Data-Driven Reliable Facility Location Design

Hao Shen,[a] Mengying Xue,[b]* Zuo-Jun Max Shen[c,d]

[a]School of Business, Renmin University of China, Beijing, China, 100872; [b]International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, Anhui, China, 230026; [c]Faculty of Engineering, The University of Hong Kong, Hong Kong 999077, China; [d]Faculty of Business and Economics, The University of Hong Kong, Hong Kong 999077, China

*Corresponding author

Contact: shenhao@rmbs.ruc.edu.cn (HS); xmy2021@ustc.edu.cn (MX); maxshen@hku.hk (Z-JMS);

We study the reliable (uncapacitated) facility location (RFL) problem in a data-driven environment where historical observations of random demands and disruptions are available. Owing to the combinatorial optimization nature of the RFL problem and the mixed-binary randomness of parameters therein, the state-of-the-art RFL models applied to the data-driven setting either suggest overly conservative solutions, or become computationally prohibitive for large- or even moderate-size problems. In this paper, we address the RFL problem by presenting an innovative prescriptive model aiming to balance solution conservatism with computational efficiency. In particular, our model selects facility locations to minimize the fixed costs plus the expected operating costs approximated by a tractable data-driven estimator, which equals to a probabilistic upper bound on the intractable Kolmogorov distributionally robust optimization estimator. The solution of our model is obtained by solving a mixed-integer linear program that does *not* scale in the training data size. Our approach is proved to be asymptotically optimal, and offers a theoretical guarantee for its out-of-sample performance in situations with limited data. In addition, we discuss the adaptation of our approach when facing data with covariate information. Numerical results demonstrate that our model significantly outperforms several important RFL models with respect to both in-sample and out-of-sample performances as well as computational efficiency.

Key words: facility location, supply chain disruption, data-driven optimization, prescriptive analytics

History:

1. Introduction

Since 2020, we have been witnessing the frequent and drastic turbulence on supply chains caused by natural disasters and social disorders. In February, 2021, Winter Storm Uri hit North America, paralyzed the electricity system of Texas, and disrupted supply chains in many industries such as grocery retail. During the storm, many people grasped for the most basic of needs, as the extreme weather disturbed the balance between supply and demand of essential goods. Notably, grocery retailers were unable to receive shipments of food (Kapadia 2021), and some of them had to throw out refrigerated items due to days of power outages (Speare-Cole 2021). On the other hand, stores were overwhelmed by soaring demand, as families rushed to stock up on the household essentials to survive in the dire situation (Ledbetter 2021). After experiencing the extreme weather, it is more crucial than ever for companies to revamp their supply chains to be better-prepared for the upheavals in both supply and demand, under the risk of potential disruptions.

The reliable facility location (RFL) problem is one of the most important reliable supply chain design problems, and has been extensively studied in the area of operations research and management science (e.g., Snyder and Daskin 2005, Cui et al. 2010, Lu et al. 2015, Li et al. 2022). The RFL problem focuses on the design of both reliable and cost-efficient supply networks to strike a balance between normal operational efficiency and emergency service quality, in the face of random disruptions and demands. In particular, normal operations such as locating facilities and shipping products incur initial setup (or fixed) costs as well as routine transportation costs. When facility failures occur due to disruptions, customers either are reassigned to survived facilities that may require higher transportation costs, or experience service outages, which incur penalty costs reflecting the service level. The RFL problem seeks the optimal contingency plan by determining facility locations and customer assignments before the realization of disruptions and demands, in order to minimize the expected overall network cost.

Addressing RFL problems in a data-driven setting is an important research question. As the joint probability distribution of random disruptions and demands is usually unknown in practice, the distributional information exploited from data is critical to contingency planning. For example, before Hurricane Patricia hit Mexico's coastline in 2015, IBM had received the forecast made by the AI Watson using huge troves of weather and location data, and avoided catastrophic damages by rerouting the inbound shipments of a production center in Guadalajara to the backups in the US (Banker 2016). Nonetheless, most of the existing RFL models–namely, stochastic models (e.g., Snyder and Daskin 2005, Cui et al. 2010, Xie et al. 2019) and robust optimization models (e.g., An et al. 2014, Lu et al. 2015, Li et al. 2022)–often become restrictive in a data-driven setting, as they rely on *exact* partial or even complete information of the data-generating distribution, which is difficult to obtain in practice. Moreover, for computational tractability, demand uncertainty is either ignored (e.g., Snyder and Daskin 2005, Cui et al. 2010, Lu et al. 2015) or only considered in a stylized form (e.g., An et al. 2014, Li et al. 2022), which can be limiting in real-world applications.

The key to data-driven RFL problem is to reasonably approximate the unknown true expected network cost in the objective function of the RFL problem, based on historical data of disruptions and demands. A natural approach to this end is the well-known sample average approximation (SAA), which estimates the true expected network cost by its sample mean. The SAA approach is shown to be asymptotically optimal under some mild assumptions, as its optimal value and optimal solutions will converge almost surely to their counterparts of the underlying RFL problem as the size of the data set tends to infinity (Kleywegt et al. 2002). However, one might need an impractically large data set to reach such optimality. Even if a sufficiently large data set were available, solving the SAA RFL model exactly would be prohibitive, as it is equivalent to solving a mixed-integer linear program (MILP) that scales in the size of the data set. In the event that only limited historical data can be acquired, the SAA tends to significantly underestimate the optimal value of the RFL problem (Xie 2020), and display a poor out-of-sample performance.

To improve the performance of SAA, a recent approach is to apply the (type-∞) Wasserstein distributionally robust optimization (DRO) framework to the RFL problem (Xie 2020). The objective of this framework is to minimize the expected network cost under a probability distribution adversarially chosen from a neighborhood of the empirical distribution[1]. The neighborhood, termed as the Wasserstein ambiguity set, consists of probability distributions whose distance from the empirical distribution, measured by the type-∞ Wasserstein metric, is below a given threshold. The Wasserstein DRO approach not only is asymptotically optimal, but also improves the SAA estimation of the optimal value of the RFL problem with limited data. However, the pitfalls of this approach are twofold. For one, it retains the same level of scalability of the SAA approach, which leads to high computational cost when facing a large data set. Moreover, our results show that in situations with relatively small data sets, the solution of the Wasserstein DRO RFL model can be overly conservative, and thus too expensive to implement in practice (Section 5.1.1).

In this paper, we investigate an alternative data-driven approach to the RFL problem that aims to achieve significant improvements on existing RFL models in both computational efficiency and statistical performance. Our approach stems from the Kolmogorov DRO framework (Luo and Mehrotra 2020), which is close in spirit to the Wasserstein DRO but uses the Kolmogorov metric (instead of the Wasserstein one) to measure the distance between probability distributions. However, when applied to the RFL problem directly, the Kolmogorov DRO yields an intractable estimator of the true expected network cost (Proposition 1). By contrast, we approximate the true expected network cost by using a tractable probabilistic upper bound (PUB) on the Kolmogorov DRO estimator. Our model is guaranteed to be asymptotically optimal, and admits an MILP formulation with attractive scalability. We also provide a theoretical finite sample performance guarantee that can help improve the out-of-sample performance of our data-driven RFL design. Below we summarize our main results in greater detail:

- **A novel data-driven approach to the RFL problem:** We address the RFL problem in a generic setting where both disruptions and demands are random and governed by an unknown joint distribution. By using historical observations of disruptions and demands, we construct a novel PUB estimator to approximate the objective function of the RFL problem (Definition 2). In particular, the PUB estimator is derived by exploiting the unique structure of a cumulative distribution function (CDF)-based representation of the Kolmogorov DRO estimator, where the decision variable is the CDF of probability distributions, and the objective function is a linear functional of the CDF (Lemma 2). In contrast to the intractable Kolmogorov DRO estimator, the PUB estimator has an analytical form which can be evaluated in polynomial time (Lemma 3). The data-driven RFL design is then obtained by minimizing the PUB estimator.

- **Solution algorithm:** We show that the PUB estimator is supermodular as a function of the facility location decisions (Lemma 3). This leads to an MILP formulation of our data-driven RFL model where the number of decision variables and constraints does *not* scale in the number of data points (Theorem 1). We develop a constraint generation algorithm that can solve the MILP formulation to optimality within practically reasonable time even when facing large networks and large data sets. Numerical results show that our algorithm achieves superior computational efficiency, compared with previous RFL models (Section 5.2).

- **Theoretical justification:** We present theoretical performance guarantees for our data-driven RFL model. First, our approach is guaranteed to be asymptotically optimal (Theorem 2). That is, under some mild assumptions, the optimal value and optimal solutions of our model will converge almost surely to their counterparts of the RFL problem as more data are obtained. Second, in situations where only limited data are available, the optimal value of our model will upper bound the true expected network cost incurred by the data-driven RFL design with high probability (Theorem 3). In addition, we provide a practical approach that can efficiently determine the parameters for our model to yield cost-efficient solutions.

- **An extension that incorporates covariate information:** Our data-driven RFL model can be further extended to incorporate covariate information, which, if available, can indicate valuable information of network uncertainty. To our knowledge, the RFL problem with covariate information has drawn little attention from literature. We construct a generalized PUB estimator based on an event-wise extension of the Kolmogorov DRO estimator, and provide a generalized version of our previous theoretical results (Theorem 4). We believe that our approach offers a promising step toward new data-driven approaches to a broader class of supply chain design problems.

1.1. Related Literature

Coping with supply chain disruptions has drawn unprecedented attention from academia, while a rich body of mitigation strategies has been proposed in the literature. Snyder et al. (2016) provide a comprehensive review of the literature on supply chain disruptions and mitigation strategies. According to their classification, existing disruption mitigation models fall into four major categories: (i) inventory models (e.g., Qi et al. 2009), (ii) sourcing and demand flexibility models (e.g., Shen et al. 2019), (iii) RFL models, and (iv) game models (e.g., Yang et al. 2009). This paper only reviews the classical results and recent progresses in the RFL literature; for a detailed survey of the other three categories, please refer to Snyder et al. (2016). In addition, we provide a brief review of the related DRO studies.

We focus on two fast-growing streams of RFL models in the literature, namely, stochastic models and robust optimization models. Stochastic models assume to have the knowledge of the *complete* disruption distribution, i.e., either independent known marginal distributions (e.g., Snyder and Daskin 2005, Cui et al. 2010), or mass probabilities for all possible scenarios (Xie et al. 2019), to minimize the expectation of the overall network cost. In contrast, most of the robust optimization models leverage limited *marginal* information of disruptions, such as the maximum number of simultaneous disruptions (e.g., Church and Scaparra 2007, An et al. 2014, Cheng et al. 2018, 2021), or moment information of disruption probabilities (Lu et al. 2015, Li et al. 2022). Their objective is usually minimizing the overall cost incurred in the worst-case scenario, given partial information of the disruption probability distribution.

A major limitation of the stochastic model is that the formulation relies heavily on a specific structure of the disruption distribution. For example, the implicit formulation models, termed by Lu et al. (2015), use the so-called "transitional probability" equations as constraints to denote the probability of a customer to be served by a facility; see, for example, Snyder and Daskin (2005), Berman et al. (2007), Cui et al. (2010), Chen et al. (2011), Shen et al. (2011), Aboolian et al. (2013), Zhang et al. (2016). However, the "transitional probability" equations only hold by assuming independent disruptions. Efforts have been made to relax this assumption. Xie et al. (2019) convert the facility network under correlated disruptions to an equivalent virtual station network with independent station disruptions. This enables an implicit formulation with transitional equations on the virtual station network. Nonetheless, if the disruptions are not subject to the "local" correlations (i.e., correlations restricted within several local areas), the transformed virtual network will potentially include an exponential number of supporting stations (Xie et al. 2015), leading to a computationally prohibitive model. Another method to tackle non-independent disruptions is the continuous approximation approach (Li and Ouyang 2010, Lim et al. 2013). However, this approach may imply opposite or deviated results compared with discrete location models, which are more suitable for capturing real-world supply chain design problems (Lu et al. 2015). Furthermore, the stochastic model requires an accurate estimation of the joint distribution of disruptions and demands, which can be challenging to obtain in practice.

Robust optimization models are designed for situations where only limited disruption information is available, and concerned with the worst-case disruption scenario. For instance, the interdiction median models, e.g., Church and Scaparra (2007), Liberatore et al. (2012), An et al. (2014), Cheng et al. (2018, 2021), design reliable networks under the adversarial attacks disrupting up to a certain number of facilities; the moment-based DRO model (Lu et al. 2015) and its variant (Li et al. 2022) introduce ambiguity sets constructed by given marginal or higher-order moments of the disruption probability distribution. Nonetheless, limited discussions are provided to address the over-conservatism of their solutions, which is a major criticism of robust optimization models. Efforts have been made by Lu et al. (2015), who consider a weighted-average objective function combining a robust optimization model and a traditional facility location model with no disruption, and Li et al. (2022), who use higher-order distributional information such as cross-moments to downsize the ambiguity set. However, the effectiveness of their approaches is only tested numerically, whereas few theoretical performance guarantees are provided for their applications to data-driven problems.

Recently, a data-driven RFL model built by Xie (2020) applies the type-∞ Wasserstein DRO framework, whose performance guarantees can be found in Bertsimas et al. (2022). Xie (2020) derives an MILP formulation of the data-driven RFL model with variables growing linearly with the size of data. This formulation, analogous to the SAA approach, will become computationally expensive as when applied to large networks and data sets. In addition, it is known that the conservatism of the type-∞ Wasserstein DRO framework can be adjusted by varying the size of the ambiguity set. However, as is shown in our numerical results, when applied to the RFL problem, there are situations where the type-∞ Wasserstein DRO model may fail to achieve low conservatism if a certain performance guarantee is required.

The type-∞ Wasserstein DRO falls into the category of the metric-based DRO, where the objective is to find a solution that achieves some goal under the worst-case distribution, chosen from a set of distributions close to a reference distribution (e.g., the empirical distribution) with respect to a statistical metric. Alternative statistical metrics in the metric-based DRO literature include Prokhorov metric (Erdoğan and Iyengar 2006), $\phi-$divergence (Ben-Tal et al. 2013, Hu and Hong 2013, Bayraksan and Love 2015, Jiang and Guan 2016), type-p Wasserstein metric (Pflug and Wozabal 2007, Wozabal 2012, Esfahani and Kuhn 2018) with $p \in [1, \infty)$, Kolmogorov metric (Lim et al. 2006, Bertsimas et al. 2018a,b, Luo and Mehrotra 2020) among others. However, when applied to the NP-hard RFL problem, the major bottleneck of the metric-based DRO is the prohibitive computational cost required to achieve a satisfactory performance. In the context of the RFL problem, the type-∞ Wasserstein DRO is possibly state-of-the-art in computational efficiency among all metric-based DRO approaches studied to date.

