Distributionally Robust Two-Stage Stochastic Programming

Daniel Duque, Sanjay Mehrotra, and David P. Morton
Department of Industrial Engineering and Management Sciences
Northwestern University
September 28, 2020

Abstract

Distributionally robust optimization is a popular modeling paradigm in which the underlying distribution of the random parameters in a stochastic optimization model is unknown. Therefore, hedging against a range of distributions, properly characterized in an ambiguity set, is of interest. We study two-stage stochastic programs with linear recourse in the context of distributional ambiguity, and formulate several distributionally robust models that vary in how the ambiguity set is built. We focus on the Wasserstein distance under a p-norm, and an extension, an optimal quadratic transport distance, as mechanisms to construct the set of probability distributions, allowing the support of the random variables to be a continuous space. We study both unbounded and bounded support sets, and provide guidance regarding which models are meaningful in the sense of yielding robust first-stage decisions. We develop cutting-plane algorithms to solve two classes of problems, and test them on a supply-allocation problem. Our numerical experiments provide further evidence as to what type of problems benefit the most from a distributionally robust solution.

Keywords: Distributionally robust optimization, two-stage stochastic programming, Wasserstein distance, optimal transport distance.

1 Introduction

Two-stage stochastic programming with recourse is a natural modeling framework for applications in which a here-and-now decision needs to be made, followed by the realization of random parameters. For such two-stage models, the recourse action consists of additional decisions to be made after observing the random parameters. The objective is typically to optimize a cost function in expectation, where often the expectation is estimated using a sample mean via sample average approximation (SAA). Alternatively, a distributionally robust two-stage stochastic program (DRTSP) assumes ambiguity in the probability distribution, and the goal is to optimize for the "worst-case" expected cost. The worst-case expectation is with respect to a well-defined set of probability distributions. The problem lends itself to a min-max formulation, where the minimization problem is with respect the original here-and-now decision and the maximization problem is with respect a probability distribution in an ambiguity set. This approach to decision making under uncertainty is often data-driven because the ambiguity set is informed by limited historical data regarding the random parameters.

If the ambiguity set comprises all probability distributions on a given support—including degenerate distributions with unit mass on a single support point—then DRTSP reduces to a parameter

robust variant of the problem [3]. Instead, we are interested in more tightly constrained ambiguity sets for three possible settings. First, we may wish to limit probability distributions to those that are statistically and physically plausible in the context of available data and other contextual information on the random parameters. Second, we may seek risk-averse decisions and view distributionally robust optimization (DRO) as an attractive way to model risk aversion. Third, we may not subscribe to the view that nature will select a probability distribution that adapts to our decisions and is designed to maximize our expected cost, but we recognize the potential regularizing benefits of DRO assessed through out-of-sample testing, when nature's choice is appropriately constrained.

There are several modeling choices needed to characterize the ambiguity set of a DRTSP. Depending on the application of interest, the first choice for the ambiguity set is whether candidate distributions are supported on a given finite set (e.g., scenarios from historical data) or an infinite and possibly unbounded set (e.g., $\mathbb{R}^m$ or an unbounded subset thereof). A second choice further specifies which distributions belong to the ambiguity set. For example, ambiguity sets can be built based on assumed moments [8, 9, 18], likelihood functions [17], goodness-of-fit measures from hypothesis tests [4, 5], divergence "distances" [2, 12], and Wasserstein distances [6, 10, 14, 19, 21], among others. See Rahimian and Mehrotra [15] for a review of such ambiguity sets.

In this paper we focus on Wasserstein distances, and more general optimal transport distances, due to their modeling flexibility to describe various types of ambiguity sets with established asymptotic and finite-sample guarantees [14]. In contrast to divergence-type distances, optimal transport distances allow us to model ambiguity sets containing probability distributions supported at atoms that need not be in the available data. This modeling choice is useful in a risk-averse setting in which available data may not suffice to cover realizations of the random parameters that a decision maker would like to hedge against. Moment-based ambiguity sets also allow us to go beyond the support suggested by available data, but most of the practical approaches in this category do not enjoy asymptotic or finite-sample guarantees. The Wasserstein distance has been extensively used in recent literature for various DRO problems; see, e.g., [6, 10, 14, 19, 20, 21]. For two-stage stochastic programs, there are several problem reformulations that leverage duality [11, 14, 19, 21]. The main difficulty in solving the type of DRTSP we consider manifests in one of two ways: either the problem admits a row-generation procedure in which the subproblem involves bilinear terms, or the reformulation requires knowing all the extreme points of the second-stage dual problem. Hanasusanto and Kuhn [11] take a major step in bridging theory and computational tractability with a polynomially sized linear programming reformulation that applies for a particular type of ambiguity set. Xie [19] revisits the setting in [11] and proposes several reformulations that are computationally tractable, including models in which the random parameters are supported on a mixed-binary set.

