictions," *International Journal of Production Economics*, Vol. 161, 2015, pp. 167–180.

[110] Bam, M., Zhang, Z., Denton, B., Van Oyen, M. P., and Duck, M., "Planning Models for Skills-Sensitive Surgical Nurse Staffing," *Available Onlin:* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959005, 2017.

[111] Gaynor, M. and Anderson, G. F., "Uncertain demand, the structure of hospital costs, and the cost of empty hospital beds," *Journal of Health Economics*, Vol. 14, No. 3, 1995, pp. 291–317.

[112] Boutsioli, Z., "Demand variability, demand uncertainty and hospital costs: a selective survey of the empirical literature," *Global Journal of Health Science*, Vol. 2, No. 1, 2010, pp. 138.

[113] Maass, K. L., Liu, B., Daskin, M. S., Duck, M., Wang, Z., Mwenesi, R., and Schapiro, H., "Incorporating nurse absenteeism into staffing with demand uncertainty," *Health Care Management Science*, Vol. 20, No. 1, 2017, pp. 141–155.

[114] of Labor Statistics, B., "Absences from work of employed full-time wage and salary workers by occupation and industry," 2016.

[115] Wang, W.-Y. and Gupta, D., "Nurse absenteeism and staffing strategies for hospital inpatient units," *Manufacturing & Service Operations Management*, Vol. 16, No. 3, 2014, pp. 439–454.

[116] Inman, R. R., Blumenfeld, D. E., and Ko, A., "Cross-training hospital nurses to reduce staffing costs," *Health Care Management Review*, Vol. 30, No. 2, 2005, pp. 116–125.

[117] Ryu, M. and Jiang, R., "Nurse Staffing under Absenteeism: A Distributionally Robust Optimization Approach," *arXiv preprint arXiv:1909.09875*, 2019.

[118] Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., and De Boeck, L., "Personnel scheduling: A literature review," *European Journal of Operational Research*, Vol. 226, No. 3, 2013, pp. 367–385.

[119] Kao, E. P. and Tung, G. G., "Aggregate nursing requirement planning in a public health care delivery system," *Socio-Economic Planning Sciences*, Vol. 15, No. 3, 1981, pp. 119–127.

[120] Brusco, M. and Showalter, M., "Constrained nurse staffing analysis," *Omega*, Vol. 21, No. 2, 1993, pp. 175–186.

[121] Trivedi, V. M., "A mixed-integer goal programming model for nursing service budgeting," *Operations Research*, Vol. 29, No. 5, 1981, pp. 1019–1034.

[122] Yang, K.-K., Webster, S., and Ruben, R. A., "An evaluation of worker cross training and flexible workdays in job shops," *IIE Transactions*, Vol. 39, No. 7, 2007, pp. 735–746.

[123] Maenhout, B. and Vanhoucke, M., "An integrated nurse staffing and scheduling analysis for longer-term nursing staff allocation problems," *Omega*, Vol. 41, No. 2, 2013, pp. 485–499.

[124] Stewart, B. D., Webster, D. B., Ahmad, S., and Matson, J. O., "Mathematical models for developing a flexible workforce," *International Journal of Production Economics*, Vol. 36, No. 3, 1994, pp. 243–254.

[125] Kao, E. P. and Queyranne, M., "Budgeting costs of nursing in a hospital," *Management Science*, Vol. 31, No. 5, 1985, pp. 608–621.

[126] Campbell, G. M., "A two-stage stochastic program for scheduling and allocating cross-trained workers," *Journal of the Operational Research Society*, Vol. 62, No. 6, 2011, pp. 1038–1047.

[127] Lu, S. F. and Lu, L. X., "Do mandatory overtime laws improve quality? Staffing decisions and operational flexibility of nursing homes," *Management Science*, Vol. 63, No. 11, 2016, pp. 3566–3585.