By contrast, we present a data-driven approach outside the scope of, albeit closely related to, the metric-based DRO to address the RFL problem. Our approach not only retains (and even improves) the attractive statistical performance of the Wasserstein DRO, but also offers significant computational benefits. In addition, our approach allows the flexibility of efficiently incorporating covariate information to aid data-driven RFL design, which is a novel feature contributed to the RFL literature.

1.2. Organization and Notation

The remainder of this paper is organized as follows. In Section 2, we present our data-driven approach to the RFL problem, including preliminary results related to the Kolmogorov DRO, and the derivation of our PUB estimator. In Section 3, we analyze the structural properties of our model, and provide a constraint generation solution algorithm. In Section 4, we prove finite sample guarantees and asymptotic optimality of our data-driven approach. In Section 5, we conduct a thorough comparison of the numerical performance between our model and several important RFL models. In Section 6, we present an extension of our model by incorporating the covariate information. In Section 7, we summarize this paper. The proofs of all statements, several supplementary results and the psuedocode of algorithms are provided in the online appendix.

Notation. We denote by $\mathbb{R}$, $\mathbb{R}_-$, $\mathbb{R}_+$ and $\mathbb{N}$ the sets of real numbers, non-positive real numbers, non-negative real numbers, and positive integers, respectively. Let $[n]$ denote the finite set $\{0,1,\dots,n\}$ for a non-negative integer n; in particular, let $\mathbb{B}$ the binary set $[1]=\{0,1\}$. If Ω is a finite set, then 2^Ω denotes the collection of all subsets of Ω. Let $\mathbb{P}\{\cdot\}$, $\mathbb{F}(\cdot)$, and $\mathbb{E}^{\mathbb{P}}[\cdot]$ denote a probability measure, a CDF, and the expected value function under distribution $\mathbb{P}$. Let $\mathbb{I}\{\cdot\}$ be the usual indicator function. Denote $\mathbf{1}_n$ and $\mathbf{0}_n$ as the n-dimensional vector of all ones and the one of all zeros; for simplicity, we will omit the subscript n if their dimensions are clear according to the context. Denote $\boldsymbol{e}_n^i$ as the i-th standard unit vector in $\mathbb{R}^n$; that is, an n-dimensional vector whose i-th component is one while all others are zeros. We use $a\wedge b$ (or $a\vee b$) to denote the minimum (or the maximum) value between $a,b\in\mathbb{R}$, and $\boldsymbol{a}\wedge\boldsymbol{b}$ (or $\boldsymbol{a}\vee\boldsymbol{b}$) to denote the component-wise minimum (or the component-wise maximum) vector between $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^n$.

2. Data-Driven Reliable Facility Location Model

We consider the problem of opening a subset of J candidate facilities to serve I customers. Let $\mathcal{I}$ denote the customer set, and $\mathcal{J}$ the candidate facility set. For all $i \in \mathcal{I}$ and $j \in \mathcal{J}$, let f_j denote the fixed cost of opening facility j, $\boldsymbol{f} = (f_j)_{j\in\mathcal{J}}$, and d_{ij} the distance between customer i and facility j (or the unit transportation cost incurred by using facility j to serve customer). Define $\boldsymbol{x} = (x_i)_{i\in\mathcal{I}}$ as the facility location decision, where $x_j = 1$ if facility j is opened, and $x_j = 0$ otherwise.

The uncertainty in our problem is twofold. First, the facilities are under the risk of random disruptions. Denote by $\widetilde{\boldsymbol{\xi}} = (\widetilde{\xi}_j)_{j\in\mathcal{J}}$ the disruption scenario, where $\widetilde{\xi}_j = 0$ if facility j is disrupted, and $\widetilde{\xi}_j = 1$ if it is operational/not disrupted. Second, customer demands are random variables that may be correlated with disruptions. For notational convenience, let $-\widetilde{\boldsymbol{\zeta}} = (-\widetilde{\zeta}_i)_{i\in\mathcal{I}}$ denote the demand vector, where each random variable $\widetilde{\zeta}_i$ denotes the "negative" demand supported on a bounded subset of $\mathbb{R}_-$. Define $\widetilde{\boldsymbol{\omega}} := (\widetilde{\boldsymbol{\zeta}}, \widetilde{\boldsymbol{\xi}})$ as the network state vector, which is then supported on a mixed-binary set $\Omega \subseteq \mathbb{R}^I_- \times \mathbb{B}^J$.

Given a facility location design $\boldsymbol{x} \in \mathbb{B}^J$ and a realization $\boldsymbol{\omega} \in \Omega$ of the network state, either each customer is served by an opened and operational facility, or the customer's demand is lost. Following the convention in the RFL literature (i.e., Cui et al. 2010, Lu et al. 2015), we use an "emergency" facility $\bar{j} \notin \mathcal{J}$ to denote an outside option for customers, and assign $\bar{j}$ to customer i if and only if customer i's demand is lost. Let $d_{i\bar{j}}$ denote the penalty cost of losing per unit of customer i's demand, and assume $d_{i\bar{j}} \geq d_{ij}$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. In addition, we assume that facility $\bar{j}$ is always opened and operational, i.e., $x_{\bar{j}} \equiv 1$ and $\widetilde{\xi}_{\bar{j}} \equiv 1$, and let $f_{\bar{j}} = 0$ and $\overline{\mathcal{J}} = \mathcal{J} \cup \{\bar{j}\}$. Then the operating cost, defined as the total transportation and penalty cost to serve customers, is given by

$$\phi(\boldsymbol{x}, \boldsymbol{\omega}) := \min_{\boldsymbol{y}} \left\{ \sum_{i\in\mathcal{I}} \sum_{j\in\overline{\mathcal{J}}} (-\zeta_i) d_{ij} y_{ij} \;\middle|\; \begin{array}{ll} \sum_{j\in\overline{\mathcal{J}}} y_{ij} = 1 & \forall i \in \mathcal{I} \\ y_{ij} \leq x_j \xi_j & \forall i \in \mathcal{I}, \forall j \in \mathcal{J} \\ \boldsymbol{y} \in \mathbb{R}_+^{I(J+1)} & \end{array} \right\} = \sum_{i\in\mathcal{I}} (-\zeta_i) \min_{j\in\overline{\mathcal{J}}: x_j\xi_j = 1} d_{ij}. \tag{1}$$

Let $\mathbb{P}^\star \in \mathcal{P}(\Omega)$ be the true probability distribution of the network state vector $\widetilde{\boldsymbol{\omega}}$, where $\mathcal{P}(\Omega)$ denotes the set of probability distributions supported on Ω. The RFL problem is defined as the following stochastic program that minimizes the fixed cost of locating facilities plus the expected operating cost under distribution $\mathbb{P}^\star$:

$$z^\star := \min_{\boldsymbol{x}\in\mathbb{B}^J} \left\{ Z^\star(\boldsymbol{x}) := \boldsymbol{f}^\intercal \boldsymbol{x} + \Phi^\star(\boldsymbol{x}) \;\middle|\; \Phi^\star(\boldsymbol{x}) := \mathbb{E}^{\mathbb{P}^\star} \left[\phi(\boldsymbol{x}, \widetilde{\boldsymbol{\omega}}) \right] \right\}. \tag{RFL}$$

Nonetheless, in most cases of practical interest, we do not precisely know the distribution $\mathbb{P}^\star$, so that Problem (RFL) cannot be solved exactly. Instead, we often have past realizations of the network state vector $\widetilde{\boldsymbol{\omega}}$ which contain partial information of the true distribution $\mathbb{P}^\star$. To be specific, assume that we have N samples $\widehat{\boldsymbol{\omega}}^n \in \Omega$, $n \in \mathcal{N} := \{1, 2, \ldots, N\}$, that are independently drawn from the true distribution $\mathbb{P}^\star$. Let $\widehat{\mathcal{S}}_N := \{\widehat{\boldsymbol{\omega}}^n\}_{n\in\mathcal{N}} \subseteq \Omega$ denote the training data set. We then consider to obtain a data-driven RFL design for Problem (RFL) by solving the following problem

$$\widehat{z}_N := \min_{\boldsymbol{x}\in\mathbb{B}^J} \boldsymbol{f}^\intercal \boldsymbol{x} + \widehat{\Phi}_N(\boldsymbol{x}), \tag{DD-RFL}$$

where $\widehat{\Phi}_N(\boldsymbol{x})$ is an estimator constructed from $\widehat{\mathcal{S}}_N$ to approximate the true expected operating cost function $\Phi^\star(\boldsymbol{x})$.

In this paper, we propose a novel estimator $\widehat{\Phi}_N(\boldsymbol{x})$ of the true expected operating cost function such that the optimal value $\widehat{z}_N$ and any optimal solution $\widehat{\boldsymbol{x}}_N$ of Problem (DD-RFL) satisfy the following conditions:

(C1) *Computational efficiency:* For any given $\boldsymbol{x} \in \mathbb{B}^J$, the value of $\widehat{\Phi}_N(\boldsymbol{x})$ can be obtained in time that is linear in the size of the data set N and polynomial in the size of the problem. In addition, the operating cost function $\widehat{\Phi}_N(\boldsymbol{x})$ is equipped with favorable structural properties, which facilitate a computationally efficient solution algorithm.

(C2) *Asymptotic optimality:* Provided that $N \to \infty$ as more data are obtained, the optimal value $\widehat{z}_N$ and an optimal solution $\widehat{\boldsymbol{x}}_N$ of Problem (DD-RFL) respectively converge to the optimal value $z^\star$ and an optimal solution $\boldsymbol{x}^\star$ of Problem (RFL) almost surely.

(C3) *Finite sample guarantee:* With probability at least $1-\beta$, a data set $\widehat{\mathcal{S}}_N$ is sampled such that $\Phi^\star(\widehat{\boldsymbol{x}}_N) \le \widehat{z}_N$, for the optimal value $\widehat{z}_N$ and any optimal solution $\widehat{\boldsymbol{x}}_N$ of Problem (DD-RFL).

For the combinatorial optimization problem (DD-RFL), the computational efficiency in practice can substantially affect the empirical performance of the obtained solutions. For instance, consider an estimator with which the optimal solution $\widehat{\boldsymbol{x}}_N$ and the optimal value $\widehat{z}_N$ of Problem (DD-RFL) satisfy properties (C2) and (C3), but are difficult to obtain in a practically reasonable time. Then in practice, we can only expect to derive some heuristic solution $\widehat{\boldsymbol{x}}'_N$ of Problem (DD-RFL) and its objective value $\widehat{z}'_N$, which do not necessarily provide good approximation of the true optimal solution $\boldsymbol{x}^\star$ and optimal value $z^\star$ (even though $\widehat{\boldsymbol{x}}_N$ and $\widehat{z}_N$ do). Therefore, incorporating property (C1) in addition to properties (C2) and (C3) is critical to the development of data-driven approaches for the RFL problem, and is one of the distinguishable features of our work.

The remainder of this section will focus on the derivation of $\widehat{\Phi}_N(\boldsymbol{x})$. We discuss structural properties, complexity results and the solution algorithm related to (C1) in Section 3. The rigorous analysis of properties (C2) and (C3) will be provided in Section 4. To begin the derivation of $\widehat{\Phi}_N(\boldsymbol{x})$, we first present several preliminary results concerning the Kolmogorov DRO RFL model that are closely related to our approach.

2.1. Preliminaries

In the context of the RFL problem, the Kolmogorov DRO model selects a facility location design that minimizes the expected total cost under the distribution chosen by an adversary from the Kolmogorov ambiguity set. Here the Kolmogorov ambiguity set contains all distributions close to the empirical distribution with respect to the Kolmogorov metric, whose definition is presented as follows.

Definition 1 (Kolmogorov Metric). Let $\mathcal{P}(\Theta)$ be the set of all probability distributions supported on $\Theta \subseteq \mathbb{R}^m$. The Kolmogorov metric $\mathscr{K} : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \to [0,1]$ is defined by

$$\mathscr{K}(\mathbb{P}_1, \mathbb{P}_2) := \sup_{\boldsymbol{\theta} \in \mathbb{R}^m} |\mathbb{P}_1\{(-\infty, \boldsymbol{\theta}]\} - \mathbb{P}_2\{(-\infty, \boldsymbol{\theta}]\}| = \|\mathbb{F}_1 - \mathbb{F}_2\|_\infty,$$

for all distributions $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\Theta)$, where $\mathbb{F}_k(\boldsymbol{\theta}) := \mathbb{P}_k\{(-\infty, \boldsymbol{\theta}]\}$ denotes the cumulative distribution function (CDF) of $\mathbb{P}_k$ for $k \in \{1,2\}$.

The intuition is that the Kolmogorov metric measures the distance between two distributions by using the largest absolute difference between their CDFs across all values in their support.

Let $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{n\in\mathcal{N}} \delta_{\widehat{\boldsymbol{\omega}}^n} \in \mathcal{P}(\Omega)$ denote the empirical probability distribution, where $\delta_{\widehat{\boldsymbol{\omega}}^n}$ denotes the unit mass on $\widehat{\boldsymbol{\omega}}^n$. Then the Kolmogorov DRO model generates data-driven RFL designs by solving

$$\widehat{z}_N^{\mathrm{K}} := \min_{\boldsymbol{x}\in\mathbb{B}^J} \left\{ \widehat{Z}_N^{\mathrm{K}}(\boldsymbol{x}) := \boldsymbol{f}^\intercal \boldsymbol{x} + \widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x}) \,\middle|\, \widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x}) := \sup_{\mathbb{P}\in\widehat{\mathcal{P}}_N^{\mathrm{K}}} \mathbb{E}^{\mathbb{P}}[\phi(\boldsymbol{x}, \widetilde{\boldsymbol{\omega}})] \right\}, \tag{KDRO-RFL}$$

where $\widehat{\mathcal{P}}_N^{\mathrm{K}} := \{\mathbb{P} \in \mathcal{P}(\Omega) : \mathscr{K}(\mathbb{P}, \widehat{\mathbb{P}}_N) \le \epsilon_N\}$ is the Kolmogorov ambiguity set with $\epsilon_N \in [0,1]$, and $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$ is the *Kolmogorov DRO estimator* of the true operating cost function. The Kolmogorov metric enjoys the following measure concentration result, which implies a finite sample performance guarantee for Problem (KDRO-RFL)[2].

