Second-order Conic Programming Approach for Wasserstein Distributionally Robust Two-stage Linear Programs

Zhuolin Wang\(^a\), Keyou You\(^a,\ast\), Shiji Song\(^a\), Yuli Zhang\(^b\)

\(^a\)Department of Automation and BNRist, Tsinghua University, Beijing 100084, P.R. China
\(^b\)School of Management and Economics, Beijing Institute of Technology, Beijing 100081, P.R. China

Abstract

This paper proposes a second-order conic programming (SOCP) approach to solve distributionally robust two-stage stochastic linear programs over 1-Wasserstein balls. We start from the case with distribution uncertainty only in the objective function and exactly reformulate it as an SOCP problem. Then, we study the case with distribution uncertainty only in constraints, and show that such a robust program is generally NP-hard as it involves a norm maximization problem over a polyhedron. However, it is reduced to an SOCP problem if the extreme points of the polyhedron are given as a prior. This motivates to design a constraint generation algorithm with provable convergence to approximately solve the NP-hard problem. In sharp contrast to the exiting literature, the distribution achieving the worst-case cost is given as an “empirical” distribution by simply perturbing each sample for both cases. Finally, experiments illustrate the advantages of the proposed model in terms of the out-of-sample performance and the computational complexity.

Keywords: Uncertainty modelling, Distribution Uncertainty, Data-driven Robust, Wasserstein Ball, Two-stage Linear Program

1. Introduction

The two-stage program is one of the most fundamental tasks in mathematical optimization and has broad applications, see e.g., [Neyshabouri & Berg (2017); Paul & Zhang (2019)]. It is observed that its coefficients are usually uncertain and ignoring their uncertainties may lead to poor decisions [Hanasusanto & Kuhn 2018; Xie 2019]. In the literature, classical robust optimization (RO) has been proposed to handle the uncertainty in the two-stage problem. It restricts the uncertain coefficients to some given sets and minimizes the worst-case cost over all possible realizations of the coefficients [Ben-Tal et al. 2009]. However,
the RO ignores their distribution information and hence tends to lead a conservative solution (Bertsimas et al., 2011). To this end, the stochastic program (SP) is used to address the uncertainty. Specifically, an exact distribution of uncertain coefficients is required (Shapiro et al., 2009). In practice, the sample-average approximation (SAA) method (Shapiro & Homem-de Mello, 1998) is adopted to solve the SP by approximating the true distribution with a discrete empirical distribution over the sample dataset. However, this method is effective only when adequate and high-quality samples can be obtained cheaply (Shapiro & Homem-de Mello, 1998). When samples are of low quality, the empirical distribution may significantly deviate from the true distribution and leads to poor out-of-sample performance of the SAA method. Hence, the SAA method is not always reliable.

An alternative approach is to apply the distributionally robust (DR) optimization technique to the two-stage program with stochastic uncertainty (Shapiro & Kleywegt, 2002). The DR method assumes that the true distribution belongs to an ambiguity set of probability distributions, over which an optimal solution in the worst-case sense is to be found. This method overcomes inherent drawbacks of the SP and RO, since it does not require the exact distribution and can exploit the sample information. In fact, numerous evidence implies that the DR method can yield high-quality solutions within a reasonable computation cost (Luo & Mehrrotra, 2019; Wiesemann et al., 2014). Thus, our exposition concentrates on DR two-stage programs over an ambiguity set of distributions.

The ambiguity set is essential in the DR two-stage programs, as the associated DR programs are generically NP-hard if it is not appropriately constructed (Xu & Burer, 2018). Moreover, it should be large enough to include the true distribution with a high probability but cannot be too “large” to avoid very conservative decisions. Bertsimas et al. (2010, 2018); Hanasusanto et al. (2016); Ling et al. (2017) adopt the moment-based ambiguity set, which is composed of distributions with specified moment constraints. If the first- and second-order moment are known exactly, the DR two-stage linear program with uncertainty only in the objective function over this ambiguity set can be reformulated either as a semidefinite program (Bertsimas et al., 2010) or the mixed-integer linear program of a polynomial size (Hanasusanto et al., 2016) under different settings. Observe that the moment mismatch of insufficient samples may lead to poor decisions, and Ling et al. (2017) further considers the moment uncertainty, which results in an intractable model.

In this work, we study a data-driven DR two-stage linear program over an ambiguity set, which is a ball centered at the empirical distribution over a finite sample dataset, and its radius reflects our confidence in the empirical distribution. Particularly, the lower the confidence in the empirical distribution, the larger the
ball radius. Hence, the sample dataset can be utilized in a flexible way to handle the distribution uncertainty, e.g., the degree of conservatism can be controlled by tuning the radius. Moreover, our model applies to the situation where the exact distribution is slowly time-varying.

Note that the empirical distribution is discrete and the true distribution is usually continuous. We adopt the 1-Wasserstein metric to measure the distance between two distributions, which is different from the Kullback-Leibler divergence in Love & Bayraksan (2015) and $L^1$-norm in Jiang & Guan (2018). Since the Wasserstein ball contains the true distribution with a high probability (Esfahani & Kuhn 2018), the proposed DR problem is expected to exhibit good out-of-sample performance. Particularly, the Wasserstein ball can asymptotically degenerate to the true distribution as the sample size increases to infinity (Esfahani & Kuhn 2018). Then, we obtain the DR two-stage stochastic linear program over 1-Wasserstein balls and develop an SOCP approach to solve it.

This work considers that the distribution uncertainty arises either in the objective function or constraints of the two-stage linear programs, which has been widely studied in the literature, see e.g. Bertsimas et al. (2010, 2018); Ling et al. (2017); Xie (2019). Specifically, we first study the case with distribution uncertainty only in the objective function and exactly reformulate it as an SOCP problem. Then we proceed to the case with the distribution uncertainty only in constraints and show that such a program is generally NP-hard as it requires to solve a norm maximization problem over a polyhedron. The good news is that the resulting program can be reduced to an SOCP problem if the extreme points of the polyhedron are given as a prior. Motivated by this and also inspired by Zeng & Zhao (2013), we design a novel constraint generation algorithm with provable convergence to approximately solve it.

It should be noted that the DR two-stage linear programs have also been studied in Hanasusanto & Kuhn (2018) and Xie (2019) with the 2-Wasserstein and $\infty$-Wasserstein metrics, respectively. In Hanasusanto & Kuhn (2018), the distribution uncertainty arises simultaneously in the objective function and constraints, which renders their model NP-hard, and a sequence of co-positive programs are constructed to approximately solve it. In Xie (2019), the DR model is reformulated as a mixed-integer problem, which is again computational demanding. In comparison, we equivalently reformulate our model with distribution uncertainty in the objective as an SOCP problem and design an SOCP approach with provable convergence to solve the NP-hard problem with uncertainty in constraints. Moreover, the distribution that achieves the worst-case cost in our model can be explicitly given as an “empirical” distribution for both cases by simply perturbing each sample, by which we can further assess the quality of an optimal decision. This is clearly in
sharp contrast to Hanasusanto & Kuhn (2018); Xie (2019) and Bertsimas et al. (2018) with moment-based ambiguity set. We highlight our contributions as follows:

- We propose a novel SOCP approach to solve data-driven DR two-stage linear programs over 1-Wasserstein balls.

- The proposed model with distribution uncertainty only in the objective function is exactly reformulated as a solvable SOCP problem.

- The model with uncertainty only in constraints is proved to be NP-hard. Then, an SOCP-based constraint generation algorithm with provable convergence is developed to approximately solve it.

- The good out-of-sample performance and the low computational complexity of our model are demonstrated by experiments.

The rest of this paper is organized as follows. Section 2 proposes the DR two-stage linear program over the 1-Wasserstein ball. In Section 3 we reformulate our model with the distribution uncertainty only in the objective function as a tractable SOCP problem. Section 4 studies the model with uncertainty only in constraints and presents an SOCP-based constraint generation algorithm. Section 5 derives the distribution achieving the worst-case cost. Section 6 reports numerical results to illustrate the performance of the proposed model and the paper is concluded in Section 7.

**Notation.** We denote the set of real numbers by $\mathbb{R}$ and the set of positive real number by $\mathbb{R}_+$. The boldface lowercase letter denotes a vector of appropriate dimensions, e.g, $\mathbf{x} \in \mathbb{R}^n$ is a vector represented by $[x_1, \ldots, x_n]^T$. Special vectors with compatible dimensions include $\mathbf{0}$ and $\mathbf{e}$, respectively corresponding to the zero vector and the vector of all ones. $\| \cdot \|_p$ denotes the standard $l_p$-norm. We use the shorthand $[N] = \{1, 2, \ldots, N\}$ to represent a set of all integers up to $N$ and $|\mathcal{E}|$ to denote the cardinality of the set $\mathcal{E}$. The letters s.t. are an abbreviation of the phrase “subject to”.