We study a class of two-stage stochastic programs in which random parameters appear only on the right-hand side of the constraints of the second-stage linear programming (LP) subproblem.

We identify cases wherein—somewhat surprisingly—an unbounded support defining the ambiguity set yields an optimal solution identical to that of the underlying SAA problem. In particular, we find that when using the Wasserstein distance and the support of the random parameters is an unbounded convex cone, there is no point in solving the DRTSP. With this negative result in hand, we give guidance in formulating meaningful DRTSPs. In particular, we next analyze the case in which the support of the random parameters is a hyper-rectangle, and we provide a mixed-integer programming reformulation of the second-stage subproblems, which generalizes the approach in [20] for the unit commitment problem. Then, we investigate the role of optimal transport functions, which generalize the Wasserstein distance, and we revisit our analysis of unbounded support sets in which a quadratic transport function is used, in the spirit of Blanchet et al. [7].

The structure of the paper is as follows. Section 2 formalizes the DRTSP that we consider, and summarizes relevant results from the literature. In Section 3, we analyze the cases of convex-cone and box-type support sets, and propose a computationally tractable algorithm for the latter case. Section 4 extends the analysis to a quadratic transport function, and we study its relative merits in the context of DRTSP. Section 5 presents computational results on a supply-allocation problem. Section 6 concludes.

Notation

We denote the set $\{1, \ldots, n\}$ by $[n]$. Subindices denote a component of a vector while superindices enumerate vectors. All vectors and matrices are properly sized so that algebraic operations make sense. For two vectors $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^m$, $a^\top b$ denotes its inner product. The set $\mathbb{R}_+$ denotes the non-negative reals. We use $\|\cdot\|$ to denote an L_p-norm with $p \geq 1$, and $c(\cdot, \cdot)$ is a more general cost function. For a polytope, Π, $\text{ext}(\Pi)$ denotes its finite set of extreme points.

2 Distributionally robust two-stage stochastic programming

In this paper we focus on the following problem:

$$\min_{x \in \mathcal{X}} \left\{ c^\top x + \max_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[V(x, \xi) \right] \right\}, \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^{d_x}$, $\mathcal{P}$ is an ambiguity set of probability distributions supported on $\Xi \subseteq \mathbb{R}^m$, ξ is a random vector that follows probability distribution $\mathbb{P}$, and the cost function is defined as:

$$V(x, \xi) = \min_{y \geq 0} f^\top y \tag{2a}$$

$$\text{s.t.} \quad Dy = Bx + \xi, \tag{2b}$$

where $y \in \mathbb{R}^{d_y}$ and f, D, and B are of appropriate dimensions. The first-stage feasible region, $\mathcal{X}$, can be polyhedral but could instead include, e.g., integer restrictions. In what follows we make use of the following linear programming dual of (2):

$$V(x, \xi) = \max_{\pi \in \Pi} \pi^\top (Bx + \xi), \tag{3}$$

where $\Pi \equiv \{\pi \,:\, D^\top \pi \leq f\}$.

Given a nominal distribution of the random vector, e.g., from historical data, we focus on a class of ambiguity sets, $\mathcal{P}$, that are centered on this distribution. Let $\mathbb{Q}$ be the nominal distribution of the random vector, ξ, defined on the finite support $\Xi^n \equiv \{\xi^1, \ldots, \xi^n\}$ with $\mathbb{Q}(\xi = \xi^\sigma) = \frac{1}{n}$ for all $\sigma \in [n]$. Further let $\mathcal{M}(\Xi)$ be the set of all probability distributions supported on Ξ and let $d : \mathcal{M}(\Xi^n) \times \mathcal{M}(\Xi) \to \mathbb{R}_+$ be a distance between probability distributions. The ambiguity set $\mathcal{P}$ is defined as:

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}(\Xi) : d(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}, \tag{4}$$

where $\varepsilon > 0$ specifies the radius of the neighborhood of $\mathbb{Q}$ that we consider.