Lemma 1 **(Measure Concentration, Naaman 2021)**. *For any given $\epsilon > 0$, $N \ge 1$, let $\widehat{\Theta}_N = \{\widehat{\boldsymbol{\theta}}^n\}_{n\in\{1,2,\dots,N\}} \subseteq \mathbb{R}^m$ be a set of m-dimensional i.i.d. samples, each governed by distribution $\mathbb{P}$. Let $\widehat{\mathbb{P}}_N$ be the empirical distribution produced by $\widehat{\Theta}_N$. Then we have*

$$\mathbb{P}_{\widehat{\Theta}_N}\left\{\mathscr{K}(\mathbb{P}, \widehat{\mathbb{P}}_N) > \epsilon\right\} \le m(N+1)\exp(-2N\epsilon^2). \tag{2}$$

Unfortunately, Problem (KDRO-RFL) is challenging to solve exactly. Indeed, even evaluating $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$ for any given $\boldsymbol{x}$ is computationally demanding. The following proposition provides a method to evaluate $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$ when customer demands are continuous and bounded.

Proposition 1. *Suppose $\widetilde{\boldsymbol{\zeta}}$ is continuous and $\Omega \subseteq [\underline{\boldsymbol{\zeta}}, \mathbf{0}] \times \mathbb{B}^J$ for some $\underline{\boldsymbol{\zeta}} \in \mathbb{R}_-^I$. Then $\widehat{\Phi}_N^K(\boldsymbol{x})$ equals the optimal value of a linear program with $\mathcal{O}(N^J 2^I)$ decision variables and constraints.*

The details of the formulation for $\widehat{\Phi}^{\mathrm{K}}_N(\boldsymbol{x})$ are provided in the proof of Proposition 1. To demonstrate the scale of this formulation, consider a small-scale problem instance with 10 customers/candidate locations and 10 samples. Then $\widehat{\Phi}^{\mathrm{K}}_N(\boldsymbol{x})$ is the optimal value of a linear program with approximately 10^{13} variables and constraints. Recognizing the intractability of $\widehat{\Phi}^{\mathrm{K}}_N(\boldsymbol{x})$, we are thus motivated to develop a tractable surrogate for $\widehat{\Phi}^{\mathrm{K}}_N(\boldsymbol{x})$ that preserves the attractive properties of the Kolmogorov DRO.

2.2. Our Approach

Our approach stems from an alternative formulation of $\widehat{\Phi}^{\mathrm{K}}_N(\boldsymbol{x})$. We first observe from Definition 1 that the Kolmogorov metric can be represented in a cleaner format in terms of CDFs rather than the probability measures. Thus we consider to reformulate $\widehat{\Phi}^{\mathrm{K}}_N(\boldsymbol{x})$ as an equivalent optimization problem where decision variables are CDFs. For expositional convenience, we define the following notation. For any subset $\mathcal{X}\subseteq\mathcal{J}$ and vector $\boldsymbol{\xi}\in\mathbb{R}^J$, define $\boldsymbol{\xi}(\mathcal{X}):=(\xi_j)_{j\in\mathcal{X}}$. Let $\mathcal{F}(\Omega)$ denotes the set of CDFs supported on Ω. Given a CDF $\mathbb{F}\in\mathcal{F}(\Omega)$, $i\in\mathcal{I}$, and $\mathcal{X}\subseteq\mathcal{J}$, let $\mathbb{F}^{i,\mathcal{X}}$ be the marginal CDF of random variables $\widetilde{\zeta}_i$ and $\widetilde{\boldsymbol{\xi}}(\mathcal{X})$. Then we have $\mathbb{P}\in\mathcal{P}^{\mathrm{K}}_N$ if and only if $\mathbb{F}$, the CDF of $\mathbb{P}$, is a member of

$$\widehat{\mathcal{F}}^{\mathrm{K}}_N:=\left\{\mathbb{F}\in\mathcal{F}(\Omega)\,\middle|\,\left\|\mathbb{F}-\widehat{\mathbb{F}}_N\right\|_\infty\le\epsilon_N\right\},\tag{3}$$

where $\widehat{\mathbb{F}}_N$ denotes the empirical CDF.

Next we reformulate the objective function $\mathbb{E}^{\mathbb{P}}[\phi(\boldsymbol{x},\widetilde{\boldsymbol{\omega}})]$ as a function of $\mathbb{F}$. Let $\mathcal{X}:=\{j\in\mathcal{J}:x_j=1\}$ and $\overline{\mathcal{X}}:=\mathcal{X}\cup\{\bar{j}\}$. By (1), the operating cost $\phi(\boldsymbol{x},\boldsymbol{\omega})$ can be decomposed as $\phi(\boldsymbol{x},\boldsymbol{\omega})=\sum_{i\in\mathcal{I}}\phi_i(\boldsymbol{x},\boldsymbol{\omega})$, where $\phi_i(\boldsymbol{x},\boldsymbol{\omega}):=(-\zeta_i)\min_{j\in\overline{\mathcal{X}}:\xi_j=1}d_{ij}$ represents the operating cost incurred by serving customer i. For each $i\in\mathcal{I}$, we rearrange the facilities in $\overline{\mathcal{X}}$ as a sequence of facilities $j_{(0)},j_{(1)},\ldots,j_{(|\mathcal{X}|)}$ in an increasing order of the distance to customer i; that is, $d_{ij_{(0)}}\le d_{ij_{(1)}}\le\cdots\le d_{ij_{(|\mathcal{X}|)}}(=d_{i\bar{j}})$. We then define a random index $\widetilde{r}^*_i:=\min\{r\in[|\mathcal{X}|]:\widetilde{\xi}_{j_{(r)}}=1\}$, so that $j_{(\widetilde{r}^*_i)}$ denotes the operational facility in $\overline{\mathcal{X}}$ that is the closest to customer i. Thus we have

$$\phi_i(\boldsymbol{x},\widetilde{\boldsymbol{\omega}})=(-\widetilde{\zeta}_i)d_{ij_{(\widetilde{r}^*_i)}},\tag{4}$$

which leads to the following description of the expected value $\mathbb{E}^{\mathbb{P}}[\phi_i(\boldsymbol{x},\widetilde{\boldsymbol{\omega}})]$.

Lemma 2 **(CDF-Based Representation of Expected Operating Cost)**. *Let $\boldsymbol{x}$ be any given facility location design, $\mathcal{X}:=\{j\in\mathcal{J}:x_j=1\}$ and $\overline{\mathcal{X}}:=\mathcal{X}\cup\{\bar{j}\}$. For each $i\in\mathcal{I}$, let $j_{(0)},j_{(1)},\ldots,j_{(|\mathcal{X}|)}$ be a permutation of facilities in $\overline{\mathcal{X}}$ in an increasing order of the distance to customer i. Then, for any given distribution $\mathbb{P}\in\mathcal{P}(\Omega)$ with CDF $\mathbb{F}\in\mathcal{F}(\Omega)$, we have*

$$\mathbb{E}^{\mathbb{P}}[\phi_i(\boldsymbol{x},\widetilde{\boldsymbol{\omega}})]=\sum_{r=0}^{|\mathcal{X}|}(\Delta d)_{ir}\int_{\Omega_i}\mathbb{F}^{i,\mathcal{X}^{r-1}_i}(\zeta,\mathbf{0}_r)d\zeta\tag{5}$$

where $(\Delta d)_{ir} := d_{ij_{(r)}} - d_{ij_{(r-1)}}$, $d_{ij_{(-1)}} := 0$, $\mathcal{X}_i^r := \{j_{(0)}, j_{(1)}, \ldots, j_{(r)}\}$, $\mathcal{X}_i^{-1} := \emptyset$, *and* $\Omega_i := [\underline{\zeta}_i, 0]$ *is the support of* $\widetilde{\zeta}_i$.

An important implication of Lemma 2 is that for any given subset of facilities assigned to a customer, the optimal way of determining their backup levels only depends on the distances (or the transportation costs) from the customer to these facilities, *regardless* of their disruption risks. In particular, if facilities in $\mathcal{X} \cup \{\bar{j}\}$ are assigned to customer i, the optimal service rule suggests to assign the r-th closest facility $j_{(r)}$ to serve the customer only if no closer facilities $j_{(0)}, j_{(1)}, \ldots, j_{(r-1)}$ are operational. The optimal expected operating cost is then equal to $\sum_{r=0}^{|\mathcal{X}|} \sum_{\zeta_i} (-\zeta_i) d_{ij_{(r)}} \mathbb{P}\left\{\widetilde{\zeta}_i = \zeta_i, \widetilde{r}_i^* = r\right\}$, whose CDF-based form is given by the right hand side of Equation (5). This is referred to as the distance-based service rule, whose optimality was only proved under the assumption of independent disruptions and deterministic demands (Snyder and Daskin 2005, Cui et al. 2010), and seemed to fail if certain correlation is incorporated into disruptions, e.g., under the "worst case disruption distribution" proposed by Lu et al. (2015). Nonetheless, Lemma 2 extends the optimality of the distance-based service rule to the case where disruptions and demands can be generally correlated, and thus reconciles the seemingly controversial results in previous RFL works. In Online Appendix EC.2, we show the optimality of the distance-based service rule under correlated disruptions as considered in Lu et al. (2015).

Having obtained the CDF-based ambiguity set (3) and the CDF-based objective function (Lemma 2), we shall now present a CDF-based formulation for $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$, which further implies a tractable upper bound. In particular, we have

$$\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x}) = \sup_{\mathbb{F} \in \widehat{\mathcal{F}}_N^{\mathrm{K}}} \sum_{i \in \mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} (\Delta d)_{ir} \int_{\Omega_i} \mathbb{F}^{i, \mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r) \mathrm{d}\zeta \tag{6}$$

$$\leq \sum_{i \in \mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} (\Delta d)_{ir} \int_{\Omega_i} \left((\widehat{\mathbb{F}}_N^{i, \mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r) + \epsilon_N) \wedge 1 \right) \mathrm{d}\zeta, \tag{7}$$

where the CDF-based formulation (6) of the Kolmogorov estimator follows from (3) and Lemma 2, and the inequality holds because $\|\mathbb{F} - \widehat{\mathbb{F}}_N\|_\infty \leq \epsilon_N$ and $\|\mathbb{F}\|_\infty \leq 1$ for all $\mathbb{F} \in \widehat{\mathcal{F}}_N^{\mathrm{K}}$, and $(\Delta d)_{ir} \geq 0$. Indeed, the equality between (6) and (7) does not necessarily hold, as there might not exist a CDF in $\widehat{\mathcal{F}}_N^{\mathrm{K}}$ whose marginal in $(\widetilde{\zeta}_i, \widetilde{\boldsymbol{\xi}}(\mathcal{X}_i^r))$ equals $(\widehat{\mathbb{F}}_N^{i, \mathcal{X}_i^r} + \epsilon_N) \wedge 1$ for all i and r. One can thus view (7) as the upper bound obtained by solving a relaxation of Problem (6) which replaces the constraint $\mathbb{F} \in \mathcal{F}(\Omega)$ with a looser one $\|\mathbb{F}\|_\infty \leq 1$.

Nonetheless, the upper bound (7) is not suitable for data-driven models, as the support Ω_i of each $\widetilde{\zeta}_i$ is usually unknown. By replacing each Ω_i in (7) with a data-driven support $[\widehat{\zeta}_i^{(1)}, \widehat{\zeta}_i^{(N)}]$, where

$\widehat{\zeta}_i^{(1)} := \min\{\widehat{\zeta}_i^n : n \in \{1,2,\ldots,N\}\}$ and $\widehat{\zeta}_i^{(N)} := \max\{\widehat{\zeta}_i^n : n \in \{1,2,\ldots,N\}\}$, we obtain a surrogate upper bound as follows:

$$
\begin{aligned}
(7) \approx & \sum_{i\in\mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} (\Delta d)_{ir} \int_{\widehat{\zeta}_i^{(1)}}^{\widehat{\zeta}_i^{(N)}} \left(\widehat{\mathbb{F}}_N^{i,\mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r) + \epsilon_N \right) \wedge 1 \; \mathrm{d}\zeta && (8) \\
\leq & \sum_{i\in\mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} (\Delta d)_{ir} \left(\int_{\widehat{\zeta}_i^{(1)}}^{\widehat{\zeta}_{i,1-\epsilon_N}} \left(\widehat{\mathbb{F}}_N^{i,\mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r) + \epsilon_N \right) \mathrm{d}\zeta + \int_{\widehat{\zeta}_{i,1-\epsilon_N}}^{\widehat{\zeta}_i^{(N)}} \mathrm{d}\zeta \right) \\
= & \sum_{i\in\mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} (\Delta d)_{ir} \int_{\widehat{\zeta}_i^{(1)}}^{\widehat{\zeta}_{i,1-\epsilon_N}} \widehat{\mathbb{F}}_N^{i,\mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r) \mathrm{d}\zeta + \sum_{i\in\mathcal{I}} d_{i\bar{j}} \widehat{\lambda}_{i,\epsilon_N}, && (9)
\end{aligned}
$$

where

$$
\begin{cases}
\widehat{\zeta}_{i,1-\epsilon_N} := \sup\{\zeta \in [\widehat{\zeta}_i^{(1)}, \widehat{\zeta}_i^{(N)}] : \widehat{\mathbb{F}}_N^i(\zeta) \leq 1 - \epsilon_N\}, \\
\widehat{\lambda}_{i,\epsilon_N} := \widehat{\zeta}_i^{(N)} - \widehat{\zeta}_{i,1-\epsilon_N}(1-\epsilon_N) - \epsilon_N \widehat{\zeta}_i^{(1)},
\end{cases}
\qquad (10)
$$

and the inequality follows from the definition of $\widehat{\zeta}_{i,1-\epsilon_N}$ and that $\widehat{\mathbb{F}}_N^{i,\mathcal{X}}(\zeta, \mathbf{0}_{|\mathcal{X}|}) + \epsilon_N \leq \widehat{\mathbb{F}}_N^i(\zeta) + \epsilon_N \leq 1$ for all $\zeta \in [\widehat{\zeta}_i^{(1)}, \widehat{\zeta}_{i,1-\epsilon_N}]$ and any given $\mathcal{X} \subseteq \mathcal{J}$. Note that (9) is not a strict upper bound but rather a probabilistic upper bound (PUB) on $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$, which will be rigorously proved in Lemma 5. In addition, we call ϵ_N as the *conservatism parameter* of the PUB estimator. Analogous to the Kolmorogov DRO estimator, a smaller ϵ_N suggests a less conservative PUB estimator of the true expected operating cost. We then refer to the expression (9) as a PUB estimator, and formally define it as follows.