2. Problem Formulation

2.1. The Two-stage Stochastic Linear Optimization

Consider the classical two-stage stochastic linear program (Birge & Louveaux 2011)

$$\min_{\mathbf{x} \in \mathcal{X}} \quad c^T \mathbf{x} + \mathbb{E}_{\mathbb{P}}[Q(x, \xi)].$$

(1)
where $x \in \mathbb{R}^n$ is the first-stage decision vector chosen from a compact set $X$ and should be decided before the realization of a random vector $\xi \in \mathbb{R}^m$ with the distribution $\mathbb{F}$.

The second-stage cost is evaluated based on the expectation of the following recourse problem

$$Q(x, \xi) = \min \quad z(\xi)^Ty$$

s.t. $A(\xi)x + By \geq b(\xi)$

$y \in \mathbb{R}_+^m$,  

(2)

where $B \in \mathbb{R}^{k \times m}$ is the recourse matrix and $z(\xi) \in \mathbb{R}^m$, $A(\xi) \in \mathbb{R}^{k \times n}$ and $b(\xi) \in \mathbb{R}^k$ depend on the random vector $\xi$.

In the sequel, we study models with uncertainty only in the objective function or constraints of problem (2) respectively, each of which is motivated by two notable examples, see also [Bertsimas et al. 2010, 2018; Ling et al. 2017].

**Example 1.** ([Ling et al. 2017]) Consider a portfolio program with $n$ assets which investors can invest in two stages. Let $x_i$ and $y_i$ be the dollar invested on the $i$-th asset at the first stage and the second-stage respectively. Generally the return vector $\xi$ for assets in the second stage is random, hence a stochastic two-stage portfolio program is designed for a maximum return

$$\min_{e^T x = 1, x \geq 0} \quad -(e + c)^T x + \mathbb{E}[Q(x, \xi)].$$

(3)

where $c_i$ is the return of the $i$-th asset in the first stage and $c \in \mathbb{R}^n$ denotes the vector of the return over all assets and

$$Q(x, \xi) = \min \quad z(\xi)^Ty$$

s.t. $y \geq 0, \Delta^s \geq 0, \Delta^b \geq 0$

$$Ax + (1 - \theta)\Delta^b - (1 + \theta)\Delta^s = y,$$

where $z(\xi) = -(e + \xi)$, $A = \text{Diag}(e + c)$ and $\text{Diag}(\cdot)$ denotes a diagonal matrix with vector $(\cdot)$ being its diagonal elements. $\Delta^s_i$ and $\Delta^b_i$ are the dollar of buying and selling the $i$-th asset respectively and $\theta$ is the transaction cost.

**Example 2.** ([Kall et al. 1994]) Consider the material order problem with $n$ raw materials and $m$ desired products. Let $b_i$ denote the market demand for product $i$ and $b \in \mathbb{R}^m$ be the vector of the demand over all products. Let $a_{ij}$ represent the amount of product $i$ produced by per unit of material $j$ and $A = [a_{ij}]_{m \times n}$ be the matrix of the production amount for all materials. The demand $b_i$ is usually time-varying and the
uncertainty in \(a_{ij}\) is generally inevitable due to the quality of raw materials. Hence, it is unavoidable to introduce uncertainty \(\xi\) to \(b_i\) and \(a_{ij}\).

The order problem aims to find an optimal order \(x \in \mathbb{R}^n\) for \(n\) raw materials to minimize the cost and satisfy market demands, i.e.,

\[
\min_{x^T x \leq u, \ x \geq 0} \left\{ c^T x + \mathbb{E}_\xi [Q(x, \xi)] \right\},
\]

where \(u\) is the capacity of \(n\) materials, \(c_i\) is the cost of material \(i\) and \(c \in \mathbb{R}^n\) denotes the cost vector for all materials. \(Q(x, \xi)\) is given as

\[
Q(x, \xi) = \min \ z^T y \\
\text{s.t.} \ A(\xi)x + y \geq b(\xi) \\
y \in \mathbb{R}^m,
\]

where \(z_i\) is the penalty for per unit of undeliverable product \(i\) and \(y_i\) is the corresponding shortage amount.

Motivated by above examples, this paper considers that \(z(\xi), A(\xi)\) and \(b(\xi)\) in (1) depend affinely on \(\xi\), i.e.,

\[
z(\xi) = z_0 + \sum_{i=1}^{m} \xi_i z_i, \ A(\xi) = A_0 + \sum_{i=1}^{m} \xi_i A_i, \ b(\xi) = b_0 + \sum_{i=1}^{m} \xi_i b_i,
\]

where vectors \(z_0, z_1, \ldots, z_m \in \mathbb{R}^m, b_0, b_1, \ldots, b_m \in \mathbb{R}^k\) and matrices \(A_0, A_1, \ldots, A_m \in \mathbb{R}^{k \times n}\) are given constants. In fact, the affine uncertainty provides a strong modeling power and has been adopted in Bertsimas et al. (2018); Ling et al. (2017); Nemirovski & Shapiro (2006).

The following condition is needed to guarantee the feasibility of the second-stage problem in (2) and is satisfied by many problems, e.g., the newsvendor problem and its variants as well as the production planning problem Birge & Louveaux (2011).

**Assumption 1.** The second-stage problem in (2) is always feasible for any \(x \in \mathcal{X}\) and \(\xi\).

### 2.2. Distributionally Robust Two-stage Problems

The program in (1) generally requires an exact distribution \(F\) of \(\xi\). In practice, \(F\) can only be estimated through a finite sample dataset \(\{\hat{\xi}_i\}_{i=1}^{N}\) and a common idea is to adopt the SAA method, where \(F\) is approximated by an empirical distribution \(\hat{F}_N\) over the sample dataset, i.e.,

\[
\hat{F}_N(\xi) = \frac{1}{N} \sum_{i=1}^{N} 1[\hat{\xi} \leq \xi],
\]
where $I_A$ is the indicator of event $A$. Then the stochastic linear problem in (1) is approximated by

$$\minimize_{x \in \mathcal{X}} \left\{ c^T x + \frac{1}{N} \sum_{i=1}^{N} Q(x, \hat{\xi}) \right\}.$$ (8)

By Glivenko-Cantelli theorem (Cantelli, 1933), the distribution $F_N$ weakly converges to the true distribution $F$ as $N$ increases to infinity. This implies the asymptotic convergence to the stochastic model (1) for (8). Hence, the SAA method is sensible only if $F_N$ well approximates the true distribution $F$.

However, insufficient and/or low-quality samples may lead to an empirical distribution $F_N$ that is far from the true distribution $F$. Thus, the SAA model (8) may be not reliable and display poor out-of-sample performance.

As in Esfahani & Kuhn (2018), a data-driven approach is adopted to address the distribution uncertainty in this paper. We assume that $F$ belongs to an ambiguity set $\mathcal{F}_N$ including all distributions no further than $\epsilon_N$-distance from the empirical distribution $F_N$. Here $\epsilon_N$ indicates the confidence on $F_N$, e.g., the larger the $\epsilon_N$, the lower the confidence on $F_N$.

Since the true distribution $F$ is generally continuous and the empirical distribution $F_N$ is discrete, the 1-Wasserstein metric (Kantorovich & Rubinshtein, 1958) is adopted to measure their distance and consequently a 1-Wasserstein ball $\mathcal{F}_N$ is obtained. Then we are interested in the worst-case second-stage cost over $\mathcal{F}_N$, i.e.,

$$\beta(x) = \sup_{F \in \mathcal{F}_N} \mathbb{E}_F[Q(x, \xi)],$$ (9)

and the DR two-stage linear program is formulated as

$$\minimize_{x \in \mathcal{X}} c^T x + \beta(x).$$ (10)

To evaluate an optimal solution, we also derive the worst-case distribution $F^*$ that achieves the worst-case expectation of the recourse problem, i.e.,

$$\sup_{F \in \mathcal{F}_N} \mathbb{E}_F[Q(x, \xi)] = \mathbb{E}_{F^*}[Q(x, \xi)].$$ (11)

### 2.3. Ambiguity Set via the 1-Wasserstein Metric

The definition of the $r$-Wasserstein metric is given below.

**Definition 1.** (Kantorovich & Rubinshtein (1958)) Let $d(\xi_1, \xi_2) = \|\xi_1 - \xi_2\|_p$ denote the $l_p$-norm of $\xi_1 - \xi_2$ on $\mathbb{R}^n$, and $(\Xi, d)$ is a Polish metric space. Given a pair of distributions $F_1 \in \mathcal{M}(\Xi)$ and $F_2 \in \mathcal{M}(\Xi)$ where
\( \mathcal{M}(\Xi) \) is a set containing all probability distributions supported on \( \Xi \), the \( r \)-Wasserstein metric \( \mathcal{W}^r : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}_+ \) is defined as

\[
\mathcal{W}^r(\mathbb{F}_1, \mathbb{F}_2) = \inf \left\{ \left( \int_{\Xi} d(\xi_1, \xi_2)^r K(d\xi_1, d\xi_2) \right)^{1/r} : \right. \\
\left. \begin{align*}
\int_{\Xi} K(\xi_1, d\xi_2) &= \mathbb{F}_1(\xi_1), \\
\int_{\Xi} K(d\xi_1, \xi_2) &= \mathbb{F}_2(\xi_2)
\end{align*} \right\},
\]

where \( r \geq 1 \) and \( K \) is a joint distribution with its marginal distributions being \( \mathbb{F}_1 \) and \( \mathbb{F}_2 \).