[128] Slaugh, V. W., Scheller-Wolf, A. A., and Tayur, S. R., "Consistent Staffing for Long-Term Care Through On-Call Pools," *Production and Operations Management*, 2018.

[129] Easton, F. F., "Service completion estimates for cross-trained workforce schedules under uncertain attendance and demand," *Production and Operations Management*, Vol. 23, No. 4, 2014, pp. 660–675.

[130] Bertsimas, D., Sim, M., and Zhang, M., "Adaptive distributionally robust optimization," *Management Science*, Vol. 65, No. 2, 2018, pp. 604–618.

[131] Hanasusanto, G. A. and Kuhn, D., "Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls," *Operations Research*, Vol. 66, No. 3, 2018, pp. 849–869.

[132] Kong, Q., Lee, C.-Y., Teo, C.-P., and Zheng, Z., "Scheduling arrivals to a stochastic service delivery system using copositive cones," *Operations Research*, Vol. 61, No. 3, 2013, pp. 711–726.

[133] Ang, J., Meng, F., and Sun, J., "Two-stage stochastic linear programs with incomplete information on uncertainty," *European Journal of Operational Research*, Vol. 233, No. 1, 2014, pp. 16–22.

[134] McCormick, G. P., "Computability of global solutions to factorable nonconvex programs: Part IConvex underestimating problems," *Mathematical programming*, Vol. 10, No. 1, 1976, pp. 147–175.

[135] Nemhauser, G. L. and Wolsey, L. A., *Integer and Combinatorial Optimization*, John Wiley & Sons, 1999.

[136] Jordan, W. C. and Graves, S. C., "Principles on the benefits of manufacturing process flexibility," *Management Science*, Vol. 41, No. 4, 1995, pp. 577–594.

[137] Wang, X. and Zhang, J., "Process flexibility: A distribution-free bound on the performance of k-chain," *Operations Research*, Vol. 63, No. 3, 2015, pp. 555–571.

[138] Chen, X., Zhang, J., and Zhou, Y., "Optimal sparse designs for process flexibility via probabilistic expanders," *Operations Research*, Vol. 63, No. 5, 2015, pp. 1159–1176.

[139] Chen, X., Ma, T., Zhang, J., and Zhou, Y., "Optimal design of process flexibility for general production systems," *Operations Research*, Vol. 67, No. 2, 2019, pp. 516–531.

[140] Boteler, D., Pirjola, R., and Nevanlinna, H., "The effects of geomagnetic disturbances on electrical systems at the Earth's surface," *Advances in Space Research*, Vol. 22, No. 1, 1998, pp. 17–27.

[141] Gannon, J., Swidinsky, A., and Xu, Z., *Geomagnetically Induced Currents from the Sun to the Power Grid*, Vol. 246, John Wiley & Sons, 2019.

[142] Bolduc, L., "GIC observations and studies in the Hydro-Québec power system," *Journal of Atmospheric and Solar-Terrestrial Physics*, Vol. 64, No. 16, 2002, pp. 1793–1802.

[143] Cauley, G. and Lauby, M., "High-Impact Low-Frequency Event Risk to the North American Bulk Power System," *NERC, Atlanta, GA, USA, Tech. Rep*, 2010.

[144] Guide, A., "Computing Geomagnetically-Induced Current in the Bulk-Power System, NERC. NERC," .

[145] Bolduc, L., Granger, M., Pare, G., Saintonge, J., and Brophy, L., "Development of a DC current-blocking device for transformer neutrals," *IEEE Transactions on power delivery*, Vol. 20, No. 1, 2005, pp. 163–168.

[146] Zhu, H. and Overbye, T. J., "Blocking device placement for mitigating the effects of geomagnetically induced currents," *IEEE Trans. on Power Systems*, Vol. 30, No. 4, 2014, pp. 2081–2089.

[147] Lu, M., Nagarajan, H., Yamangil, E., Bent, R., Backhaus, S., and Barnes, A., "Optimal transmission line switching under geomagnetic disturbances," *IEEE Trans. on Power Systems*, Vol. 33, No. 3, 2017, pp. 2539–2550.