Definition 2 (Probablistic Upper Bound Estimator). Given a facility location design $\boldsymbol{x} \in \mathbb{B}^J$, a data set $\widehat{\mathcal{S}}_N = \{\widehat{\boldsymbol{\omega}}^n = (\widehat{\boldsymbol{\zeta}}^n, \widehat{\boldsymbol{\xi}}^n) : n \in \{1,2,\ldots,N\}\}$, and a conservatism parameter $\epsilon_N \in [0,1]$, the PUB estimator is defined as $\widehat{\Phi}_N(\boldsymbol{x}) := \sum_{i\in\mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} (\Delta d)_{ir} \int_{\widehat{\zeta}_i^{(1)}}^{\widehat{\zeta}_{i,1-\epsilon_N}} \widehat{\mathbb{F}}_N^{i,\mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r) \mathrm{d}\zeta + \sum_{i\in\mathcal{I}} d_{i\bar{j}} \widehat{\lambda}_{i,\epsilon_N}$, where the definition of related notation can be found in Lemma 2 and (10).

In contrast to the Kolmogorov DRO estimator $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$, which is implicitly defined by an optimization problem, the PUB estimator $\widehat{\Phi}_N(\boldsymbol{x})$ has an analytical form. This offers threefold advantages: First, it is tractable to evaluate the value of $\widehat{\Phi}_N(\boldsymbol{x})$ for any given facility location design $\boldsymbol{x}$ (see Lemma 3 in Section 3). Second, the PUB estimator $\widehat{\Phi}_N(\boldsymbol{x})$ preserves useful structural properties of the true operating cost function $\Phi^\star(\boldsymbol{x})$, such as monotonicity and supermodularity (Lemma 3), which facilitates efficient solutions (Section 3). Third, although $\widehat{\Phi}_N(\boldsymbol{x})$ may be more conservative than $\widehat{\Phi}_N^{\mathrm{K}}(\boldsymbol{x})$, the conservatism of $\widehat{\Phi}_N(\boldsymbol{x})$ will disappear as more data are obtained. This is referred to as the asymptotic optimality of $\widehat{\Phi}_N(\boldsymbol{x})$, and will be discussed in Section 4. Moreover, even when the number of samples are limited, using $\widehat{\Phi}_N(\boldsymbol{x})$ to guide RFL designs is guaranteed to have favorable out-of-sample performances (Section 4). In the remainder, we focus on describing results concerning Problem (DD-RFL) with the PUB estimator $\widehat{\Phi}_N(\boldsymbol{x})$ defined in Definition 2.

3. Solution Algorithm

In this section, we provide tractability results for Problem (DD-RFL) with the PUB estimator. Problem (DD-RFL) is an NP-hard problem, as it boils down to the NP-hard SAA version of the RFL problem when $\epsilon_N = 0$. Nonetheless, we can reformulate Problem (DD-RFL) as an MILP by exploiting structural properties of $\widehat{\Phi}_N(\boldsymbol{x})$, and develop a constraint generation algorithm to solve it. Our results are based on the following properties of the PUB estimator $\widehat{\Phi}_N(\boldsymbol{x})$.

Lemma 3 **(Properties of $\widehat{\Phi}_N(\boldsymbol{x})$).** *The following assertions hold true:*

(i) Given any $\boldsymbol{x} \in \mathbb{B}^J$, the value of $\widehat{\Phi}_N(\boldsymbol{x})$ can be obtained in polynomial time.

(ii) $\widehat{\Phi}_N(\boldsymbol{x})$ can be decomposed as $\widehat{\Phi}_N(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \widehat{\Phi}_{i,N}(\boldsymbol{x})$, where each $\widehat{\Phi}_{i,N}(\boldsymbol{x}) := \sum_{r=0}^{|\mathcal{X}|}(\Delta d)_{ir} \int_{\widehat{\zeta}_i^{(1)}}^{\widehat{\zeta}_{i,1-\epsilon_N}} \widehat{\mathbb{F}}_N^{i,\mathcal{X}_i^{r-1}}(\zeta, \mathbf{0}_r)d\zeta + d_{i\bar{j}}\widehat{\lambda}_{i,\epsilon_N}$ is nonincreasing and supermodular[3] in $\boldsymbol{x}$.

By Assertion (ii) of Lemma 3, Problem (DD-RFL) can be rewritten as

$$\widehat{z}_N = \min_{(\boldsymbol{x},\boldsymbol{\gamma})\in\mathbb{B}^J\times\mathbb{R}^I} \left\{ \boldsymbol{f}^\intercal \boldsymbol{x} + \mathbf{1}^\intercal \boldsymbol{\gamma} : \gamma_i \geq \widehat{\Phi}_{i,N}(\boldsymbol{x}) \text{ for all } i \in \mathcal{I} \right\}. \tag{11}$$

As each $\widehat{\Phi}_{i,N}(\boldsymbol{x})$ is a nonincreasing supermodular function, the constraint $\gamma_i \geq \widehat{\Phi}_{i,N}(\boldsymbol{x})$ is equivalent to a series of supermodular inequalities (see, e.g., lemma 1 of Nemhauser and Wolsey 1981). This leads to the following MILP formulation.

Theorem 1 **(Reformulation of Problem (DD-RFL)).** *Problem (DD-RFL) is equivalent to the following MILP problem:*

$$\widehat{z}_N = \min_{(\boldsymbol{x},\boldsymbol{\gamma})\in\mathbb{B}^J\times\mathbb{R}^I} \ \boldsymbol{f}^\intercal \boldsymbol{x} + \mathbf{1}^\intercal \boldsymbol{\gamma} \tag{12a}$$

$$s.t. \ \gamma_i \geq \widehat{\Phi}_{i,N}(\boldsymbol{v}) + \sum_{j\in\mathcal{J}} \widehat{\Phi}_{i,N}(j|\boldsymbol{v})x_j \quad \forall i \in \mathcal{I}, \ \forall \boldsymbol{v} \in \mathbb{B}^J, \tag{12b}$$

where $\widehat{\Phi}_{i,N}(j|\boldsymbol{v}) := \widehat{\Phi}_{i,N}(\boldsymbol{v} \vee \boldsymbol{e}_J^j) - \widehat{\Phi}_{i,N}(\boldsymbol{v})$.

Problem (12) is an MILP with an exponential number of constraints, as the constraint (12b) is enforced over $\mathbb{B}^J$. It is noteworthy that the number of both decision variables and constraints of Problem (12) does not scale in the number of data points[4] N. Thus Problem (12) can be applied to the setting of big data. A constraint generation algorithm is appropriate to solve this MILP. Below we present the details of this algorithm.

To efficiently generate a feasible solution to Problem (12), we can instead solve a relaxation where the constraint (12b) is enforced over a manageable subset $\mathcal{V}_i \subset \mathbb{B}^J$ for each $i \in \mathcal{I}$; that is, we solve the following problem:

$$\min_{\boldsymbol{x}\in\mathbb{B}^J, \boldsymbol{\gamma}\in\mathbb{R}^I} \left\{ \boldsymbol{f}^\intercal \boldsymbol{x} + \mathbf{1}^\intercal \boldsymbol{\gamma} \,\middle|\, \gamma_i \geq \widehat{\Phi}_{i,N}(\boldsymbol{v}) + \sum_{j\in\mathcal{J}} \widehat{\Phi}_{i,N}(j|\boldsymbol{v})x_j \quad \forall i \in \mathcal{I} \ \forall \boldsymbol{v} \in \mathcal{V}_i \right\}. \tag{MP}$$

Let $\overline{\boldsymbol{x}}, \overline{\boldsymbol{\gamma}}$ be the incumbent solution. Thus the incumbent objective value $\boldsymbol{f}^\intercal \overline{\boldsymbol{x}} + \mathbf{1}^\intercal \overline{\boldsymbol{\gamma}}$ is a lower bound on $\widehat{z}_N$. To tighten the relaxation and obtain an improved bound, we can add a new vector to each $\mathcal{V}_i$ (if necessary) and resolve Problem (MP). To identify such vector, for each i, we solve the following subproblem:

$$\max_{\boldsymbol{v} \in \mathbb{B}^J} \left\{ \widehat{\Phi}_{i,N}(\boldsymbol{v}) + \sum_{j \in \mathcal{J}} \widehat{\Phi}_{i,N}(j|\boldsymbol{v}) \overline{x}_j \right\}. \tag{SP$_i$}$$

Although each subproblem (SP$_i$) is a nonlinear integer problem, we can prove that it has a closed-form solution by using the structural properties shown in Lemma 3.

Proposition 2 **(Subproblem Solution)**. *For each $i \in \mathcal{I}$, the incumbent solution $\overline{\boldsymbol{x}}$ and $\widehat{\Phi}_{i,N}(\overline{\boldsymbol{x}})$ are an optimal solution and the optimal value of Problem (SP$_i$). In addition, the optimal value of Problem (SP$_i$) can be obtained in polynomial time.*

By Proposition 2, if $\overline{\gamma}_i \geq \widehat{\Phi}_{i,N}(\overline{\boldsymbol{x}})$ for all $i \in \mathcal{I}$, then it follows that $\overline{\gamma}_i \geq \widehat{\Phi}_{i,N}(\boldsymbol{v}) + \sum_{j \in \mathcal{J}} \widehat{\Phi}_{i,N}(j|\boldsymbol{v}) \overline{x}_j$ for all $i \in \mathcal{I}$ and $\boldsymbol{v} \in \mathbb{B}^J$, and the incumbent solution is optimal to the original problem (12). Otherwise, for those i with $\overline{\gamma}_i < \widehat{\Phi}_{i,N}(\overline{\boldsymbol{x}})$, we add $\overline{\boldsymbol{x}}$ to $\mathcal{V}_i$ to cut off the incumbent solution. We then resolve Problem (MP), obtain a new solution $\overline{\boldsymbol{x}}, \overline{\boldsymbol{\gamma}}$, and start a new iteration of the algorithm. As both $\mathbb{B}^J$ and $\mathcal{I}$ are finite sets, the algorithm will stop after a finite number of iterations. Below we provide an outline of the constraint generation algorithm, whose computational performance will be investigate in Section 5.

Initialization. Set $\mathcal{V}_i = \{\mathbf{1}_J\}$ for each $i \in \mathcal{I}$.

1. Solve Problem (MP) to obtain solution $\overline{\boldsymbol{x}}, \overline{\boldsymbol{\gamma}}$.
2. Set $\mathcal{I}' = \{i \in \mathcal{I} : \overline{\gamma}_i < \widehat{\Phi}_{i,N}(\overline{\boldsymbol{x}})\}$.
3. **If** $\mathcal{I}' = \emptyset$, terminate and return optimal facility location design $\overline{\boldsymbol{x}}$.
   Else, set $\mathcal{V}_i \leftarrow \mathcal{V}_i \cup \{\overline{\boldsymbol{x}}\}$ for all $i \in \mathcal{I}'$, and return to step 1.

4. Performance Guarantees

Having developed efficient algorithm for solving Problem (DD-RFL), we now investigate the performance of the obtained data-driven RFL designs. We remark that although our approach is developed by upper bounding the Kolmogorov DRO estimator $\widehat{\Phi}_N^{\mathrm{K}}$, the focus of the analysis in this section is on the gap between the PUB estimator $\widehat{\Phi}_N$ and the *true optimal expected operating cost* $\Phi^\star$ (rather than the gap between $\widehat{\Phi}_N$ and $\widehat{\Phi}_N^{\mathrm{K}}$), as $\Phi^\star$ is the primary goal that we intend to approach. Our results are twofold: First, as more data are obtained, our approach can produce data-driven solutions that are asymptotically optimal to the true RFL problem. Second, with finite samples, we provide sufficient conditions for our approach to generate data-driven RFL designs whose out-of-sample cost can be bounded from above with high probability. In this section, we use the notation $\widehat{\boldsymbol{x}}_N$ and $\boldsymbol{x}^\star$ to denote the optimizers of Problems (DD-RFL) and (RFL), respectively.

4.1. Asymptotic Optimality

Essentially, Problem (DD-RFL) provides an approximation for the true RFL problem by using limited information contained in the available data. However, provided that $\epsilon_N \to 0$ as more data are obtained, it can be proved that the optimal cost and any optimal facility location design generated by Problem (DD-RFL) will converge almost surely to those of the true RFL problem (Theorem 2). This is referred to as the asymptotic optimality of Problem (DD-RFL). The practical implication of this result is that any suboptimality of Problem (DD-RFL) disappears as long as we have sufficient data.

Theorem 2 **(Asymptotic Optimality of Problem (DD-RFL)).** *Let $\epsilon_N \to 0$ as $N \to \infty$. Then almost surely we have (i) $\widehat{z}_N \to z^\star$ as $N \to \infty$, and (ii) any limit point of $\{\widehat{\boldsymbol{x}}_N\}_{N\in\mathbb{N}}$ is an optimal solution for Problem (RFL).*

A key result to prove Theorem 2 is that we can construct both upper and lower bounds on $\widehat{\Phi}_N(\boldsymbol{x})$ by using the SAA operating cost

$$\widehat{\Phi}_N^{\mathrm{SAA}}(\boldsymbol{x}) := \frac{1}{N}\sum_{n=1}^{N} \phi(\boldsymbol{x}, \widehat{\boldsymbol{\omega}}^n), \tag{13}$$

which is asymptotically consistent with the true operating cost $\Phi^\star(\boldsymbol{x})$ (Kleywegt et al. 2002). We highlight this result in the following lemma.

Lemma 4. *Given any facility location design $\boldsymbol{x} \in \mathbb{B}^J$, we have that*

$$\underline{\phi}_N \le \widehat{\Phi}_N(\boldsymbol{x}) - \widehat{\Phi}_N^{SAA}(\boldsymbol{x}) \le \overline{\phi}_N,$$

where $\{\underline{\phi}_N\}_{N\in\mathbb{N}}$ and $\{\overline{\phi}_N\}_{N\in\mathbb{N}}$ are two sequences that both converge to 0 as $N \to \infty$.

By Lemma 4, Assertion (i) of Theorem 2 immediately follows by noting that $\underline{\phi}_N \le \widehat{z}_N - \widehat{z}_N^{\mathrm{SAA}} \le \overline{\phi}_N$ and $\lim_{N\to\infty} \widehat{z}_N^{\mathrm{SAA}} \overset{\text{a.s.}}{=} z^\star$ (Kleywegt et al. 2002), where “a.s.” denotes “almost surely”. Then the convergence result of $\{\widehat{\boldsymbol{x}}_N\}_{N\in\mathbb{N}}$ presented in Assertion (ii) of Theorem 2 follows from Assertion (i) by applying some well-known inequalities in the theory of limits.