In order to avoid the infinity value of \( \mathcal{W}^r \) and ensure it exactly a real metric \( \text{[Ambrosio & Gigli, 2013]} \), we need the following requirement on the set \( \mathcal{M}(\Xi) \) without sacrificing much modeling power.

**Assumption 2.** For any distribution \( \mathbb{F} \in \mathcal{M}(\Xi) \), it holds that

\[
\int_{\Xi} \|x\|^p \mathbb{F}(dx) < \infty.
\]

Different from \text{[Hanasusanto & Kuhn, 2018]} and \text{[Xie, 2019]}, we adopt the 1-Wasserstein metric and \( l_2 \)-norm, i.e., \( r = 1 \) and \( p = 2 \) in (12) to construct the ambiguity ball \( \mathcal{F}_N \),

\[
\mathcal{F}_N = \{ \mathbb{F} \in \mathcal{M}(\Xi) : \mathcal{W}^1(\mathbb{F}_N, \mathbb{F}) \leq \epsilon_N \},
\]

where \( \epsilon_N > 0 \) is the ball radius, i.e., \( \mathcal{F}_N \) is the set of distributions no further than \( \epsilon_N \)-distance away from the \( \mathbb{F}_N \).

### 2.4. Comparisons with the state-of-the-art methods

In \text{[Bertsimas et al., 2018]}, the ambiguity set in (13) of the DR two-stage linear programs is defined as a set of distributions with the specified second-order moment constraints.

\text{[Hanasusanto & Kuhn, 2018]} considers DR two-stage linear programs of the form (10) with 2-Wasserstein balls, i.e., \( r = p = 2 \) in (12), and \( Q(x, \xi) \) is defined as

\[
Q(x, \xi) = \min \ (Q \xi + q)^T y \\
\text{s.t.} \ T(x) \xi + h(x) \leq By
\]

where \( T(\cdot) \) and \( h(\cdot) \) are two affine functions.

In \text{[Xie, 2019]}, the DR two-stage program is defined via the \( \infty \)-Wasserstein metric, i.e., \( r = \infty \) and \( p = 1, \infty \) in (12) with the uncertainty only in the objective function or constraints separately, i.e., \( Q \) or \( T(x) \) in (14) is set to 0 respectively.

Comparisons with these state-of-art models are further summarized as follows:
• **Model differences:** Clearly, $Q(x, \xi)$ in (2) of this work and [Bertsimas et al., 2010, 2018; Ling et al., 2017] is mathematically different from (14) in [Bertsimas et al., 2018] and [Xie, 2019]. Our model is motivated from a wide range of real applications, see e.g. Examples 1-2.

• **Solving approaches:** [Hanasusanto & Kuhn, 2018] derives a sequence of co-positive programs to approximate their NP-hard DR two-stage model. [Xie, 2019] reformulates his model as a generally NP-hard mixed-integer problem. [Bertsimas et al., 2018] approximate their model by linear decision rule techniques. In this work, we *equivalently* reformulate our model with distribution uncertainty only in the objective as an SOCP problem and design an SOCP-based constraint generation algorithm for the problem with distribution uncertainty only in constraints.

• **Approximation gaps:** There is no approximation gap in [Hanasusanto & Kuhn, 2018] and [Bertsimas et al., 2018], under the condition that for any $t \in \mathbb{R}^k$, there exists a solution $y$ to solve the inequality $By \geq t$ (aka complete recourse). In this work, the feasibility condition in Assumption 1 (aka relatively complete recourse) is weaker and is satisfied by numerous real application models (Birge & Louveaux, 2011).

As explicitly stated in [Bertsimas et al., 2018], “there are also problems that would generally not satisfy complete recourse, such as a production planning problem where a manager determines a production plan today to satisfy all uncertain demands for tomorrow instead of incurring penalty”, see Example 2 for details. However, it satisfies relatively complete recourse.

• **The worst-case distribution:** In sharp contrast to these state-of-art models, this work derives the worst-case distribution attaining the worst-case second-stage cost with distribution uncertainty either in the objective function or constraints, respectively.

### 3. Uncertainty in the Objective Function

We first consider the distribution uncertainty only in the objective function of (2) via the following form

$$Q(x, \xi) = \min \ z(\xi)^T y$$

s.t. $Ax + By \geq b$

$$y \in \mathbb{R}^m_+.$$
where $z(\xi)$ is defined as (7) in Section 2.1.

In this section, we convert problem (10) with $Q(x, \xi)$ given by (15) over the 1-Wasserstein ball $\mathcal{F}_N$ to an SOCP problem which can be solved efficiently by general-purpose commercial-grade solvers such as CPLEX.

**Theorem 1.** Under Assumptions 1-2, the worst-case second-stage cost (9) with $Q(x, \xi)$ given in (15) over the 1-Wasserstein ball $\mathcal{F}_N$ can be computed by a finite SOCP problem

$$\beta(x) = \inf \left\{ \lambda \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} s_i \right\}$$

s.t. $\lambda \geq \|Zy\|_2$

$$s_i \geq z_0^Ty + y^T Z^T \hat{\xi}^i, \forall i \in [N]$$

$$Ax + By \geq b, y \geq 0.$$  

(16)

Moreover, the associated DR problem (10) is equivalent to the following SOCP problem

$$\min_{x \in X} \left\{ c^T x + \lambda \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} s_i \right\}$$

subject to $\lambda \geq \|Zy\|_2$

$$s_i \geq z_0^Ty + y^T Z^T \hat{\xi}^i, \forall i \in [N]$$

$$Ax + By \geq b, y \geq 0.$$  

(17)

Proof. For any feasible first-stage decision vector $x$, the $\beta(x)$ over the 1-Wasserstein ball can be obtained by solving the following conic linear program

$$\beta(x) = \sup \sum_{i=1}^{N} \int_{\Xi} Q(x, \xi) K(d\xi, \hat{\xi}^i)$$

s.t. $\int_{\Xi} K(d\xi, \hat{\xi}^i) = \frac{1}{N}, \forall i \in [N]$  

$$\int_{\Xi} \sum_{i=1}^{N} d(\xi, \hat{\xi}^i) K(d\xi, \hat{\xi}^i) \leq \epsilon_N.$$  

(18)

Consider a Lagrangian function for (18), i.e.,

$$L(\xi, \lambda, s) = \int_{\Xi} \sum_{i=1}^{N} Q(x, \xi) K(d\xi, \hat{\xi}^i) - \int_{\Xi} \sum_{i=1}^{N} s_i K(d\xi, \hat{\xi}^i) - \int_{\Xi} \sum_{i=1}^{N} \lambda d(\xi, \hat{\xi}^i) K(d\xi, \hat{\xi}^i) + \frac{1}{N} \sum_{i=1}^{N} s_i + \lambda \epsilon_N.$$  

(10)
The Lagrange dual function is represented as
\[ g(\lambda, s) = \sup_{\xi \in \Xi} L(\xi, \lambda, s) = \sup_{\xi \in \Xi} \left( \int_{\Xi} \sum_{i=1}^{N} \left( Q(x, \xi) - s_i - \lambda d(\xi, \hat{\xi}^i) \right) K(d(\xi, \hat{\xi}^i)) \right) + \frac{1}{N} \sum_{i=1}^{N} s_i + \lambda \epsilon_N. \]

Consequently, the dual problem of (18) is given as
\[ \beta(x) = \inf \lambda \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} s_i \]
\[ \text{s.t.} \quad Q(x, \xi) - \lambda d(\xi, \hat{\xi}^i) \leq s_i, \forall i \in [N], \xi \in \Xi \]
\[ \lambda \geq 0. \]  

Note that \( K = F_N \times F_N \) is a strictly feasible solution to (18) due to \( \epsilon_N > 0 \), the Slater condition for the strong duality of primal problem (18) and its dual problem (19) is satisfied (Shapiro, 2001).