In particular, we first focus on ambiguity sets based on the Wasserstein distance as a mechanism to measure the distance between probability distributions. The following definition specifies the Wasserstein distance for this special case in which one of the distributions has the finite support of an empirical distribution.

Definition 1 (Wasserstein distance)**.** *Let $\mathbb{Q} \in \mathcal{M}(\Xi^n)$ be the empirical distribution on $\Xi^n = \{\xi^1, \ldots, \xi^n\}$, let $\mathbb{P} \in \mathcal{M}(\Xi)$, and let $\|\cdot\|$ be the L_p-norm. Then,*

$$\begin{aligned} d(\mathbb{Q}, \mathbb{P}) = \inf_{\mathbb{Z}} & \sum_{\sigma \in [n]} \int_\Xi \|\xi - \xi^\sigma\| \mathbb{Z}(\xi^\sigma, \xi) d\xi \\ \text{s.t.} & \int_\Xi \mathbb{Z}(\xi^\sigma, \xi) d\xi = \frac{1}{n} \quad \forall \sigma \in [n], \\ & \sum_{\sigma \in [n]} \mathbb{Z}(\xi^\sigma, \xi) = d\mathbb{P}(\xi) \quad \forall \xi \in \Xi, \\ & \mathbb{Z}(\xi^\sigma, \xi) \geq 0 \qquad \sigma \in [n], \xi \in \Xi. \end{aligned}$$

Embedding Definition 1 in the ambiguity set (4), we now restate the inner maximization problem for a given $x \in \mathcal{X}$:

$$\max_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_\mathbb{P}\left[V(x, \xi)\right] = \max_{\mathbb{P}, \mathbb{Z}} \int_\Xi V(x, \xi) d\mathbb{P}(\xi) \tag{5a}$$

$$\text{s.t.} \sum_{\sigma \in [n]} \int_\Xi \|\xi - \xi^\sigma\| \mathbb{Z}(\xi^\sigma, \xi) d\xi \leq \varepsilon, \tag{5b}$$

$$\int_\Xi \mathbb{Z}(\xi^\sigma, \xi) d\xi = \frac{1}{n} \qquad \forall \sigma \in [n], \tag{5c}$$

$$\sum_{\sigma \in [n]} \mathbb{Z}(\xi^\sigma, \xi) = d\mathbb{P}(\xi) \qquad \forall \xi \in \Xi, \tag{5d}$$

$$\mathbb{Z}(\xi^\sigma, \xi) \geq 0 \qquad \sigma \in [n], \xi \in \Xi. \tag{5e}$$

Following the developments in [13, 14] when Ξ is a compact set or in [6, 10, 11] when Ξ can be an unbounded set, we know that strong duality holds for problem (5). Let γ and ν^σ be the dual

variables corresponding to constraints (5b) and (5c), respectively, and replace $d\mathbb{P}(\xi)$ from (5d) in (5a). Then, the dual problem to (5) is:

$$\min_{\gamma \geq 0, \nu} \left[\varepsilon \gamma + \frac{1}{n} \sum_{\sigma \in [n]} \nu^\sigma \right] \tag{6a}$$

$$\text{s.t.} \quad \|\xi - \xi^\sigma\| \gamma + \nu^\sigma \geq V(x, \xi) \qquad \forall \sigma \in [n], \xi \in \Xi. \tag{6b}$$

As a result, model (1) can be reformulated as a single-level minimization problem:

$$\min_{x \in \mathcal{X}, \gamma \geq 0, \nu} \left[c^\top x + \varepsilon \gamma + \frac{1}{n} \sum_{\sigma \in [n]} \nu^\sigma \right] \tag{7a}$$

$$\text{s.t.} \quad \|\xi - \xi^\sigma\| \gamma + \nu^\sigma \geq V(x, \xi) \qquad \forall \sigma \in [n], \xi \in \Xi. \tag{7b}$$

We can also express model (1) in the following equivalent form:

$$\min_{x \in \mathcal{X}, \gamma \geq 0} \left[c^\top x + \varepsilon \gamma + \frac{1}{n} \sum_{\sigma \in [n]} \mathcal{V}(x, \gamma, \xi^\sigma) \right], \tag{8}$$

where for each $\sigma \in [n]$, $x \in \mathcal{X}$ and $\gamma \geq 0$

$$\mathcal{V}(x, \gamma, \xi^\sigma) = \max_{\xi \in \Xi} \{ V(x, \xi) - \gamma \|\xi - \xi^\sigma\| \}. \tag{9}$$

We make the following assumptions throughout the paper, and incorporate additional assumptions for special cases.