4.2. Finite Sample Performance Guarantees

Next we focus on evaluating the performance of our data-driven RFL designs produced by using a limited number of samples. Notably, we are concerned with the *out-of-sample cost* of the data-driven RFL design $\widehat{\boldsymbol{x}}_N$, defined as $Z^\star(\widehat{\boldsymbol{x}}_N) = \boldsymbol{f}^\intercal \widehat{\boldsymbol{x}}_N + \mathbb{E}^{\mathbb{P}^\star}[\phi(\widehat{\boldsymbol{x}}_N, \widetilde{\boldsymbol{\omega}})]$, the expected total cost under a new sample independent of the training data set. However, the exact value of $Z^\star(\widehat{\boldsymbol{x}}_N)$ cannot be computed because $\mathbb{P}^\star$ is unknown. We are thus hope to bound the out-of-sample cost $Z^\star(\widehat{\boldsymbol{x}}_N)$ from above based on the training data set. In particular, we demonstrate a finite sample guarantee that the optimal value $\widehat{z}_N$ of Problem (DD-RFL), with a carefully chosen conservatism parameter, can upper bound the out-of-sample cost of $\widehat{\boldsymbol{x}}_N$ with high probability.

Recall that the training data set $\widehat{\mathcal{S}}_N$ consists of N data points independently generated from the true distribution $\mathbb{P}^\star$. Thus, before its realization, the data set $\widehat{\mathcal{S}}_N$ can be viewed as a random "object" governed by the N-fold product distribution denoted by $\mathbb{P}^\star_{\widehat{\mathcal{S}}_N}$. Then, to retain mathematical rigor, our finite sample guarantee can be restated as follows: For any given $\beta \in (0,1)$, there exists an $\epsilon_N = \epsilon_N(\beta)$ such that

$$\mathbb{P}^\star_{\widehat{\mathcal{S}}_N}\left\{Z^\star(\widehat{\boldsymbol{x}}_N) \le \widehat{z}_N\right\} \ge 1-\beta, \tag{14}$$

where $\widehat{z}_N$ and $\widehat{\boldsymbol{x}}_N$ are the optimal value and an optimizer of Problem (DD-RFL) with conservatism parameter $\epsilon_N(\beta)$. In other words, $\widehat{z}_N$ provides a $1-\beta$ confidence upper bound on the out-of-sample cost of $\widehat{\boldsymbol{x}}_N$. Note that similar bounds of type (14) are also used in the DRO literature for evaluating different data-driven models (e.g., Esfahani and Kuhn 2018, Bertsimas et al. 2018a, Van Parys et al. 2021). In (14), the parameter $\beta \in (0,1)$ is termed as a significance parameter with respect to the distribution $\mathbb{P}^\star_{\widehat{\mathcal{S}}_N}$, and the left-hand side of (14) is termed as the reliability of the optimal value $\widehat{z}_N$.

Before we proceed to prove that (14) holds for Problem (DD-RFL), we need the following lemma, which shows that the PUB estimator can bound the Kolmogorov DRO estimator from above with high probability.

Lemma 5. *Assume that the probability distribution of each $\widetilde{\zeta}_i$ is supported on $[\underline{\zeta}_i, 0]$ for some $\underline{\zeta}_i \in \mathbb{R}_-$, and the probability of the event $\widetilde{\zeta}_i = \zeta$ is strictly positive for all $\zeta \in \{\underline{\zeta}_i, 0\}$ and $i \in \mathcal{I}$. Then there exists $\rho \in (0,1)$ such that*

$$\mathbb{P}^\star_{\widehat{\mathcal{S}}_N}\left\{\widehat{\Phi}_N(\boldsymbol{x}) \ge \widehat{\Phi}^K_N(\boldsymbol{x})\right\} \ge 1 - 2I\rho^N,$$

for any given $\boldsymbol{x} \in \mathbb{B}^J$ and $\epsilon_N \in [0,1]$.

Combining the results of Lemma 5 and the previous measure concentration result (Lemma 1), we obtain the following finite sample performance guarantee for Problem (DD-RFL).

Theorem 3 **(Finite Sample Guarantee for Problem (DD-RFL))**. *Let the conditions of Lemma 5 hold. Then there exists a constant $\rho \in (0,1)$ such that, for any given $\beta \in (0,1)$, setting*

$$\epsilon_N(\beta) = \sqrt{\frac{\ln\left(2(I+J)(N+1)\beta^{-1}\right)}{2N}} \wedge 1. \tag{15}$$

implies the finite sample guarantee (14) for all $N \ge \lceil \ln(\beta/4I)/\ln\rho \rceil$.

Usually in practice, the parameter $\epsilon_N(\beta)$ given by (15) may be significantly larger than necessary and thus yield overly conservative solutions. Note that the level of conservatism of $\widehat{z}_N$ reduces as ϵ_N decreases. Given a data set $\widehat{\mathcal{S}}$ and a significance parameter β (or, equivalently, a performance requirement $1-\beta$), we can implement the following method to find the smallest $\widehat{\epsilon}(\beta)$ that achieves reliability of at least $1-\beta$. Below we use $\widehat{\mathbb{P}}_{\widehat{\mathcal{S}}}$ to denote the empirical distribution associated with a data set $\widehat{\mathcal{S}}$.

The algorithm is a combination of two procedures, namely, bootstrap and binary search. For one, the algorithm independently bootstraps K training-test pairs from the original data set $\widehat{\mathcal{S}}$, where K is a pre-determined parameter. Here each training-test pair $l \in \{1,2,\dots,L\}$ consists of a training data set $\widehat{\mathcal{S}}^l$, also termed as a resample, bootstrapped from $\widehat{\mathcal{S}}$ such that $|\widehat{\mathcal{S}}^l| = |\widehat{\mathcal{S}}|$, and a test data set $\widetilde{\mathcal{S}}^l = \widehat{\mathcal{S}} \setminus \widehat{\mathcal{S}}^l$. For each l, we use the training data $\widehat{\mathcal{S}}$ to obtain the optimal value $\widehat{z}^l$ and the optimal solution $\widehat{\boldsymbol{x}}^l$ of Problem (DD-RFL), and the test data $\widetilde{\mathcal{S}}^l$ to compute the out-of-sample cost $\widetilde{z}^l := \boldsymbol{f}^\intercal \widehat{\boldsymbol{x}}^l + \mathbb{E}^{\widehat{\mathbb{P}}_{\widetilde{\mathcal{S}}^l}}[\phi(\widehat{\boldsymbol{x}}^l, \widetilde{\boldsymbol{\omega}})]$ incurred by solution $\widehat{\boldsymbol{x}}^l$. Using the L training-test pairs, we can evaluate whether a given radius $\epsilon \in [0,1]$ achieves the performance guarantee $1-\beta$. That is, ϵ satisfies the performance guarantee $1-\beta$, if for at least $\lceil (1-\beta)L \rceil$ training-test pairs, our model suggests an optimal value that upper bounds the out-of-sample cost of its solution. Then, the algorithm applies a binary search method to find the smallest radius $\widehat{\epsilon}(\beta)$ that satisfies the performance guarantee $1-\beta$. The psuedocode of the proposed algorithm is presented in Algorithm 1 in Appendix EC.3.

5. Numerical Results

In this section, we demonstrate the practical value of using the PUB estimator in RFL design. In particular, we analyze the effectiveness (Section 5.1) and computational efficiency (Section 5.2) of our data-driven RFL model with the PUB estimator, compared with the following benchmark RFL models:

1. Wass: The type-∞ Wasserstein DRO model of Xie (2020);
2. MM: The marginal moment-based RFL model of Lu et al. (2015);
3. CM: The cross moment-based RFL model used in the numerical experiments of Li et al. (2022).

The subsequent experiments use the same network data set as in many RFL studies including Snyder and Daskin (2005), Cui et al. (2010), Lu et al. (2015), and Li et al. (2022). This data set contains information of facility locations, (deterministic) demand levels, fixed costs, and transportation costs on networks based on the US map with up to 150 cities. Observations of random demands and disruptions are generated synthetically. Details of the data-generating process as well as other experiment settings will be discussed in the following. All the computational experiments are conducted using Gurobi 9.5 with Python API on an Ubuntu server equipped with 20 processors and 40G RAM.

5.1. Effectiveness of the PUB Estimator

In this section, we conduct experiments to compare the performance of aforementioned models when the size of the data set (of demands and disruptions) varies. All problem instances are based on a fixed network with 10 locations of candidate facilities/customers. Problem instances on larger-scale networks are presented in Section 5.2. We consider two types of demand distributions, namely, high demand (denoted by **H**) and low demand (denoted by **L**), with different levels of expected values. For a given number N, we generate N observations of the network state vector as follows. We first generate N vectors $\widehat{\boldsymbol{u}}^n \in \mathbb{R}^{I+J}$, $n \in \{1,2,\dots,N\}$, independently drawn from a multivariate normal distribution with mean $\mathbf{0}$ and a random covariance matrix[5] $\boldsymbol{\Sigma}$. Then we obtain N samples of the network state vector $\widehat{\boldsymbol{\omega}}^n$, $n \in \{1,2,\dots,N\}$, by

$$\widehat{\zeta}_i^n = \begin{cases} -\max\{\mu_i(\widehat{u}_i^n + 1.6), 0\} \text{ if } \mathbf{H}\text{-type} \\ -\max\{\mu_i(\widehat{u}_i^n + 0.4), 0\} \text{ if } \mathbf{L}\text{-type} \end{cases} \quad \text{and} \quad \widehat{\xi}_j^n = \mathbb{I}\left\{\widehat{u}_{I+j}^n \geq \rho_{I+j,1/10}\right\}.$$

where μ_i is the demand at node i from the network data set, $\rho_{i,\alpha}$ is the α-quantile of the i-th marginal distribution of the normal distribution $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$. In all benchmark models, Wass, CM, and MM, we set the demand as $\widehat{d}_i^n = -\widehat{\zeta}_i^n$. In addition, as suggested in Xie (2020), we use the nomalized demand data $\widehat{d}_i^n / \max_n\{\widehat{d}_i^n\} \in [0,1]$ to ensure data consistency.

The remainder of the experiment settings is described as follows. Given the type of the underlying distribution, we first generate a testing data set of size $N^{\text{out}} = 10000$. Then for varying values of N from 10 to 1000, we generate $M = 100$ training data sets of size N. For each training data set $s \in \{1,2,\dots,M\}$, we solve each model $\mathscr{M} \in \{\text{PUB}, \text{Wass}, \text{MM}, \text{CM}\}$ to obtain a facility location design $\widehat{\boldsymbol{x}}_N^{\mathscr{M},s}$ and the associated objective value $\widehat{z}_N^{\mathscr{M},s}$. We then use the SAA cost of $\widehat{\boldsymbol{x}}_N^{\mathscr{M},s}$ over the testing data set, defined by

$$Z^{\text{out}}(\widehat{\boldsymbol{x}}_N^{\mathscr{M},s}) = \boldsymbol{f}^\intercal \widehat{\boldsymbol{x}}_N^{\mathscr{M},s} + \frac{1}{N^{\text{out}}} \sum_{n=1}^{N^{\text{out}}} \phi(\widehat{\boldsymbol{x}}_N^{\mathscr{M},s}, \widehat{\boldsymbol{\omega}}^{\text{out},n}),$$

to estimate the out-of-sample cost of $\widehat{\boldsymbol{x}}_N^{\mathscr{M},s}$. Furthermore, we solve the SAA RFL model for an independent data set of size 10^6, and use its optimal value $\widehat{z}^\star (\approx z^\star)$ as the *true optimum* of the RFL problem.

To evaluate the effectiveness, we compare each method $\mathscr{M}$ along the following metrics:

1. **In-sample performance**: The ratio of the average optimal value over the M training data sets to the true optimum: $M^{-1}\sum_{s=1}^{M} \widehat{z}_N^{\mathscr{M},s} / \widehat{z}^\star$.

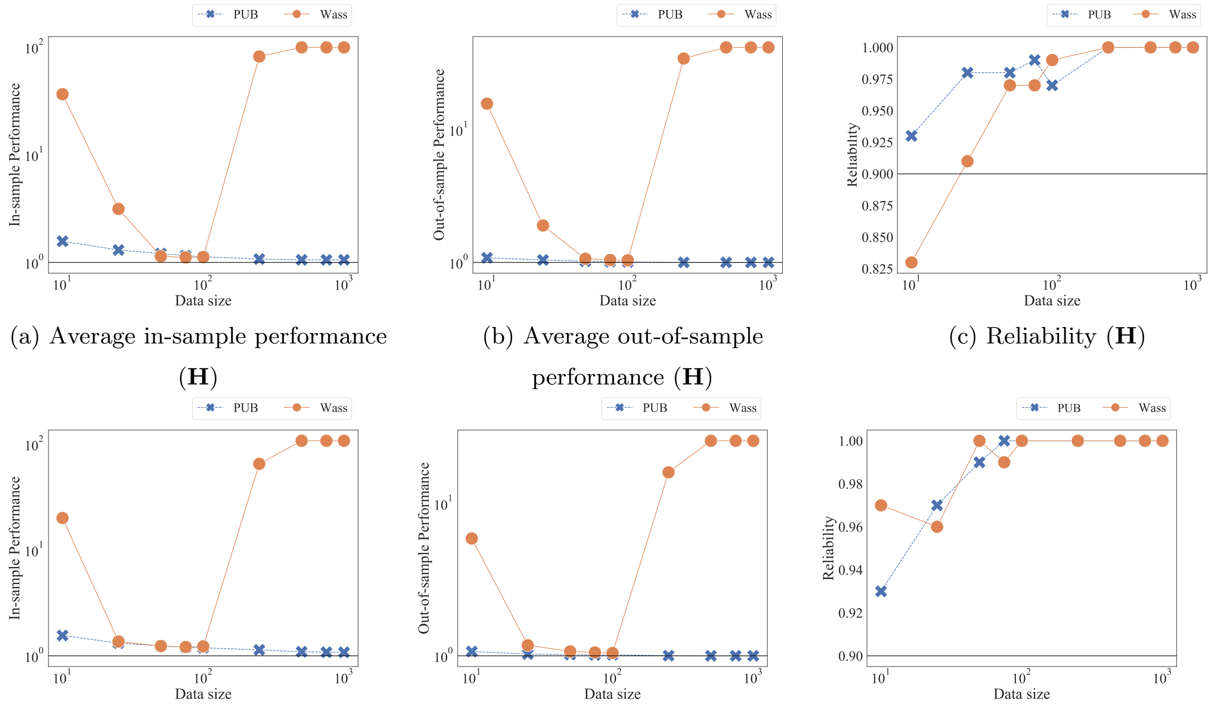
2. **Out-of-sample performance**: The ratio of the average out-of-sample cost over the M training data sets to the true optimum: $M^{-1}\sum_{s=1}^{M} Z^{\mathrm{Out}}(\widehat{\boldsymbol{x}}_N^{\mathscr{M},s})/\widehat{z}^{\star}$.
3. **Reliability**: The percentage of realizations in the M training data sets for which the optimal value does not underestimate the out-of-sample cost: $M^{-1}\sum_{s=1}^{M}\mathbb{I}\{\widehat{z}_N^{\mathscr{M},s} \geq Z^{\mathrm{Out}}(\widehat{\boldsymbol{x}}_N^{\mathscr{M},s})\}$.