The constraints in (20) require a feasible second-stage decision vector \( \hat{y} \) to guarantee the feasibility of the following inequality for all \( \xi \in \Xi \)
\[ z(\xi)^T \hat{y} - \lambda d(\xi, \hat{\xi}^i) \leq s_i, \forall i \in [N]. \]

Note that Assumption 1 ensures the existence of such a \( \hat{y} \). Hence, constraints (20) can be expressed as
\[ s_i \geq z(\xi)^T \hat{y} - \lambda d(\xi, \hat{\xi}^i), \forall i \in [N], \xi \in \Xi. \]  

The following equality
\[ z(\xi)^T \hat{y} = \left( z_0 + \sum_{i=1}^{m} \xi_i z_i \right)^T \hat{y} = z_0^T \hat{y} + \xi^T Z \hat{y} \]
leads to
\[ \sup_{\xi} \left\{ z(\xi)^T \hat{y} - \lambda \| \xi - \hat{\xi}^i \|_2 \right\} = \sup_{\xi} \left\{ z_0^T \hat{y} + \xi^T Z \hat{y} - \lambda \| \xi - \hat{\xi}^i \|_2 \right\} = \begin{cases} z_0^T \hat{y} + \hat{y}^T Z \hat{y}, & \text{if } \| Z \hat{y} \|_2 \leq \lambda \\ +\infty, & \text{if } \| Z \hat{y} \|_2 > \lambda \end{cases} \]
where the equality follows from Lemma 1 in Wang et al. (2019).

Consequently, constraints in (20) admit an equivalent form
\[ \begin{cases} s_i \geq z_0^T \hat{y} + \hat{y}^T Z \hat{y}, \forall i \in [N], \\ \lambda \geq \| Z \hat{y} \|_2. \end{cases} \]
Substituting above inequities into (20) leads to the equivalence of (16) and (9). Hence, the two-stage problem (10) can be equivalently reformulated as the SOCP problem (17).
Theorem 1 shows that the optimization program (10) can be reformulated as a tractable SOCP problem. Furthermore, different $l_p$-norms lead to different equivalent forms of the DR two-stage optimization program. For example, (10) can be reformulated as a linear programming (LP) problem when we adopt $l_1$-norm in the 1-Wasserstein metric, see Table 1 for details.

| Norm $p$ | $p = 1$ | $p = 2$ | $p = \infty$ | Otherwise |
|----------|---------|---------|--------------|-----------|
| DR Problem | LP | SOCP | LP | Convex Program |

Table 1: Equivalent problems of the our DR problem, where $p$ represents the $l_p$-norm in (12).

4. Uncertainty in the Constraints

In this section we consider the distribution uncertainty only in constraints of (2), i.e.,

$$Q(x; \xi) = \min z^T y$$

s.t. $A(\xi)x + By \geq b(\xi)$

$$y \in \mathbb{R}^m_+,$$

where $A(\xi)$ and $b(\xi)$ are defined in (7) of Section 2.1.

4.1. Reformulation of the DR Problem

In this subsection we prove the equivalence for the two-stage program (10) with $Q(x; \xi)$ given by (22) and an NP-hard single-stage problem.

**Theorem 2.** Under Assumptions 1-2, the worst-case second-stage cost (9) with $Q(x; \xi)$ given in (22) over the 1-Wasserstein ball $\mathcal{F}_N$ is equivalent to the following NP-hard problem

$$\beta(x) = \inf \left\{ \lambda \varepsilon_N + \frac{1}{N} \sum_{i=1}^{N} s_i \right\}$$

s.t. $\lambda \geq \|Cp\|_2, \forall p \in \mathcal{P}$

$$s_i \geq (Cp)^T \xi_i + p^T (b_0 - A_0 x), \forall i \in [N], p \in \mathcal{P},$$

where

$$C = [b_1 - A_1 x, \ldots, b_m - A_m x]^T$$

and $\mathcal{P}$ is a polyhedron given by

$$\mathcal{P} = \{ p \in \mathbb{R}^k_+: B^T p \leq d \}.$$
Proof. The strong duality stills holds for the $\beta(x)$, which allows us to rewrite it as

$$\beta(x) = \inf \lambda \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} s_i$$

(28)

subject to

$$Q(x, \xi) - \lambda d(\xi, \hat{\xi}) \leq s_i, \forall i \in [N], \xi \in \Xi$$

(29)

$$\lambda \geq 0.$$ Under the strong duality of the LP problem, $Q(x, \xi)$ in (22) is equivalent to

$$Q(x, \xi) = \max p^T (b(\xi) - A(\xi)x)$$

(30)

subject to

$$z \geq B^T p$$

$$p \geq 0.$$ Then the constraints in (29) can be expressed as

$$s_i \geq p^T (b(\xi) - A(\xi)x) - \lambda d(\xi, \hat{\xi}), \forall \xi \in \Xi, p \in P.$$ (31)

Further, the right-hand side of (31) can be re-expressed as

$$\sup_{\xi} \left\{ p^T (b(\xi) - A(\xi)x) - \lambda d(\xi, \hat{\xi}) \right\}$$

$$= \sup_{\xi} \left\{ (Cp)^T \xi + p^T (b_0 - A_0x) - \lambda d(\xi, \hat{\xi}) \right\}$$

$$= \begin{cases} (Cp)^T \hat{\xi}i - p^T (b_0 - A_0x), & \text{if } \|Cp\| \leq \lambda \\ +\infty, & \text{if } \|Cp\| > \lambda, \end{cases}$$

where $C$ is defined in (26) and the equality holds due to Lemma 1 in Wang et al. (2019).

Consequently, constraints in (29) are equivalent to

$$\begin{cases} s_i \geq (Cp)^T \hat{\xi}i - p^T (b_0 - A_0x), & \forall i \in [N], p \in P \\ \lambda \geq \|Cp\|, & \forall p \in P. \end{cases}$$

Thus, the $\beta(x)$ in (9) is reformulated as problem (23). This completes the first part of the proof.

Next we prove the NP-hardness for the problem (10) with $Q(x, \xi)$ given by (22) over the 1-Wasserstein ball. The constraint (24) in problem (23) can be expressed as

$$\lambda \geq \max_{p \in P} \|Cp\|,$$
The norm maximization problem over the polyhedron is shown to be NP-complete in Bodlaender et al. (1990) and hence checking the feasibility of constraint (24) is NP-hard. Thus, the problem (23) and the corresponding two-stage optimization program (10) are both NP-hard due to their equivalence.

Theorem 2 proves the NP-hardness of the DR problem (10) with $Q(x,\xi)$ given by (22). However, when the extreme point set $E = \{p_1, \ldots, p_P\}$ of the polyhedron $P$ in (27) is explicitly known, (10) can be reformulated as an SOCP problem.

**Corollary 1.** Under Assumptions 1-2 and given the extreme point set $E$ of the polyhedron $P$ in (27), the DR problem (10) with $Q(x,\xi)$ given in (22) over the 1-Wasserstein ball $T_N$ is equivalent to the following SOCP problem

$$
\min_{x \in \mathbb{X}} \left\{ c^T x + \lambda \epsilon + \frac{1}{N} \sum_{i=1}^{N} s_i \right\}
$$

subject to

$$
s_i \geq (Cp)^T \hat{\xi}^i - p^T (b^0 - A^0 x), \quad \forall i \in [N], \ p \in E,
$$

$$
\lambda \geq \|Cp\|_2, \quad \forall p \in E.
$$

**Proof.** Given the extreme point set $E$, the following equality

$$
Q(x,\xi) = \max_{peE} p^T (b(\xi) - A(\xi) x),
$$

holds since the LP problem (30) attains its optimal value at an extreme point of its feasible set $P$.

Then the constraints in (24) and (25) can be explicitly expressed as

$$
\begin{align*}
& s_i \geq (Cp)^T \hat{\xi}^i - p^T (b^0 - A^0 x), \quad \forall i \in [N], \ p \in E, \\
& \lambda \geq \|Cp\|_2, \quad \forall p \in E,
\end{align*}
$$

which leads to the equivalence of (32) and (10). Hence we complete the proof.

Corollary 1 shows that we can solve the DR two-stage problem by explicitly enumerating the extreme points of the polyhedron $P$. Motivated by this, we design an algorithm to approximately solve the NP-hard DR two-stage problem by a constraint generation approach.

### 4.2. Approximately Solving the DR Two-stage Problem with Uncertainty in Constraints

In this subsection, we propose a constraint generation algorithm to solve (10). Inspired by Corollary 1, the proposed DR problem can be efficiently solved given the set of extreme points. While the direct enumeration of all extreme points is computational demanding, we can gradually select sets of “good”
extreme points by solving a sequence of second-stage problems $Q(x, \xi)$ with different given first-stage decision vectors $x$. Hence, we utilize a master-subproblem framework to approximately solve (10).

In the master problem (MP), an optimal solution under a selected subset of extreme points is derived. Then a subproblem (SuP), i.e., the second-stage problem based on this sub-optimal solution is solved to find a better subset of extreme points. We add these points to the subset in MP as feasible cuts. Note that the optimal value of the MP and SuP are the lower and upper bounds for (10) respectively, by iteratively introducing the extreme points and computing MP. Both the lower and upper bounds will converge and a good solution to (10) can be obtained. The algorithm based on such an MP-SuP framework is given in the sequel.