(A.1) The first stage feasible region, $\mathcal{X}$, is nonempty and compact.

(A.2) The dual feasible region, $\Pi = \{\pi : D^\top \pi \leq f\} \neq \emptyset$, is a polytope and satisfies $\Pi \neq \{0\}$.

(A.3) The empirical support, $\Xi^n = \{\xi^1, \ldots, \xi^n\}$, and candidate support, Ξ, satisfy $\Xi^n \subseteq \Xi$.

(A.4) The ambiguity set, $\mathcal{P}$, in equation (4) is specified using Definition 1 with $p \geq 1$ and $\varepsilon > 0$.

Problem (7) serves as a first step towards a tractable approach to solve problem (1). In Sections 3 and 4 we discuss different reformulations and algorithms to solve (1) depending on the structure of Ξ and the form of $d(\mathbb{Q}, \mathbb{P})$. In Section 3 we begin with the Wasserstein distance and with the case in which Ξ is unbounded. In particular, we assume Ξ is a convex cone, which includes the possibility that $\Xi = \mathbb{R}^m$. Then we discuss the case in which Ξ is a bounded hyper-rectangle. Section 4 revisits $\Xi = \mathbb{R}^m$ under an optimal transport distance, which replaces the L_p-norm in Definition 1 with a Mahalanobis-style distance.

3 Reformulations and algorithms

3.1 Ambiguity sets with unbounded support

In this section we consider ambiguity sets of probability distributions supported on unbounded sets. In particular, we assume that the support of the random parameters is a convex cone, e.g., $\Xi = \mathbb{R}^m$, $\Xi = \mathbb{R}^m_+$, or Ξ is a second-order cone. In what follows we show that the DRTSP reduces to the SAA problem. In other words, we establish the negative result that there is no reason to employ DRO in this manner. The two problems differ by a constant in the value of the objective function, which is, in general, hard to compute. Theorem 1 formalizes the result.

Theorem 1. *Assume (A.1)–(A.4), where (A.4) uses the L_p-norm, $\|\cdot\|$, for $p \geq 1$. Let Ξ be a convex cone; i.e., Ξ is convex and $\xi \in \Xi$ implies $\mu\xi \in \Xi$ for all $\mu \geq 0$. Then,*

$$\min_{x\in\mathcal{X}} \left\{ c^\top x + \max_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[V(x,\xi)\right] \right\} = K + \min_{x\in\mathcal{X}} \left\{ c^\top x + \frac{1}{n}\sum_{\sigma\in[n]} V(x,\xi^\sigma) \right\}, \tag{10}$$

where

$$\begin{aligned} K = \varepsilon \min_{z} \; & \|z\|_* \\ \text{s.t.} \; & z - \pi \in \Xi^* \equiv \{v \,:\, v^\top \xi \geq 0, \; \forall \xi \in \Xi\}, \; \forall \pi \in \text{ext}(\Pi) \end{aligned}$$

and where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proof. We start by rewriting the single-level dual reformulation (6) of nature's inner maximization problem:

$$\min_{\gamma\geq 0} \; \varepsilon\,\gamma + \frac{1}{n}\sum_{\sigma\in[n]} \max_{\xi\in\Xi} \{V(x,\xi) - \gamma\|\xi - \xi^\sigma\|\}.$$

Using the linear programming subproblem (2) and strong duality implied by (A.2), we can write $V(x,\xi)$ in terms of its dual:

$$\min_{\gamma\geq 0} \; \varepsilon\,\gamma + \frac{1}{n}\sum_{\sigma\in[n]} \max_{\xi\in\Xi} \left\{ \max_{\pi\in\Pi} \{\pi^\top(Bx+\xi)\} - \gamma\|\xi - \xi^\sigma\| \right\}. \tag{11}$$