5.1.1. PUB v.s. Wass. We first compare the performance between two data-driven RFL models, i.e., PUB and Wass. The parameters of both models are determined by Algorithm 1 with resample size $L = 50$, stopping criterion $\delta = 10^{-3}$, significance parameter $\beta = 0.1$, and computational time limit $\bar{T} = 750$ seconds. We restrict the conservatism parameters of both models, $\epsilon_N^{\mathrm{PUB}}$ and $\epsilon_N^{\mathrm{Wass}}$ (the size of the Wasserstein ambiguity set), within [0,1]. This is because, by the formulation of Xie (2020), Wass would suggest a trivial and overly-conservative solution $\widehat{\boldsymbol{x}}_N^{\mathrm{Wass}} = \mathbf{0}$ (given any data set) if setting $\epsilon_N^{\mathrm{Wass}} \geq 1$. Figure 1 summarizes the performance results for PUB and Wass under both **H**-type and **L**-type underlying distributions.

In-sample performance. Figures 1a and 1d illustrate the attractive in-sample performance of PUB, which is sufficiently close to the true optimum of the RFL problem, compared with the one of Wass. More interestingly, we observe significant gaps between the in-sample performance of Wass and the true optimum in both small-sample ($N \leq 25$ under **H** and $N = 10$ under **L**) and large-sample ($N \geq 100$) scenarios. In the small-sample scenario, owing to the limited number of training data, we find that there is a noticeable chance of an "extreme" event that the reliability of $\widehat{\boldsymbol{x}}_N^{\mathrm{Wass}}$ cannot reach the required significance level $1-\beta$, even though the conservatism parameter $\epsilon_N^{\mathrm{Wass}}$ is sufficiently close to 1. In this event, Algorithm 1 sets $\epsilon_N^{\mathrm{Wass}} = 1$, which yields the overly-conservative solution $\widehat{\boldsymbol{x}}_N^{\mathrm{Wass}} = \mathbf{0}$. Consequently, this leads to a considerable increase in the average in-sample performance of Wass, which suggests that Wass can provide overly conservative estimate of the true expected network cost when data size is small.

When the training data size grows, the chance of the aforementioned extreme event decreases. Thus the in-sample performance of Wass becomes comparable with that of PUB. However, when the data size continues increasing, e.g., $N \geq 100$, we observe a sudden surge in the in-sample performance of Wass. This is because the computational cost of solving Wass significantly increases as data size grows, so that Algorithm 1 will be terminated way before finding the smallest conservatism parameter for Wass due to the time limit. Thus, the in-sample performance of Wass is far away from its optimum. In particular, as is shown in Figures 1a, 1d, 1b, and 1e, solutions of Wass for problem instances with data size $N \in \{500, 750, 1000\}$ share (roughly) the same in-sample and out-of-sample performances, identical to the ones of the trivial solution where no facility is built. This is because solving Wass is so time-consuming that, for any training data set of size

N, Algorithm 1 never completes the first iteration when reaching the time limit, and outputs $\epsilon_N^{\text{Wass}} = 1$ as the "best" conservatism parameter. By contrast, solving PUB is computationally efficient, which allows Algorithm 1 to always find the best conservatism parameter within the time limit. Therefore, we can conclude that PUB consistently outperforms Wass with respect to the in-sample performance for varying sizes of data sets.



(a) Average in-sample performance (**H**)

(b) Average out-of-sample performance (**H**)

(c) Reliability (**H**)

(d) Average in-sample performance (**L**)

(e) Average out-of-sample performance (**L**)

(f) Reliability (**L**)

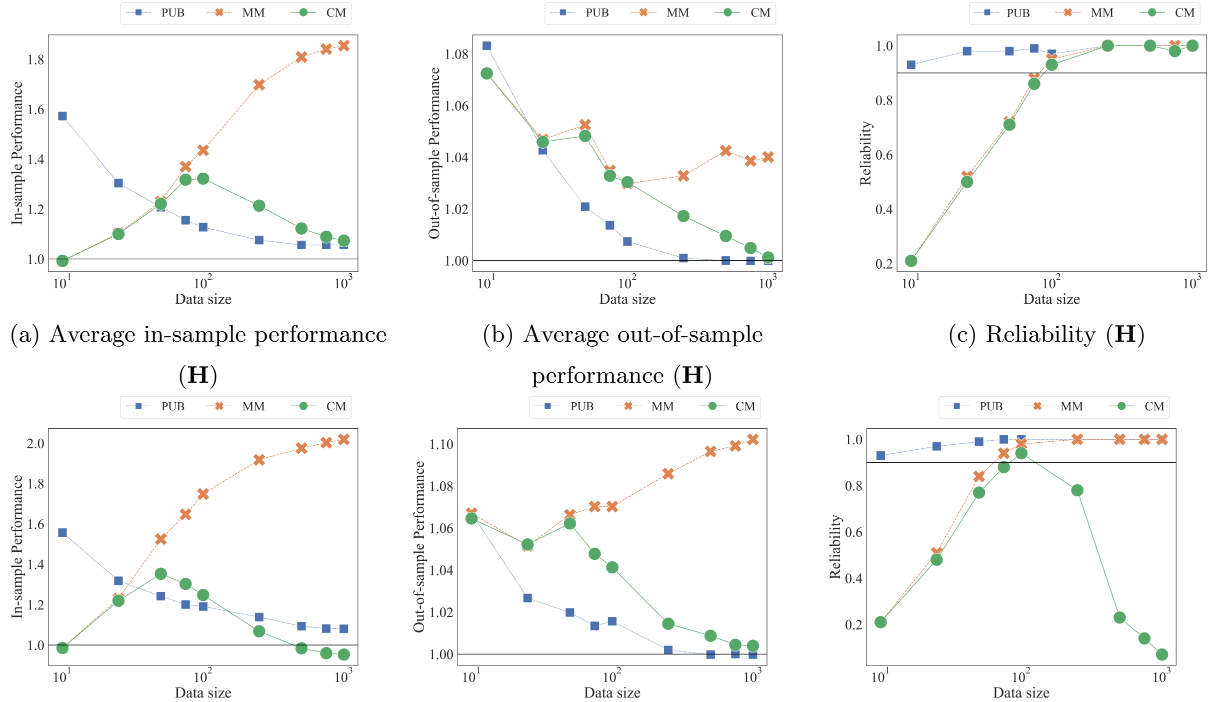
Figure 1 Performance comparison between PUB and Wass under H-type and L-type underlying distributions

Out-of-sample performance. Figures 1b and 1d show that, as the size of data set increases, PUB can consistently generate more cost-efficient RFL designs compared with Wass. The extreme values of Wass in both small-sample and large-sample scenarios can be explained by identical reasoning as in the previous discussion of its in-sample performance.

Reliability. As observed from Figures 1c and 1f, the reliablity of both models are comparable in general. In particular, PUB can achieve the required reliability (the black horizontal line in both figures) in all problem instances, whereas Wass may fail in small-sample scenarios (e.g., when data size is $N = 10$ under **H**).

5.1.2. PUB v.s. moment-based models. We then focus on the comparison between PUB and two moment-based RFL models, MM and CM. As the latter two models are not originally designed to be data-driven, all moment information of disruptions are obtained using their empirical estimation. Note that both MM and CM assumes deterministic demands. As a remedy, we use their empirical means as surrogates in the following experiments. Figure 2 illustrates the results of PUB, MM, and CM under both **H**-type and **L**-type distributions.

In-sample performance. The results in Figures 2a and 2d numerically verify the asymptotic optimality of the PUB estimator, that is, the in-sample performance of PUB converges to 1 when the data size increases. This equivalently suggests that the in-sample cost of PUB converges to the true optimum as more data are obtained. However, the moment-based methods do not seem to be asymptotically optimal. The in-sample performance of CM first increases as the data size grows, then decreases to a level that can be even lower than the true optimum when data size is large (see Figure 2d). This implies that CM can be sometimes too optimistic, and unnecessarily underestimate the true optimal cost of the RFL problem. On the other hand, the in-sample performance of MM constantly increases as more data are obtained, which indicates that MM can hardly benefit from the increase of data size.



(a) Average in-sample performance (**H**)

(b) Average out-of-sample performance (**H**)

(c) Reliability (**H**)

(d) Average in-sample performance (**L**)

(e) Average out-of-sample performance (**L**)

(f) Reliability (**L**)

Figure 2 **Performance comparison between PUB and moment-based models under H-type and L-type distributions**

Out-of-sample performance. As is shown in Figures 2b and 2e, the out-of-sample performance of both PUB and CM converge to the true optimum when we have more training data, whereas the one of MM fails to exhibit such trend (see Figure 2e). Furthermore, PUB has the lowest out-of-sample costs compared with MM and CM in nearly all problem instances.

Reliability. As the moment-based models are not originally developed to be data-driven, both MM and CM is outperformed by PUB with respect to the reliability. As shown in Figures 2c and 2f, the PUB estimator can meet the reliability requirement in all cases, whereas the reliability of MM or CM can be as low as 0.2 when the data size is $N = 10$. Under the **H**-type distribution, the moment-based methods can achieve the required reliability only in the large-sample scenario. Under the **L**-type distribution, the reliability of CM even decreases when the data size grows beyond 100. We then conclude that, when applied to the RFL problem, the moment-based models can not guarantee to provide reliable (or safe) estimates for the true expected cost of their RFL designs.

5.2. Computational Efficiency of the PUB Estimator

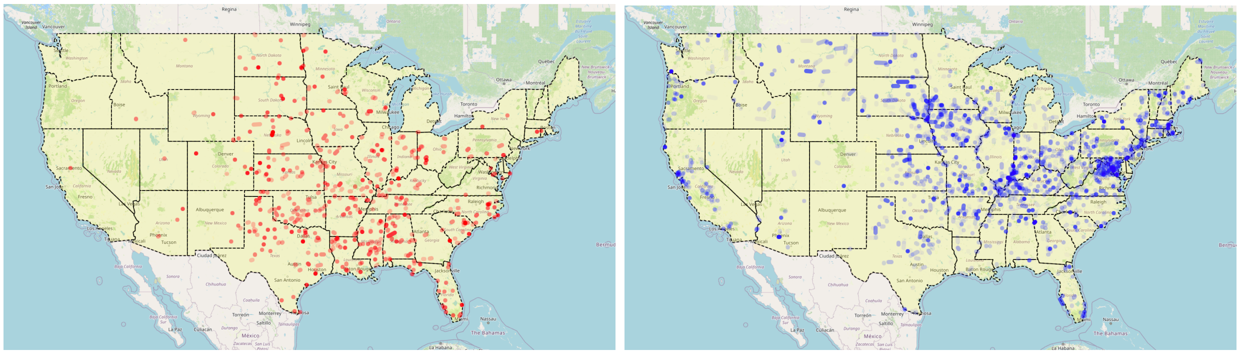
In this section, we show the computational benefits of the PUB approach. We consider the problem instances with network size $I = J \in \{10, 20, 50\}$, data size $N \in \{100, 500, 1000, 5000, 10000\}$. The data generation process is similar to the one in Section 5.1. In addition, we only consider **H**-type scenario, as the demand distribution does not essentially affect the evaluation of computational efficiency.

Solution algorithms for all the models, namely, PUB, Wass, MM, and CM, are terminated at either a 0.05% optimality gap or a maximum CPU time of 900 seconds in Gurobi Solver.[6] To keep the computational burden manageable, for a given network, the conservatism parameters of PUB and Wass for data sets of larger sizes ($N > 100$) are assumed to be the same as the ones derived for the data set of size $N = 100$ (by using Algorithm 1 with $\beta = 0.1$). Results are summarized in Table EC.1.

It is observed from Table EC.1 that the in-sample and out-of-sample performances of our PUB estimator are comparable with Wass, and noticeably better than those of CM and MM. This is consistent with the observation on the large-sample performance for all the models in Section 5.1. Moreover, to achieve such performance, PUB requires significantly lower computational cost, compared with Wass and CM, especially when the network size or the data size is large. In particular, the computational time of Wass and CM increases dramatically when the network size reaches 50. Both models can not even be solved to optimality within the time limit when $N \geq 5000$. Compared with MM, the PUB considerably improves both in-sample and out-of-sample performance at a practically reasonable computational cost.

6. Extension: Data-Driven RFL Model with Covariate Information

In this section, we discuss how to extend our data-driven RFL model by incorporating covariate information associated with both demand and disruptions. For example, in a network facing disruptions caused by natural disasters, the affected areas tend to exhibit different patterns under different types of disasters. Figure 3 illustrates the trajectory of two different disasters, namely, tornado and flood, across the U.S. mainland from 1950 to 2021. The red and blue points in Figure 3 highlight the locations of the disasters with estimated property damages exceeding 50,000 dollars. The opacity of points denotes the severity of disasters; that is, a less transparent point represents a disaster that caused a more severe damage. An obvious observation is that locations affected by flood were more concentrated around big cities (see, for example, the areas near Washington, D.C. in Figure 3b), while tornados were evenly distributed across the middle U.S. mainland. There is growing evidence on the value of covariate information in data-driven decision-making (e.g., Hao et al. 2020, Chen et al. 2020). In the remainder, we first present the data-driven RFL model based on an extended PUB estimator that incorporates covariate information, and then numerically investigate its effectiveness.



(a) Tornado (b) Flood

Figure 3 Trajectory of Tornado and Flood with damage over 50,000 dollars in U.S. mainland during 1950-2021.

7

6.1. Model and Theoretical Results

Mathematically, we assume that there are random covariate variables associated with both disruptions and demands. We assume that all the covariates are finitely supported, and thus can be denoted by a single random variable $\widetilde{c}$ supported on $\mathcal{C} := \{c_k\}_{k\in\mathcal{K}}$, where $\mathcal{K} := \{1, 2, \ldots, K\}$ and $c_1 < c_2 < \cdots < c_K$. Note that for covariates with infinite supports, we can apply a wide range of machine learning methods to partition the observations of those covariates into finite scenarios, and define a scenario-specific variable as a surrogate covariate (e.g., Hao et al. 2020, Chen et al. 2020).