By Corollary [1] the master problem (MP) is an SOCP problem and given as

$$
\begin{align*}
\text{minimize} & \quad \left\{ c^T x + \lambda \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} s_i \right\} \\
\text{subject to} & \quad s_i \geq (Cp)^T \hat{\xi}_i + p^T (b_0 - A_0 x_m), \forall i \in [N], \quad p \in E_s, \\
& \quad \lambda \geq \|Cp\|_2, \quad \forall p \in E_s,
\end{align*}
$$

(33)

where $E_s$ is a given subset of extreme points of $P$.

After obtaining an optimal solution $x^m$ of the MP, an SuP is derived as follows

$$
Q(x^m, \xi) = \min_{d^i, s^i} \left\{ \lambda^i \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} s^i \right\} \\
\text{s.t.} \quad \lambda \geq \|Cp\|_2, \quad \forall p \in P \\
\quad s^i \geq (Cp)^T \hat{\xi}_i + p^T (b_0 - A_0 x^m), \forall p \in P, i \in [N].
$$

(34)

The following condition is needed to obtain an good solution of the SuP.

**Assumption 3.** The polyhedron $P = \{ p \in \mathbb{R}^k_+ : B^T p \leq z \}$ is nonempty and bounded.

By (34), the decision variables $\lambda^i$ and $s^i$ are completely decoupled, which implies that we can separately find their optimal solutions. To achieve it, we have the following steps.

1. An optimal solution $s^i$ to SuP is obtained by solving a group of linear programs, i.e.,

$$
\begin{align*}
s^i = \max & \quad (Cp)^T \hat{\xi}_i + p^T (b_0 - A_0 x^m) \\
\text{s.t.} & \quad p \in P,
\end{align*}
$$

(35)
2. An optimal $\lambda^*$ is obtained by solving a norm maximization problem, i.e.,

$$\lambda^* = \max \|Cp\|_2$$

s. t. $p \in \mathcal{P}$.  \hfill (36)

The LP problem (35) can be solved efficiently and a sequence of optimal solutions \{p_i^*\}_{i=1}^{N}$ from (35) are then added to the set of extreme point set $E_s$ of (33), since the LP problem (35) obtains its optimal value at extreme points in the feasible region $\mathcal{P}$.

To solve the norm maximization problem, which clearly is a non-convex quadratic problem, we adopt the consensus alternating direction method of multipliers (ADMM) method (Huang & Sidiropoulos, 2016). Particularly, (36) is reformulated as a consensus form by introducing $m$ auxiliary variables \{g_1, \ldots, g_m\}, i.e.,

$$\lambda^* = \min -p^T C^T Cp$$

s. t. $b_i^T g_i \leq z_i, g_i \geq 0$

$$g_i = p, \quad \forall i \in [m],$$

where $b_i$ is the $i$-th column of matrix $B$. The detailed consensus-ADMM algorithm is given in Algorithm 1.

Note that the parameter $\rho$ in Algorithm 1 should be carefully to ensure $-C^T C/\rho + mI$ being a semi-positive matrix. We omit the proof for the convergence of Algorithm 1 for brevity, which can be found in Huang & Sidiropoulos (2016).

By Assumption 3, a solution $p^*$ is ensured to exist as an extreme point of polyhedron $\mathcal{P}$ and hence is also added to the extreme point set $E_s$ of (33) as a feasible cut (Malyscheff et al., 2002).

The algorithm based on the MP-SuP framework to solve the DR two-stage problem (10) is given in Algorithm 2. In Theorem 3, we show that Algorithm 2 terminates in a finite number of iterations.

**Theorem 3.** Under Assumption 3, Algorithm 2 generates an optimal solution of (10) in $O(|E_s|)$ iterations.

**Proof.** Assume that in the $k$-th iteration, \{x_k, \lambda_k, s_k\} are optimal solutions of MP, \{\lambda_k^i, s_k^i\} are optimal solutions of SuP and \{p_k\}_{i=1}^{N} \cup \{p_k\}$ are the corresponding extreme points derived in SuP. We show that \{p_k\}_{i=1}^{N} \cup \{p_k\} $\subseteq E_s$, implies the convergence of Algorithm 2, i.e., $LB = UB$.

Step 2 in Algorithm 2 implies that

$$UB \leq c^T x_k + \frac{1}{N} \sum_{i=1}^{N} s_k^i + \epsilon_N \lambda_k^i.$$

Moreover, since \{p_k\}_{i=1}^{N} \cup \{p_k\} $\subseteq E_s$, MP in the $k$-th iteration is identical to that in the $(k - 1)$-th iteration. Thus, $x_k$ is an optimal solution to the $(k - 1)$-th MP as well. From the step 1 in Algorithm 2, we find that
Algorithm 1 The consensus-ADMM for (37)
Input: matrix $B$, $C$, vector $z$, $g_i$ and $u_i$, tolerance $\tau$
Output: An optimal solution $p^*$ and the corresponding optimal value $\lambda^*$
1: Initialize $g_i$ and $u_i$
2: repeat
3: $p \leftarrow \left(-\frac{c^T C}{\rho} + mI\right) \left(\sum_{i=1}^m (g_i + u_i)\right)$
4: for each $i \in [m]$ do
5: $z_i \leftarrow \arg\min_{z_i} \|z_i - x + u_i\|^2$
6: subject to $z_i = p$, $g_i \geq 0$ $u_i \leftarrow g_i + u_i - p$
7: end for
8: until The successive difference of $p$ is smaller than $\tau$
9: Return $p^* \leftarrow p$, $\lambda^* \leftarrow \|Cp^*\|_2$ and terminate

Algorithm 2 Solve the Robust program
Input: Extreme point set $E_s = \{\text{initial extreme points}\}$, $UB = +\infty$, $LB = -\infty$, iteration index $k = 0$
Output: Optimal Solution $x^*$
1: repeat
2: Solve the master problem (33) and derive an optimal solution $(x_k, s_k, \lambda_k)$, update the lower bound
   \[ LB = c^T x_k + \lambda_k \epsilon_N + \frac{1}{N} \sum_{i=1}^N s_{ki} \]
3: Solve the subproblem (34) and obtain the optimal solution $(s^e_{ki})_{i=1}^N \cup \{\lambda^e_k\}$ as well as the corresponding optimal solution $(p^e_{ki})_{i=1}^N \cup \{p_k\}$, update the upper bound
   \[ UB = \min\{UB, c^T x_k + Q(x_k, \xi)\} \]
   where $Q(x_k, \xi) = \frac{1}{N} \sum_{i=1}^N s^e_{ki} + \epsilon_N \lambda^e_k$
4: Add extreme points $(p^e_{ki})_{i=1}^N \cup \{p_k\}$ to set $E_s$ and set $k = k + 1$
5: until $UB - LB \leq \epsilon$
6: Return $x^* \leftarrow x_k$ and terminate
\[ LB \geq e^T x_k + \epsilon_N \lambda_k + \frac{1}{N} \sum_{i=1}^{N} s_{ki} \geq e^T x_k + \frac{1}{N} \sum_{i=1}^{N} s_{ki}^i + \epsilon_N \lambda_k, \] where the last inequality holds due to the fact that \( \{p_i^j\}_{i=1}^N \cup \{p_k\} \subseteq E \), and hence the related constraints are already added to MP before or in the \((k - 1)\)-th iteration.

Consequently, we have \( UB = LB \). The conclusion of the convergence in \( O(|E|) \) iterations follows immediately from the finite number of extreme points for the polyhedron \( P \).

5. The Worst-case Distribution and the Asymptotic Consistency

5.1. The Worst-case Distribution

In this subsection we derive the distribution achieving the worst-case \( \beta(x) \) in (9) of Section 2.2 for any feasible vector \( x \in X \).

**Lemma 1.** Fix a feasible first-stage decision vector \( x \), the \( \beta(x) = \sup_{F \in \mathcal{F}_N} \mathbb{E}_F \{ Q(x, \xi) \} \) is equivalent to

\[
\sup_{\tilde{\xi} \in \mathcal{B}} \left\{ \frac{1}{N} \sum_{i=1}^{N} Q(x, \tilde{\xi}^{(i)}) \right\},
\]

where \( \mathcal{B} = \left\{ (\tilde{\xi}^{(1)}, \ldots, \tilde{\xi}^{(N)}) \mid \frac{1}{N} \sum_{i=1}^{N} d(\tilde{\xi}^{(i)}, \hat{\xi}^{(i)}) \leq \epsilon_N, \tilde{\xi}^{(i)} \in \Xi \right\} \).

**Proof.** Given a solution \( x \), it follows that

\[
\sup_{\tilde{\xi} \in \mathcal{B}} \left\{ \frac{1}{N} \sum_{i=1}^{N} Q(x, \tilde{\xi}^{(i)}) \right\} \leq \sup_{F \in \mathcal{F}_N} \mathbb{E}_F \{ Q(x, \xi) \},
\]

which holds due to the Lemma 2 in Wang et al. (2019). Hence \( \beta(x) \) is no less than (38).