[148] Kazerooni, M. and Overbye, T. J., "Transformer Protection in Large-Scale Power Systems During Geomagnetic Disturbances Using Line Switching," *IEEE Trans. on Power Systems*, Vol. 33, No. 6, 2018, pp. 5990–5999.

[149] Lu, M., Eksioglu, S. D., Mason, S. J., Bent, R., and Nagarajan, H., "Distributionally Robust Optimization for a Resilient Transmission Grid During Geomagnetic Disturbances," *arXiv preprint:1906.04139*, 2019.

[150] Zeng, B. and Zhao, L., "Solving two-stage robust optimization problems using a column-and-constraint generation method," *Operations Research Letters*, Vol. 41, No. 5, 2013, pp. 457–461.

[151] Ryu, M., Nagarajan, H., and Bent, R., "Algorithms for Mitigating the Effect of Uncertain Geomagnetic Disturbances in Electric Grids," *Power Systems Computation Conference (PSCC)*, IEEE, 2020, pp. 1–7.

[152] Kocuk, B., Dey, S. S., and Sun, X. A., "New formulation and strong MISOCP relaxations for AC optimal transmission switching problem," *IEEE Trans. on Power Systems*, Vol. 32, No. 6, 2017, pp. 4161–4170.

[153] Woodroffe, J. R., Morley, S., Jordanova, V., Henderson, M., Cowee, M., and Gjerloev, J., "The latitudinal variation of geoelectromagnetic disturbances during large (Dst$\leq$-100 nT) geomagnetic storms," *Space Weather*, Vol. 14, No. 9, 2016, pp. 668–681.

[154] Lu, M., Nagarajan, H., Bent, R., Eksioglu, S. D., and Mason, S. J., "Tight piecewise convex relaxations for global optimization of optimal power flow," *Power Systems Computation Conference (PSCC)*, IEEE, 2018, pp. 1–7.

[155] Horton, R., Boteler, D., Overbye, T. J., Pirjola, R., and Dugan, R. C., "A test case for the calculation of geomagnetically induced currents," *IEEE Transactions on Power Delivery*, Vol. 27, No. 4, 2012, pp. 2368–2373.

[156] Birchfield, A. B., Gegner, K. M., Xu, T., Shetye, K. S., and Overbye, T. J., "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *IEEE Transactions on Power Systems*, Vol. 32, No. 2, 2016, pp. 1502–1510.

[157] Shapiro, A., "On duality theory of conic linear problems," *Semi-infinite programming*, Springer, 2001, pp. 135–165.

[158] Hanasusanto, G. A., Roitch, V., Kuhn, D., and Wiesemann, W., "Ambiguous joint chance constraints under mean and dispersion information," *Operations Research*, Vol. 65, No. 3, 2017, pp. 751–767.

[159] Burer, S., "Copositive programming," *Handbook on semidefinite, conic and polynomial optimization*, Springer, 2012, pp. 201–218.

[160] Zangwill, W. I., "A deterministic multi-period production scheduling model with backlogging," *Management Science*, Vol. 13, No. 1, 1966, pp. 105–119.

[161] Zangwill, W. I., "A backlogging model and a multi-echelon model of a dynamic economic lot size production systema network approach," *Management Science*, Vol. 15, No. 9, 1969, pp. 506–527.

[162] Faigle, U. and Kern, W., "On the core of ordered submodular cost games," *Mathematical Programming*, Vol. 87, No. 3, 2000, pp. 483–499.

[163] Ben-Tal, A. and Nemirovski, A., *Lectures on Modern Convex Optimization: analysis, algorithms, and engineering applications*, Vol. 2, SIAM, 2001.

[164] Ruszczyński, A. P., *Nonlinear optimization*, Vol. 13, Princeton University Press, 2006.