The training data set is now defined as $\widehat{\mathcal{S}}_N := \{(\widehat{\boldsymbol{\omega}}^n, \widehat{c}^n)\}_{n\in\mathcal{N}} \subseteq \Omega \times \mathcal{C}$ constituting N samples i.i.d. drawn from the true distribution $\mathbb{P}_\star \in \mathcal{P}(\Omega \times \mathcal{C})$. We partition $\widehat{\mathcal{S}}_N$ into K disjoint subsets $\widehat{\mathcal{S}}^k_{N_k} := \{(\widehat{\boldsymbol{\omega}}^{kn}, c_k)\}_{n\in\mathcal{N}_k:=\{1,\ldots,N_k\}}$, $k\in\mathcal{K}$, where each $\widehat{\mathcal{S}}^k_{N_k}$ contains all samples in $\widehat{\mathcal{S}}_N$ whose covariate value equals to c_k. Thus we have $N = \sum_{k\in\mathcal{K}} N_k$, $\widehat{\mathcal{S}}_N = \bigcup_{k\in\mathcal{K}} \widehat{\mathcal{S}}^k_{N_k}$, and $\widehat{\mathcal{S}}^k_{N_k} \cap \widehat{\mathcal{S}}^{k'}_{N_{k'}} = \emptyset$ for all $k, k' \in \mathcal{K}$ and $k \neq k'$. For any distribution $\mathbb{P} \in \mathcal{P}(\Omega\times\mathcal{C})$ with a CDF $\mathbb{F}\in\mathcal{F}(\Omega\times\mathcal{C})$, denote by $\mathbb{F}^{\widetilde{c}}$ the marginal CDF of $\widetilde{c}$ under $\mathbb{F}$, and by $\mathbb{F}^{\widetilde{\boldsymbol{\omega}}|c_k}$ the distribution of $\widetilde{\boldsymbol{\omega}}$ conditioning on $\widetilde{c} = c_k$ under $\mathbb{F}$. In particular, let $\mathbb{F}^{i,\mathcal{X}|c_k}$ denote the marginal CDF of random variables $\widetilde{\zeta}_i$ and $\widetilde{\boldsymbol{\xi}}(\mathcal{X})$ conditioning on the event $\widetilde{c} = c_k$, for $i\in\mathcal{I}$ and $\mathcal{X}\subseteq\mathcal{J}$. We then extend our PUB estimator as follows

$$\widehat{\Phi}^{\mathrm{Cov}}_N(\boldsymbol{x}) := \sum_{i\in\mathcal{I}} \sum_{r=0}^{|\mathcal{X}|} \sum_{k\in\mathcal{K}} (\Delta d)_{ir} \widehat{p}^{\epsilon}_{k,N} \left(\int_{\widehat{\zeta}^{k,(1)}_i}^{\widehat{\zeta}^k_{i,1-\epsilon_{N_k}}} \widehat{\mathbb{F}}^{i,\mathcal{X}^{r-1}_i|c_k}_{N_k}(\zeta, \mathbf{0}_r)\mathrm{d}\zeta + \widehat{\lambda}^k_{i,\epsilon_{N_k}} \right), \tag{16}$$

where $\widehat{p}^{\epsilon}_{k,N} := (N_k/N + \epsilon_N)\wedge 1$, $\widehat{\lambda}^k_{i,\epsilon_{N_k}} := \widehat{\zeta}^{k,(N_k)}_i - \epsilon_{N_k}\widehat{\zeta}^{k,(1)}_i - \widehat{\zeta}^k_{i,1-\epsilon_{N_k}}(1-\epsilon_{N_k})$, $\widehat{\zeta}^{k,(N_k)}_i := \max\{\widehat{\zeta}^{kn}_i : n\in\mathcal{N}_k\}$, $\widehat{\zeta}^{k,(1)}_i := \min\{\widehat{\zeta}^{kn}_i : n\in\mathcal{N}_k\}$, $\widehat{\zeta}^k_{i,1-\epsilon_{N_k}} := \sup\{\zeta \in [\widehat{\zeta}^{k,(1)}_i, \widehat{\zeta}^{k,(N_k)}_i] : \widehat{\mathbb{F}}^{i|c_k}_{N_k}(\zeta) \le 1-\epsilon_{N_k}\}$, and ϵ_N, ϵ_{N_k}, $k\in\mathcal{K}$, are conservatism parameters. Intuitively, the estimator $\widehat{\Phi}^{\mathrm{Cov}}_N(\boldsymbol{x})$ is a probabilistic upper bound on the *event-wise* Kolmogorov DRO-based operating cost function $\widehat{\Phi}^{\text{K-Cov}}_N(\boldsymbol{x}) = \sup_{\mathbb{F}\in\widehat{\mathcal{F}}^{\text{K-Cov}}_N} \int_{\Omega\times\mathcal{C}} \phi(\boldsymbol{x},\boldsymbol{\omega})\mathrm{d}\mathbb{F}(\boldsymbol{\omega})$, where

$$\widehat{\mathcal{F}}^{\text{K-Cov}}_N := \left\{ \mathbb{F}\in\mathcal{F}(\Omega\times\mathcal{C}) \;\middle|\; \begin{array}{l} \left\|\mathbb{F}^{\widetilde{\boldsymbol{\omega}}|c_k} - \widehat{\mathbb{F}}^{\widetilde{\boldsymbol{\omega}}|c_k}_{N_k}\right\|_\infty \le \epsilon_{N_k},\ \forall k\in\mathcal{K} \\ \left|\mathbb{F}^{\widetilde{c}}(c_k) - \mathbb{F}^{\widetilde{c}}(c_{k-1}) - N_k/N\right| \le \epsilon_N,\ \forall k\in\mathcal{K} \end{array} \right\}. \tag{17}$$

Then the data-driven RFL problem with covariate information is defined as

$$\widehat{z}^{\mathrm{Cov}}_N := \min_{\boldsymbol{x}\in\mathbb{B}^J} \boldsymbol{f}^\intercal\boldsymbol{x} + \widehat{\Phi}^{\mathrm{Cov}}_N(\boldsymbol{x}). \tag{DD-RFL-COV}$$

All of our previous results, including reformulation (Theorem 1) and performance guarantees (Theorems 2 and 3), can be generalized when covariate information is incorporated. The following theorem summarizes these results. An extended version of Algorithm 1, aiming to determine the optimal conservatism parameters for Problem (DD-RFL-COV), is provided in Algorithm 2 in Appendix EC.3.

Theorem 4. *Let $\widehat{\boldsymbol{x}}^{Cov}_N$ be the optimizer of Problem (DD-RFL-COV). The following statements hold true:*

(i) ***Reformulation and solution algorithm****: We have that* $\widehat{\Phi}_N^{Cov}(\boldsymbol{x})$ *can be decomposed as* $\widehat{\Phi}_N^{Cov}(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \widehat{\Phi}_{i,N}^{Cov}(\boldsymbol{x})$*. Then* $\widehat{\boldsymbol{x}}_N^{Cov}$ *minimizes the following MILP problem:*

$$\begin{aligned} \widehat{z}_N^{Cov} = \min_{(\boldsymbol{x},\boldsymbol{\gamma})\in\mathbb{B}^J\times\mathbb{R}^I} \quad & \boldsymbol{f}^\intercal\boldsymbol{x} + \boldsymbol{1}^\intercal\boldsymbol{\gamma} \\ \text{s.t. } & \gamma_i \geq \widehat{\Phi}_{i,N}^{Cov}(\boldsymbol{v}) + \sum_{j\in\mathcal{J}} \widehat{\Phi}_{i,N}^{Cov}(j|\boldsymbol{v})x_j \quad \forall i\in\mathcal{I}\ \forall \boldsymbol{v}\in\mathbb{B}^J, \end{aligned} \tag{18}$$

which can be solved by a constraint generation algorithm.

(ii) ***Asymptotic optimality****: Suppose that* $\mathbb{P}_\star\{\widetilde{c} = c_k\} > 0$ *for all* $k\in\mathcal{K}$*. As* $N\to\infty$ *and hence* $N_k\to\infty$ *almost surely, let* $\epsilon_N\to 0$ *and* $\epsilon_{N_k}\to 0$ *for all* $k\in\mathcal{K}$*. Then almost surely we have (i)* $\widehat{z}_N^{Cov}\to z^\star$ *as* $N\to\infty$*, and (ii) any limit point of* $\{\widehat{\boldsymbol{x}}_N^{Cov}\}_{N\in\mathbb{N}}$ *is an optimal solution for Problem (RFL).*

(iii) ***Finite sample guarantee****: For each* $k\in\mathcal{K}$ *and given* $\widetilde{c} = c_k$*, assume that the conditional probability distribution of each* $\widetilde{\zeta}_i$ *is supported on* $[\underline{\zeta}_i^k, 0]$ *for some* $\underline{\zeta}_i^k\in\mathbb{R}_-$*, and the conditional probability of the event* $\{\widetilde{\zeta}_i = \zeta|\widetilde{c} = c_k\}$ *is strictly positive for all* $\zeta\in\{\underline{\zeta}_i^k, 0\}$ *and* $i\in\mathcal{I}$*. Then there exist a constant* $\rho\in(0,1)$ *such that, for any given* $\beta\in(0,1)$*, setting*

$$\epsilon_N(\beta) = \sqrt{\frac{\ln\left(2(K+1)(N+1)\beta^{-1}\right)}{2N}} \wedge 1.$$

and

$$\epsilon_{N_k}(\beta) = \sqrt{\frac{\ln\left(2(I+J)(K+1)(N_k+1)\beta^{-1}\right)}{2N_k}} \wedge 1 \text{ for all } k\in\mathcal{K}$$

implies the following finite sample guarantee:

$$\mathbb{P}^\star_{\widehat{\mathcal{S}}_N}\left\{Z^\star(\widehat{\boldsymbol{x}}_N^{Cov}) \leq \widehat{z}_N^{Cov}\right\} \geq 1-\beta,$$

for all $N \geq \lceil \ln(\beta/4IK)/\ln\rho \rceil$*.*

6.2. Benefits of Using the Extended PUB Estimator

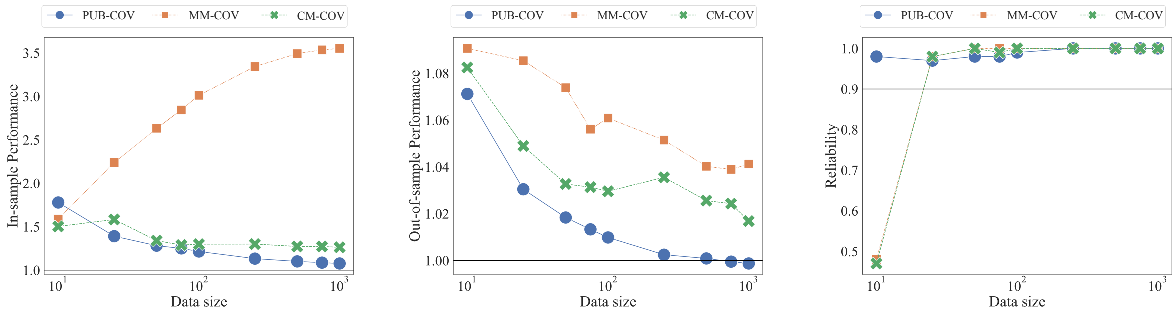
We now show the benefits of the extended PUB estimator through numerical studies using the synthetic data. We compare the performance of our model (denoted by PUB-COV) with the event-wise marginal moment-based model (denoted by MM-COV) and the event-wise cross moment-based model (denoted by CM-COV). The latter two models are event-wise extensions of MM and CM in Section 5 to incorporate covariate information; see Appendix EC.5 for their formal definitions based on the modeling techniques of Hao et al. (2020), Chen et al. (2020).

The experiments are conducted using the network with 10 nodes. Data for random disruptions and customer demands are generated as follows. We consider a two-point distribution for the covariate variable, that is, $\mathbb{P}^\star\{\widetilde{c}=1\}=1/3$ and $\mathbb{P}^\star\{\widetilde{c}=2\}=2/3$. Given the data size N, we first generate N realizations of $\widetilde{c}$. Let N_k be the number of covariate realizations that equal to $k\in\{1,2\}$. For each k, we randomly generate N_k intermediate vectors $\widehat{\boldsymbol{u}}^{kn}=(\widehat{u}_i^{kn})\in\mathbb{R}^{I+J}$, $n\in\{1,2,\ldots,N_k\}$ governed by an $(I+J)$-dimensional normal distribution $\mathcal{N}(0,\boldsymbol{\Sigma}_k)$ with a zero mean and a randomly generated covariance matrix (which depends on k). Then we obtain $(\widehat{\boldsymbol{\zeta}}^{kn},\widehat{\boldsymbol{\xi}}^{kn})$ as

$$\widehat{\zeta}_i^{kn}=\begin{cases}-\min\{\max\{0,\widehat{u}_i^{kn}+0.6\},2\}\mu_i, & \text{if } k=1\\ -\min\{\max\{0,\widehat{u}_i^{kn}+1.3\},2\}\mu_i, & \text{if } k=2,\end{cases}\qquad \widehat{\xi}_j^{kn}=\begin{cases}\mathbb{I}\left\{\widehat{u}_{I+j}^{kn}>\rho_{I+j,1/10}\right\}, & \text{if } k=1\\ \mathbb{I}\left\{\widehat{u}_{I+j}^{kn}>\rho_{I+j,3/10}\right\}, & \text{if } k=2.\end{cases}$$

That is, we assume covariate-dependent joint distributions of disruptions and demands.

Figure 4 illustrates the results of PUB-COV, MM-COV, and CM-COV, including the in-sample performance, out-of-sample performance, and the reliability. Compared with the results in Section 5.1.2, Figures 4a and 4b demonstrate an even stronger dominance of our approach over the moment-based approaches when utilizing the covariate information. In particular, PUB-COV outperforms MM-COV and CM-COV with respect to both in-sample and out-of-sample costs for nearly all problem instances. The only exception is the instance with $N=10$, where the in-sample performance of PUB-COV is slightly higher than the one of the moment-based approaches. Nonetheless, this ensures that our approach satisfies the reliability requirement, as shown in Figure 4c, whereas the low in-sample performance of the moment-based approaches leads to a poor reliability.