Next we show that \( \beta(x) \) is no greater than (38). Due to the equivalence between \( \beta(x) \) and problem (19), the following inequality

\[
\sup_{F \in \mathcal{F}_N} \mathbb{E}_F \{ Q(x, \xi) \} - \epsilon < \inf_{\lambda \geq 0} \left\{ \lambda \epsilon_N + \frac{1}{N} \sum_{i=1}^{N} \left[ Q(x, \tilde{\xi}^{(i)}) - \lambda d(\tilde{\xi}^{(i)}, \hat{\xi}^{(i)}) \right] \right\}
\]

holds for any \( \epsilon \geq 0 \) and certain \( \{\tilde{\xi}^{(i)}\}_{i \in [N]} \subseteq \Xi \).

If \( (\tilde{\xi}^{(1)}, \ldots, \tilde{\xi}^{(N)}) \notin \mathcal{B} \) and let \( \lambda > 0 \), it follows that

\[
\lambda \left\{ \epsilon_N - \frac{1}{N} \sum_{i=1}^{N} d(\tilde{\xi}^{(i)}, \hat{\xi}^{(i)}) \right\} < 0.
\]

Increasing \( \lambda \) to \( +\infty \) in (39) results in \( \sup_{F \in \mathcal{F}_N} \mathbb{E}_F \{ Q(x, \xi) \} = -\infty \). It contradicts the fact that

\[
\sup_{F \in \mathcal{F}_N} \mathbb{E}_F \{ Q(x, \xi) \} \geq \mathbb{E}_{F_x} \{ Q(x, \xi) \} > -\infty,
\]

18
where the second inequality holds due to the finite $Q(x, \xi)$ under Assumption 1. Thus, $(\tilde{\xi}^{(1)}, \ldots, \tilde{\xi}^{(N)}) \in B$.

By Lemma 2 in Wang et al. (2019), it holds that

$$\sup_{F \in \mathcal{F}_N} \mathbb{E}_F\{Q(x, \xi)\} - \epsilon < \sup_{\tilde{\xi} \in B} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( Q(x, \xi^{(i)}) \right) \right\}.$$

Decreasing $\epsilon$ to zero yields the desired inequality and we have that

$$\sup_{F \in \mathcal{F}_N} \mathbb{E}_F\{Q(x, \xi)\} = \sup_{\tilde{\xi} \in B} \left\{ \frac{1}{N} \sum_{i=1}^{N} Q(x, \xi^{(i)}) \right\}.$$

Since $Q(x, \xi)$ is concave with respect to $\xi$ and $B$ is a compact set, problem (38) allows for an optimal solution, i.e., we can substitute “sup” by “max” for problem (38). By Lemma 1, problem (38) is equivalent to $\beta(x)$ for any $x \in X$, then a worst-case distribution related to the solution for (38) is explicitly derived.

**Theorem 4.** Fix a solution $x \in X$ and let $\xi_x = (\xi^{(1)}_x, \ldots, \xi^{(N)}_x)$ denote an optimal solution to (38). Then the following distribution

$$\mathbb{F}_x^* = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\xi}^{(i)}},$$

is the distribution achieving the worst-case second-stage cost, i.e.,

$$\sup_{F \in \mathcal{F}_N} \mathbb{E}_F\{Q(x, \xi)\} = \mathbb{E}_{\mathbb{F}_x^*}\{Q(x, \xi)\}.$$

**Proof.** Obviously, the following distribution

$$\Pi_x = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{\xi}^{(i)}},$$

is a joint distribution of $F_N$ and $\mathbb{F}_x^*$. Then it holds that

$$W(F_N, \mathbb{F}_x^*) \leq \int \|\xi - \xi'\|_p \Pi_x(\xi, d\xi') = \frac{1}{N} \sum_{i=1}^{N} \|\xi^{(i)}_x - \tilde{\xi}^{(i)}\|_p \leq \epsilon_N,$$

where the first inequality follows immediately from the definition of the 1-Wasserstein metric and the last inequality follows from the fact that $(\xi^{(1)}_x, \ldots, \xi^{(N)}_x) \in B$. Hence, $\mathcal{F}_N$ includes the distribution $\mathbb{F}_x^*$. As a result, it yields that

$$\sup_{F \in \mathcal{F}_N} \mathbb{E}_F\{Q(x, \xi)\} \geq \mathbb{E}_{\mathbb{F}_x^*}\{Q(x, \xi)\} = \frac{1}{N} \sum_{i=1}^{N} \left( Q(x, \xi^{(i)}_x) \right) = \sup_{F \in \mathcal{F}_N} \mathbb{E}_F\{Q(x, \xi)\},$$

where the last equality follows from Lemma 1. Hence, $\mathbb{F}_x^*$ is a distribution in the 1-Wasserstein ball $\mathcal{F}_N$ achieving the worst-case cost. □
5.2. The Asymptotic Consistency

This subsection studies the asymptotic consistency of the DR problem (10) under the a mild assumption.

**Assumption 4.** The following inequality under some positive constant $c$

$$\int_{\Xi} \exp(\|\xi\|_2^2) F(d\xi) < \infty.$$  

holds for the true distribution $F$.

Under Assumptions 1-4, we formalize the asymptotic consistency of the proposed DR problem as follows.

**Theorem 5.** Under Assumptions 1-4 and fix $\beta_N \in (0, 1)$ with $\sum_{N=1}^{\infty} \beta_N < \infty$. Let

$$\epsilon_N(\beta_N) = \begin{cases} \left( \frac{\log(c_1 \beta^{-1})}{c_2^N} \right)^{1/\max\{n,2\}} , & \text{if } N \geq \frac{\log(c_1 \beta^{-1})}{c_2^N} \\ \left( \frac{\log(c_1 \beta^{-1})}{c_2^N} \right)^{1/e} , & \text{if } N < \frac{\log(c_1 \beta^{-1})}{c_2^N} \end{cases}$$

be the 1-Wasserstein ball radius, where $c_1 > 0$ and $c_2 > 0$ are constants related to the constant $c$ in Assumption 4. Then the minimum and minima of the DR problem (10) asymptotically converge to those of the stochastic problem (1) almost surely when the sample number increases to infinity.

**Proof.** For the problem with distribution uncertainty only in the objective function, the relatively complete recourse implies that $Q(x, \xi)$ is feasible and finite. Then there exists a finite $y$ such that $|Q(x, \xi)| = |(Zy)^T \xi| \leq \|Zy\|_2 \|\xi\|_2 \leq L(1 + \|\xi\|_2)$ holds for any $x \in X$ and $\xi \in \Xi$, where $L > 0$ is a constant.

As for the problem with the distribution uncertainty only in constraints, under the strong duality of LP problem it follows that $Q(x, \xi) = (C \bar{p})^T \xi$, where $C$ is defined as (26) in Section 4.1 and $\bar{p}$ is the extreme point of polyhedron $\mathcal{P}$. Assumption 4 implies that $\|\bar{p}\|$ is bounded and hence there also exists a positive constant $L$ such that $|Q(x, \xi)| \leq \|C \bar{p}\|_2 \|\xi\|_2 \leq L(1 + \|\xi\|_2)$ holds for $x \in X$ and $\xi \in \Xi$.

Moreover, since $Q(x, \xi)$ is continuous in $\xi$, the asymptotic consistency of our model follows from Theorem 3.6 in Esfahani & Kuhn (2018). \qed

6. Simulation

This section conducts experiments to evaluate the performance of the proposed model and the constraint generation algorithm. All experiments are conducted on a 64 bit PC with an Intel Core i5-7500 CPU at 3.4GHz and 8 GB RAM. Cplex 12.6 optimizer is used to solve the optimization programs.
6.1. Application to Two-stage Portfolio Optimization

This subsection is devoted to the application in two-stage portfolio program with uncertainty only arose in the objective function as stated in Example[1].

6.1.1. The Two-stage Portfolio program

Consider a portfolio consisting of four assets: (1) Dow Jones Industrial Average Index (DJI), (2) Dow Jones Transportation Average Index (DJT), (3) Dow Jones Composite Average Index (DJA) and (4) Dow Jones Utility Average (DJU). The daily returns of above assets over seven years spanning January 02th, 2011 to December 31th, 2018 are collected from RESSET database (http://www.resset.cn).

Note that the first-stage return $c$ is unknown in our simulation, we collect the data from January 02th, 2011 to December 31th, 2016 to approximate it by the SAA method, i.e., $c = \sum_{i=1}^{N} \hat{\xi}_i^1$, where $\hat{\xi}_i^1$ is the $i$th sample of the first-stage return. We introduce additional constraints described in [Ling et al., 2017] to the portfolio problem (3) to obtain a feasible solution. Since all constraints added in this problem are linear equalities or inequalities, it is still equivalently reformulated as a solvable SOCP problem by Theorem[1].

6.1.2. Impact of the 1-Wasserstein Radius and the Sample Size

Experiments are conducted to test the impact of the 1-Wasserstein radius $\epsilon_N$ and the sample size $N$ on the out-of-sample performance of our model in this subsection. The out-of-sample performance is measured by the loss of the proposed model on new samples, i.e.,

$$ c^T x + \frac{1}{\alpha} \mathbb{E}_F \{Q(x, \xi)\}. \tag{40} $$

We are unable to exactly calculate (40) due to the unknown true distribution $F$. Instead, we randomly choose 300 test samples from the dataset to approximate it, i.e.,

$$ c^T x + \frac{1}{\alpha N_T} \sum_{i=1}^{N_T} Q(x, \hat{\xi}_i^T), $$

where $\hat{\xi}_i^T$ is the $i$-th test sample and $N_T$ is the number of test samples.

We first test the impact of the 1-Wasserstein radius $\epsilon_N$ on the out-of-sample performance of our model. We conduct 200 independent experiments and the averaged out-of-sample performance is visualized in Figure[1]. Experimental results imply the improvement of the out-of-sample performance as the 1-Wasserstein radius increases and also its deterioration as the radius is greater than a specific value.
Figure 1: The averaged out-of-sample performance under sample dataset of different sizes as a function for 1-Wasserstein radius estimated by 200 independent simulation runs. (a) $N = 20$, (b) $N = 100$, (c) $N = 200$.

Experiments on different sample dataset is performed to evaluate the impact of the size. The out-of-sample performance averaged on 200 independent experiments is presented in Figure 2. The asymptotic consistency of our model stated in Theorem 5 is validated by the out-of-sample performance improvement with the growing sample dataset.

Figure 2: The averaged out-of-sample performance as a function of sample size $N$ for 200 independent experiments.

6.1.3. Comparisons with the State-of-the-art Methods

In this subsection, we compare the proposed DR model over the 1-Wasserstein ball (denoted as DRW) with the SAA method and the model over the moment-based ambiguity set (denoted as DRM). The moment-based ambiguity is the set of distributions with specific mean value uncertainty and ellipsoidal second-order uncertainty, we refer readers to Ling et al. (2017) for details. We set $N = \{20, 30, 50, 100, 200, 300\}$ in this subsection. Due to the dependence of the 1-Wasserstein radius $\epsilon_N$ on the sample dataset size, we tune $\epsilon_N$ for different $N$ to ensure a powerful out-of-sample performance.
We adopt the percentage difference
\[
\left( \frac{\text{DR}}{\text{SAA}} - 1 \right) \times 100\%
\]
to compare the out-of-sample performance of these models, where DR denotes the out-of-sample performance of solutions from the DR two-stage problem and SAA denotes that of the SAA method.

### Table 2: Percentage differences (in %) between the DR model and SAA in terms of out-of-sample performance

| N    | 20  | 30  | 50  | 100 | 200 | 300 |
|------|-----|-----|-----|-----|-----|-----|
| DRW  | 1.1 | 1.6 | 1.7 | 2.1 | 4.1 | 4.8 |
| DRM  | -1.3| -0.7| 0.7 | 1.5 | 3.6 | 3.5 |

### Table 3: Averaged computation time (second) of different methods

| N    | 20  | 30  | 50  | 100 | 200 | 300 |
|------|-----|-----|-----|-----|-----|-----|
| DRW  | 0.14| 0.15| 0.15| 0.17| 0.16| 0.19|
| DRM  | 0.12| 0.14| 0.14| 0.16| 0.15| 0.16|
| SAA  | 0.13| 0.15| 0.16| 0.17| 0.16| 0.16|

Comparisons in terms of the out-of-sample performance and computation time are presented in Table 2 and Table 3 respectively. A positive value in Table 2 implies a better performance of the DR method than the SAA. Table 2 indicates the best out-of-sample performance of our proposed method among all models. Importantly, our model can be solved in an acceptable time even under a large sample dataset.

### 6.2. Application to Two-stage Material Order Problem

This subsection we consider the application in material ordering problem with distribution uncertainty only in constraints as stated in Example 2, where \( A(\xi) \) and \( b(\xi) \) follows the definition in (7). The constraint generation algorithm is applied to solve the DR two-stage ordering problem.

In this section we omit the comparison with the model over the moment-based ambiguity set since there is no effective method to solve this NP-hard problem (Ling et al., 2017).

#### 6.2.1. The Two-stage Material Order Problem

Consider an example where a refinery supplies the gasoline and the fuel oil for a company weekly (see example of pp. 9-15 in Kall et al., 1994). The raw material of the products is the crude oil from two
countries, and can be viewed as different materials. The coefficients of (5) in our simulation are the same as those of the problem (2.10) in [Kall et al. (1994)], i.e.,

\[ c = [2, 3]^T, d = [7, 12]^T, u = 100, A(\xi) = \begin{bmatrix} 2 + \xi_1 & 3 \\ 6 & 3.4 + \xi_2 \end{bmatrix}, b(\xi) = \begin{bmatrix} 180 + \xi_3 \\ 162 + \xi_4 \end{bmatrix}, \]

where \( \xi \in \mathbb{R}^4 \) is a random vector with unknown distribution. In our simulation, we assume \( \xi \) follows a Gaussian distribution \( N(\mu, \Sigma) \) with \( \mu = [0, 0, 0, 0]^T \) and \( \Sigma = \text{Diag}(9, 12, 0.21, 0.16)^T \). We generate \( N \) samples to construct the 1-Wasserstein ball \( F_N \). The ordering plan problem of the refinery (6) defined in Example 2 is solved via the Algorithm 2 of Section 4.2.

6.2.2. Test the Tightness of Bounds

We test the tightness of the proposed bounds in MP and SP for optimal function value (O.F.V.) and the first-stage cost over the 1-Wasserstein ball with different radii \( \epsilon_N \). Obviously the set \( P = \{ p \geq 0 : p \leq d \} = \{ p \in \mathbb{R}_+^2 : p_1 \leq 7, p_2 \leq 12 \} \) is a box and its extreme points are \( [0, 0]^T, [0, 12]^T, [7, 0]^T \) and \( [7, 12]^T \). Hence, we also solve problem (10) directly with explicitly known extreme points and compare its solutions with those obtained from the Algorithm 2. Let \( (x_1^d, x_2^d) \) denote the solution obtained via solving (10) directly and \( (x_1^a, x_2^a) \) obtained by our algorithm. The results summarized in Table 4 indicates that the two methods under different 1-Wasserstein radius obtain identical results.

The performance of the proposed bounds for O.F.V. and the first-stage cost compared to that of the method with known extreme points under 500 samples is shown in Fig.3 and Fig.4. We observe that both the lower bound and upper bound are tight, irrespective of the size of the 1-Wasserstein ball. Thus, these bounds can be viewed as a good reference to verify the performance of our algorithm.

![Figure 3: Performance of the proposed bounds for O.F.V. under the 1-Wasserstein ball with different radii.](image-url)
Table 4: The optimal solutions obtained via different methods under different 1-Wasserstein ball radii $\epsilon_N$ when sample size $N = 500$

| $\epsilon_N$ | $(x_1^d, x_2^d)$ | $(x_1^a, x_2^a)$ |
|-------------|------------------|------------------|
| 0.01        | (42.75, 57.24)   | (42.75, 57.24)   |
| 0.11        | (43.64, 42.26)   | (43.64, 42.26)   |
| 0.21        | (41.21, 50.88)   | (41.21, 50.88)   |
| 0.31        | (39.99, 46.07)   | (39.99, 46.07)   |
| 0.41        | (38.72, 41.59)   | (38.72, 41.59)   |
| 0.51        | (36.80, 34.80)   | (36.80, 34.80)   |
| 0.61        | (36.02, 32.04)   | (36.02, 32.04)   |
| 0.71        | (35.56, 29.00)   | (35.56, 29.00)   |
| 0.81        | (34.79, 26.42)   | (34.79, 26.42)   |
| 0.91        | (33.49, 22.05)   | (33.49, 22.05)   |

Fig. 5 shows the tendency of the upper and lower bound for the proposed two-stage program in a single experiment. To eliminate the contingency in a single simulation, we record the averaged extreme point number generated in the algorithm and the iteration number of the proposed method under different sample sets over 100 independent experiments in Table 5 and Table 6. Results agree with the conclusion in Theorem 3 that our algorithm is able to converge within $O(|E|)$ iterations.