(a) Average In-sample Performance (b) Average Out-of-sample Performance (c) Reliability

Figure 4 Performance comparison between PUB-COV, MM-COV, and CM-COV

6.3. A Case Study

We close this section by comparing the performance of PUB-COV, MM-COV and CM-COV on a 49-node network ($I=J=49$), where demands and disruptions are simulated based on a real-world severe weather data set retrieved from the Storm Prediction Center of the National Oceanic and Atmospheric Administration (NOAA)[8]. This data set, referred to as the NOAA data set hereafter, contains the historical records of severe hazards from 1996 to 2021, and is also used in Lu et al. (2015) and Shen et al. (2021) to simulate network disruptions. In particular, each record in the NOAA data set documents the information of a single hazard including its type (e.g., Thunderstorm Wind, Hail, etc.), start and end times, trajectory, and estimated property damage. In our experiment, each hazard with an estimated property damage over 500,000 dollars is recognized as a severe hazard resulting in disruptions. We summarize the disruption states of all 49 locations on a monthly basis. In each month, a location is disrupted if it lies within 150km of the center of at least one severe hazard. We partition data into two groups ($K = 2$): For each data point in the group labeled c_1, at least one location is disrupted due to one of the "wind" type hazards, including "Marine Thunderstorm Wind", "Thunderstorm Wind", and "Marine High Wind"; we label the rest of data points as c_2. In addition, we assume that the distribution of demand $-\widehat{\zeta}_i$ at location i depends on the local disruption state $\widehat{\xi}_i$. In particular, given $\widehat{\xi}_i$, let $\widehat{\zeta}_i = \mu_i \max\{\min\{\widehat{u}_i, 0\}, 3\}$, where the intermediate variable $\widehat{u}_i$ is sampled from a normal distribution $\mathcal{N}(a, 1)$ with $a = 1/3$ if $\widehat{\xi}_i = 0$ and $a = 3$ if $\widehat{\xi}_i = 1$.

We consider 25 problem instances, indexed by 0,1,..., 24, each of which aims to design facility locations for a year among 1997-2021. In particular, in each problem instance $q \in \{0, 1, \ldots, 24\}$, we solve each model to generate an RFL design using data in year $1996 + q$ as the training data set, and evaluate the out-of-sample cost of the RFL design using the data in the year $1997 + q$. In implementing our model, we apply Algorithm 2 with $\beta = 0.2$ and $L = 30$ to determine conservatism parameters for PUB-COV. The results are illustrated in Figure 5, and summarized in Table 1.

Table 1 Comparison of performance between PUB-COV, MM-COV, and CM-COV tested on 25 years of real-world weather data (NOAA data)

Model	Avg. out-of-sample cost (10^6)	Reliability	Frequency of yielding least-cost RFL design
PUB-COV	**1.21**	**0.88**	**0.64**
MM-COV	1.25	0.8	0.12
CM-COV	1.23	0.8	0.24

We can observe from Table 1 that PUB-COV stands out as achieving lower average out-of-sample cost and the highest reliability. The detailed cost comparison for each of the problem instances is shown in Figure 5, which indicates that our PUB-COV estimator outperforms the moment-based models in most of problem instances. In particular, as shown in the last column of Table 1, the RFL design suggested by PUB-COV incurs the lowest out-of-sample cost for 64% of the 25 problem instances, which is noticeably higher than the proportions of MM-COV and CM-COV.

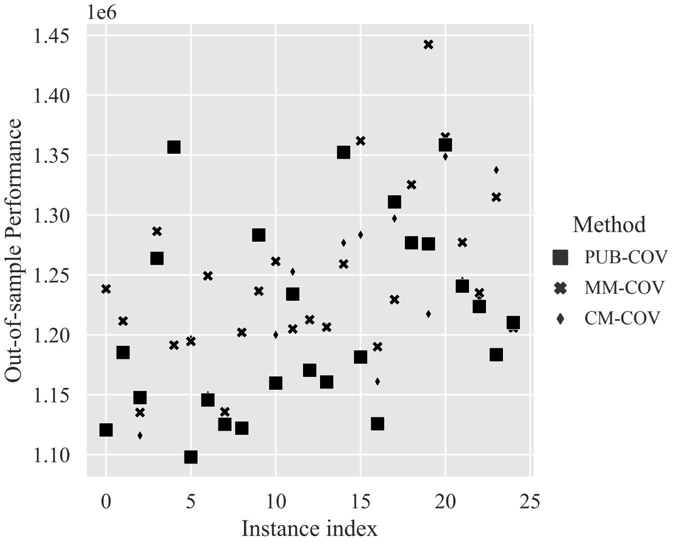


Figure 5 Comparison of out-of-sample cost in 25 problem instances over the years 1996-2021.

7. Conclusion

This paper studies the reliable uncapacitated facility location problem in a data-driven environment where past observations of random demands and disruptions are available. As an extension, we also investigate a novel situation in the context of RFL problem where covariate information is considered.

We propose an innovative prescriptive model based on the PUB on the Kolmogorov DRO estimator of the true operating cost, which is shown to be particularly effective in addressing the RFL problem. We derive the PUB estimator based on a novel CDF-based reformulation of the Kolmogorov DRO estimator. In contrast to the intractable Kolmogorov DRO estimator, our PUB estimator is tractable, and has favorable structures that yield significant computational benefits in obtaining data-driven RFL designs.

Our approach offers several attractive properties. *Scalability*: The data-driven RFL model with the PUB estimator is equivalent to an MILP problem that does *not* scale in the number of data points, and can be solved to optimality by a practically efficient constraint generation algorithm. *Performance guarantees*: Our approach is proved to be asymptotically optimal, and offers a theoretical guarantee for the out-of-sample performance in situations with limited samples. *Extendibility*: Our approach as well as the associated properties can be extended to the case where covariate information is incorporated.

We validate our theoretical results by thoroughly comparing the numerical performances between our model and several state-of-the-art RFL models, including the Wasserstein DRO model, the marginal moment-based model, and the cross moment-based model. Compared with benchmark RFL models, our model achieves not only better small-sample performance, but also considerably higher computational efficiency especially when applied to large-size networks and data sets.

However, the application of our results to some of the closely related reliable network design problems, including the capacitated RFL problem, the reliable location-inventory problem and the reliable location-routing problem, is not immediate and will be possible directions of future research. It is also of interest to investigate the gap between the PUB estimator and the Kolmogorov DRO estimator, and how to incorporate continuous-valued covariates in our data-driven optimization framework.

Acknowledgments

The authors thank the Department Editor, Prof. Chung Piaw Teo, the anonymous associate editor and reviewers for their constructive comments and suggestions. The authors thank Prof. J.George Shanthikumar for valuable discussion. Hao Shen acknowledges the support from the National Natural Science Foundation of China (72371240, 72001206). Mengying Xue acknowledges the support from National Natural Science Foundation of China (72201257). Zuo-Jun Max Shen acknowledges the support from the National Natural Science Foundation of China (71991462) and RGC Theme-based Research Scheme (No. T32-707/22-N).

Endnotes

[1] Note that the SAA approach can be viewed as a special case of the DRO where the ambiguity set is a singleton that only contains the empirical distribution.

[2] The finite sample performance guarantee for Problem (KDRO-RFL) by following identical arguments as in the proof of the theorem 3.5 of Esfahani and Kuhn (2018).

[3] We say a function $\varphi(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{B}^J$, is nonincreasing, if $\varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{x}')$ holds for all $\boldsymbol{x} \leq \boldsymbol{x}'$; $\varphi(\boldsymbol{x})$ is supermodular, if $\varphi(j|\boldsymbol{x}) \leq \varphi(j|\boldsymbol{x} \vee \boldsymbol{e}_J^k)$ holds for all $\boldsymbol{x} \in \mathbb{B}^J$ and $j, k \in \mathcal{J}$ such that $x_j = x_k = 0$, where $\varphi(j|\boldsymbol{x}) := \varphi(\boldsymbol{x} \vee \boldsymbol{e}_J^j) - \varphi(\boldsymbol{x})$.

[4] Here the data size N only affects the efficiency of computing the coefficients of constraints, which can be obtained efficiently as the time required for computing each coefficient is linear in N; see the proof of Lemma 3.

[5] The matrix is generated using the python package "data sets.make_spd_matrix"

[6] Computational time shown in Table EC.1 of some methods may exceed 900 seconds because preprocessing time in Gurobi is included.

[7] Data source: NOAA's Storm Prediction Center. A point or a short line is obtained by linking start and the end position of an event.

[8] NOAA data is downloaded from https://www.ncdc.noaa.gov/stormevents/ftp.jsp

References

Aboolian R, Cui T, Shen ZJM (2013) An efficient approach for solving reliable facility location models. *INFORMS J. Comput.* 25(4):720–729.

An Y, Zeng B, Zhang Y (2014) Reliable *p*-median facility location problem: Two-stage robust models and algorithms. *Transportation Res. Part B: Methodological* 64:54–72.

Banker S (2016) Using weather data to improve supply chain resiliency, (Jun 29, 2016) https://www.forbes.com/sites/stevebanker/2016/06/29/using-weather-to-improve-supply-chain-resiliency/#23a609a823f2.

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *The Operations Research Revolution*, 1–19 (INFORMS).

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Berman O, Krass D, Menezes M (2007) Facility reliability issues in network p-median problems: Strategic centralization and co-location effects. *Oper. Res.* 55(2):332–350.

Bertsimas D, Gupta V, Kallus N (2018a) Data-driven robust optimization. *Math. Programming* 167(2):235–292.

Bertsimas D, Gupta V, Kallus N (2018b) Robust sample average approximation. *Math. Programming* 171(1):217–282.

Bertsimas D, Shtern S, Sturt B (2022) A data-driven approach to multistage stochastic linear optimization. *Management Science* .

Chen Q, Li X, Ouyang Y (2011) Joint inventory-location problem under the risk of probabilistic facility disruptions. *Transportation Res. Part B: Methodological* 45(7):991–1003.

Chen Z, Sim M, Xiong P (2020) Robust stochastic optimization made easy with rsome. *Management Sci.* 66(8):3329–3339.

Cheng C, Adulyasak Y, Rousseau LM (2021) Robust facility location under demand uncertainty and facility disruptions. *Omega* 102429.

Cheng C, Qi M, Zhang Y, Rousseau LM (2018) A two-stage robust approach for the reliable logistics network design problem. *Transportation Research Part B: Methodological* 111:185–202.

Church RL, Scaparra MP (2007) Protecting critical assets: The r-interdiction median problem with fortification. *Geographical Anal.* 39(2):129–146.

Cui T, Ouyang Y, Shen ZJM (2010) Reliable facility location design under the risk of disruptions. *Oper. Res.* 58(4):998–1011.

Erdoğan E, Iyengar G (2006) Ambiguous chance constrained problems and robust optimization. *Math. Programming* 107(1):37–61.

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

Hao Z, He L, Hu Z, Jiang J (2020) Robust vehicle pre-allocation with uncertain covariates. *Production Oper. Management* 29(4):955–972.

Hu Z, Hong LJ (2013) Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online* .

Jiang R, Guan Y (2016) Data-driven chance constrained stochastic program. *Math. Programming* 158(1):291–327.

Kapadia S (2021) Winter storm slams texas food supply chains, logistics networks, https://www.supplychaindive.com/news/winter-storm-texas-food-grocery-heb-supply-chains-logistics/595354/.

Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12(2):479–502.

Ledbetter K (2021) Texas grocery stores emptied by winter storm, (February 23, 2021) https://today.tamu.edu/2021/02/23/texas-grocery-stores-emptied-by-winter-storm/.

Li X, Ouyang Y (2010) A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. *Transportation Res. Part B: Methodological* 44(4):535–548.

Li Y, Li X, Shu J, Song M, Zhang K (2022) A general model and efficient algorithms for reliable facility location problem under uncertain disruptions. *INFORMS J. Computing* 34(1):407–426.

Liberatore F, Scaparra MP, Daskin MS (2012) Hedging against disruptions with ripple effects in location analysis. *Omega* 40(1):21–30.

Lim AE, Shanthikumar JG, Shen ZM (2006) Model uncertainty, robust optimization, and learning. *Models, Methods, and Applications for Innovative Decision Making*, 66–94 (INFORMS).

Lim M, Daskin MS, Bassamboo A, Chopra S (2013) Facility location decisions with random disruptions and imperfect estimation. *Manufacturing Service Oper. Management* 15(2):239–249.

Lu M, Ran L, Shen ZJM (2015) Reliable facility location design under uncertain correlated disruptions. *Manufacturing Service Oper. Management* 17(4):445–455.

Luo F, Mehrotra S (2020) Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Lett.* 14(8):2565–2594.

Naaman M (2021) On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters* 173:109088.

Nemhauser GL, Wolsey LA (1981) Maximizing submodular set functions: Formulations and analysis of algorithms. *Ann. Discrete Math.* 59:279–301.

Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quantitative Finance* 7(4):435–442.

Qi L, Shen ZJM, Snyder LV (2009) A continuous-review inventory model with disruptions at both supplier and retailer. *Production Oper. Management* 18(5):516–532.

Shen H, Liang Y, Shen ZJM (2021) Reliable hub location model for air transportation networks under random disruptions. *Manufacturing & Service Operations Management* 23(2):388–406.

Shen H, Liang Y, Shen ZJM, Teo CP (2019) Reliable flexibility design of supply chains via extended probabilistic expanders. *Production Oper. Management* 28(3):700–720.

Shen ZJM, Zhan RL, Zhang J (2011) The reliable facility location problem: Formulations, heuristics, and approximation algorithms. *INFORMS J. Comput.* 23(3):470–482.

Snyder LV, Atan Z, Peng P, Rong Y, Schmitt AJ, Sinsoysal B (2016) Or/ms models for supply chain disruptions: A review. *IIE Transactions* 48(2):89–109.

Snyder LV, Daskin MS (2005) Reliability models for facility location: the expected failure cost case. *Transportation Sci.* 39(3):400–416.

Speare-Cole R (2021) Texans face empty supermarket shelves as weather crisis hits supply chains, (Feburary 22, 2021), https://www.newsweek.com/texas-empty-supermarket-shelves-weather-crisis-supply-chains-1570971.

Van Parys BP, Esfahani PM, Kuhn D (2021) From data to decisions: Distributionally robust optimization is optimal. *Management Sci.* 67(6):3387–3402.

Wozabal D (2012) A framework for optimization under ambiguity. *Ann. Oper. Res.* 193(1):21–47.

Xie S, An K, Ouyang Y (2019) Planning facility location under generally correlated facility disruptions: Use of supporting stations and quasi-probabilities. *Transportation Res. Part B: Methodological* 122:115–139.

Xie S, Li X, Ouyang Y (2015) Decomposition of general facility disruption correlations via augmentation of virtual supporting stations. *Transportation Res. Part B: Methodological* 80:64–81.

Xie W (2020) Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ wasserstein ball. *Operations Research Letters* 48(4):513–523.

Yang Z, Aydın G, Babich V, Beil DR (2009) Supply disruptions, asymmetric information, and a backup production option. *Management Sci.* 55(2):192–209.

Zhang Y, Snyder LV, Qi M, Miao L (2016) A heterogeneous reliable location model with risk pooling under supply disruptions. *Transportation Res. Part B: Methodological* 83:151–178.