Figure 4: Performance of the proposed bounds for the first-stage cost under the 1-Wasserstein ball with different radii.
6.2.3. Test in High Dimension

It is hard to enumerate all extreme points of the polyhedron \( \mathcal{P} = \{ p \in \mathbb{R}^M_+ : B^T p \leq d \} \) especially when \( p \) is in a high dimension (Khachiyan et al., 2009), e.g., it is unreliable to enumerate all extreme points of \( \mathcal{P} \) if \( p \in \mathbb{R}^{20} \) since it can have \( 2^{20} \) extreme points at most. In this subsection, we consider a problem with variables in a high dimension to verify the efficiency of our algorithm, i.e.,

\[
\begin{align*}
  u &= 1000, \ x \in \mathbb{R}^{20}, \ A(\xi) \in \mathbb{R}^{20 \times 20}, \ b(\xi) \in \mathbb{R}^{20}, \\
  c &= [2, 3, 1, 4, 5, 2, 4, 3, 4, 2, 5, 4, 2, 6, 2, 4, 3, 1, 2]^T, \\
  d &= [7, 9, 4, 6, 8, 5, 6, 8, 10, 7, 12, 10, 6, 7, 9, 5, 11, 10, 5, 8]^T,
\end{align*}
\]

where \( A(\xi) \) and \( b(\xi) \) are affinely dependent on the random vector \( \xi \) and the recourse matrix \( B \) is the identity matrix.

Note that it is unpractical to enumerate all extreme points of polyhedron \( \mathcal{P} \), hence we omit comparisons between our proposed bounds and results obtained by solving problem (10) with explicitly known extreme points in this subsection.

Fig.6 and Fig.7 report the averaged performance of our proposed bounds for O.F.V and the first-stage cost under different 1-Wasserstein radii \( \epsilon_N \) when the sample size \( N = 500 \). Similar to the results in previous
subsection, these proposed bounds are tight as well.

![Graph of O.F.V](image)

**Figure 6:** Performance of the proposed bounds for O.F.V. under the 1-Wasserstein ball with different radii.

![Graph of First-stage cost](image)

**Figure 7:** Performance of the proposed bounds for the first-stage cost under the 1-Wasserstein ball with different radii.

We record the averaged computation time, the averaged extreme point number generated in Algorithm 2 and the averaged iteration number over 100 independent simulations as sample size $N$ varies from 10 to 1000 in Table 7, Table 8 and Table 9 respectively. The convergence of the proposed algorithm in a single experiment is also visualized in Fig. 8.

| $N$  | 10  | 20  | 30  | 50  | 100 | 200 | 300 | 500 | 1000 |
|------|-----|-----|-----|-----|-----|-----|-----|-----|------|
| Time | 10.93 | 11.63 | 11.67 | 11.62 | 12.37 | 13.95 | 17.22 | 23.03 | 36.24 |

Table 7: The averaged computation time (second) under different sample sizes $N$

Results show that Algorithm 2 converges in a reasonable time even for the problem in high dimension under a large sample dataset. The number of extreme points generated in our algorithm is far smaller than the total number ($2^{20}$), which indicates the low efficiency to enumerate all extreme points to solve the DR problem.
Table 8: The averaged extreme points number under different sample sizes $N$

| $N$  | 10  | 20  | 30  | 50  | 100 | 200 | 300 | 500  | 1000 |
|------|-----|-----|-----|-----|-----|-----|-----|------|------|
| Num  | 35.28 | 46.52 | 49.02 | 60.38 | 72.44 | 100.18 | 123.76 | 156.46 | 181.10 |

Table 9: The averaged iteration number under different sample sizes $N$

| $N$  | 10  | 20  | 30  | 50  | 100 | 200 | 300 | 500  | 1000 |
|------|-----|-----|-----|-----|-----|-----|-----|------|------|
| Ite  | 10.28 | 9.58 | 9.32 | 8.92 | 8.80 | 8.54 | 8.58 | 8.16  | 8.46  |

Figure 8: The convergence of the O.F.V for the two-stage program with 500 samples.

7. Conclusion

We have proposed a novel SOCP approach to solve the data-driven DR two-stage linear programs over 1-Wasserstein balls. The model with distribution uncertainty in the objective function is reformulated as a solvable finite SOCP problem. While the DR model over the moment-based ambiguity set is generally unsolvable, we propose a constraint generation algorithm with provable convergence to approximately solve the NP-hard model with distribution uncertainty only in constraints. We explicitly derive a distribution achieving the worst-case cost. Numerical results validate the good out-of-sample performance for our model and the high efficiency of the proposed algorithm.

Acknowledgements

The authors would like to thank the Associate Editor and anonymous reviewers for their very constructive comments, which greatly improved the quality of this work. This work was supported by the National Natural Science Foundation of China [grant numbers 61722308, U1660202 and 71871023].
References

Ambrosio, L., & Gigli, N. (2013). A user’s guide to optimal transport. In Modelling and optimisation of flows on networks (pp. 1–155). Springer.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). Robust optimization volume 28. Princeton University Press.

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. SIAM review, 53, 464–501.

Bertsimas, D., Doan, X. V., Natarajan, K., & Teo, C.-P. (2010). Models for minimax stochastic linear optimization problems with risk aversion. Mathematics of Operations Research, 35, 580–602.

Bertsimas, D., Sim, M., & Zhang, M. (2018). Adaptive distributionally robust optimization. Management Science, 65, 604–618.

Birge, J. R., & Louveaux, F. (2011). Introduction to stochastic programming. Springer Science & Business Media.

Bodlaender, H. L., Gritzmann, P., Klee, V., & Van Leeuwen, J. (1990). Computational complexity of norm-maximization. Combinatorica, 10, 203–225.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita (On the empirical determination of the laws of probability). Giorn. Ist. Ital. Attuari, 4, 421–424.

Esfahani, P. M., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming, 171, 115–166.

Hanasusanto, G. A., & Kuhn, D. (2018). Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. Operations Research, 66, 849–869.

Hanasusanto, G. A., Kuhn, D., & Wiesemann, W. (2016). K-adaptability in two-stage distributionally robust binary programming. Operations Research Letters, 44, 6–11.

Huang, K., & Sidiroopoulos, N. D. (2016). Consensus-admm for general quadratically constrained quadratic programming. IEEE Transactions on Signal Processing, 64, 5297–5310.

Jiang, R., & Guan, Y. (2018). Risk-averse two-stage stochastic program with distributional ambiguity. Operations Research, 66, 1390–1405.

Kall, P., Wallace, S. W., & Kall, P. (1994). Stochastic programming. Springer.

Kantorovich, L., & Rubinshtein, G. (1958). On a space of totally additive functions, vestn. Vestn Lening. Univ, 13, 52–59.

Khachiyan, L., Boros, E., Borys, K., Gurvich, V., & Elbassioni, K. (2009). Generating all vertices of a polyhedron is hard. In Twentieth Anniversary Volume: (pp. 1–17). Springer.

Ling, A., Sun, J., Xiu, N., & Yang, X. (2017). Robust two-stage stochastic linear optimization with risk aversion. European Journal of Operational Research, 256, 215–229.

Love, D., & Bayraksan, G. (2015). $\phi$-divergence constrained ambiguous stochastic programs for data-driven optimization. Technical report, Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio,

Luo, F., & Mehrrot, S. (2019). Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. European Journal of Operational Research, 278, 20–35.

Malycheff, A., Tafalis, T., & Raman, S. (2002). From support vector machine learning to the determination of the minimum enclosing zone. Computers & industrial engineering, 42, 59–74.
Nemirovski, A., & Shapiro, A. (2006). Convex approximations of chance constrained programs. *SIAM Journal on Optimization, 17*, 969–996.

Neyshabouri, S., & Berg, B. P. (2017). Two-stage robust optimization approach to elective surgery and downstream capacity planning. *European Journal of Operational Research, 260*, 21–40.

Paul, J. A., & Zhang, M. (2019). Supply location and transportation planning for hurricanes: A two-stage stochastic programming framework. *European Journal of Operational Research, 274*, 108–125.

Shapiro, A. (2001). On duality theory of conic linear problems. *Nonconvex Optimization and Its Applications, 57*, 135–165.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

Shapiro, A., & Kleywegt, A. (2002). Minimax analysis of stochastic problems. *Optimization Methods and Software, 17*, 523–542.

Shapiro, A., & Homem-de Mello, T. (1998). A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming, 81*, 301–325.

Wang, Z., You, K., Song, S., & Zhang, Y. (2019). Wasserstein distributionally robust shortest path problem. *arXiv preprint arXiv:1902.09128*, .

Wiesemann, W., Kuhn, D., & Sim, M. (2014). Distributionally robust convex optimization. *Operations Research, 62*, 1358–1376.

Xie, W. (2019). Tractable reformulations of distributionally robust two-stage stochastic programs with $\infty-$ Wasserstein distance. *arXiv preprint arXiv:1908.08454*, .

Xu, G., & Burer, S. (2018). A copositive approach for two-stage adjustable robust optimization with uncertain right-hand sides. *Computational Optimization and Applications, 70*, 33–59.

Zeng, B., & Zhao, L. (2013). Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters, 41*, 457–461.