Re-arranging the max operators in (11) and using the dual norm, $\|\cdot\|_*$, we obtain:

$$\min_{\gamma\geq 0} \; \varepsilon\,\gamma + \frac{1}{n}\sum_{\sigma\in[n]} \max_{\pi\in\Pi} \left\{ \pi^\top Bx + \max_{\xi\in\Xi} \min_{z:\|z\|_*\leq\gamma} \left(\pi^\top\xi - z^\top\xi + z^\top\xi^\sigma \right) \right\}. \tag{12}$$

Convexity of Ξ and compactness of the convex set $\{z : \|z\|_* \leq \gamma\}$ allows us to apply Sion's min-max theorem [16] to the inner-most problem and obtain:

$$\min_{\|z\|_*\leq\gamma} \left(z^\top\xi^\sigma + \max_{\xi\in\Xi} (\pi - z)^\top\xi \right). \tag{13}$$

Optimizing over the outer min problem in (12) requires γ to be large enough so that $z - \pi \in \Xi^* = \{v : v^\top \xi \geq 0 \; \forall \xi \in \Xi\}$ for all $\pi \in \Pi$ because otherwise we can select $\pi \in \Pi$, and a ray $\hat{\xi}$ of the cone Ξ with $(\pi - z)^\top \hat{\xi} > 0$ so that the value of (13), and hence that of (12) becomes ∞. Hence, we require $\gamma \geq \{\min_z \|z\|_* \; : \; z - \pi \in \Xi^*, \forall \, \pi \in \text{ ext}(\Pi)\}$, and for such a value of γ, we note that zero is a lower bound to the maximization problem in (13) since $0 \in \Xi$. For a given $\pi \in \Pi$, and under the required value of γ, the following problem is feasible and equivalent to (13):

$$\begin{array}{lll} \min\limits_{z} z^\top \xi^\sigma & = & \pi^\top \xi^\sigma \; + \; \min\limits_{v} v^\top \xi^\sigma \\ \text{s.t. } z - \pi \in \Xi^* & & \text{s.t. } v^\top \xi \geq 0 \; \forall \xi \in \Xi, \end{array}$$

where the equality follows with the change of variables $v = z - \pi$. Since $\xi^\sigma \in \Xi$ by assumption (A.3), $v^\top \xi^\sigma \geq 0$. Thus when γ is sufficiently large, (13) reduces to $\pi^\top \xi^\sigma$ and we obtain:

$$\max_{\pi \in \Pi} \left\{ \pi^\top Bx + \max_{\xi \in \Xi} \min_{z : \|z\|_* \leq \gamma} \left(\pi^\top \xi - z^\top \xi + z^\top \xi^\sigma \right) \right\} = \max_{\pi \in \Pi} \pi^\top (Bx + \xi^\sigma).$$

As a result, reformulation (11) can be written as:

$$\begin{array}{l} \min\limits_{\gamma \geq 0} \; \varepsilon \, \gamma + \dfrac{1}{n} \displaystyle\sum_{\sigma \in [n]} \max_{\pi \in \Pi} \pi^\top (Bx + \xi^\sigma) \\ \text{s.t. } \gamma \geq \{\min\limits_{z} \|z\|_* \; : \; z - \pi \in \Xi^*, \forall \, \pi \in \text{ ext}(\Pi)\}, \end{array}$$

which yields the desired result. □

When $\Xi = \mathbb{R}^m$ and $p = 1$ in Definition 1, Hanasusanto and Kuhn [11] provide a polynomially sized linear program for the DRTSP. Their result holds for a broader class of problems in which the right-hand-side of (2b) is of the form $Bx + T(x)\xi$, where $T(x)$ is affine in x, with appropriate dimensions. In our setting, $T(x) = I$, and their result applied to our problem provides a tractable formulation to compute K.

Theorem 2 (Special case of Theorem 6. Hanasusanto and Kuhn [11])**.** *Assume (A.1)–(A.4). Let $\Xi = \mathbb{R}^m$, let e_k be the kth unit vector, and assume Definition 1 uses the L_1 norm. Then problem* (1) *can be reformulated as the linear program:*

$$\begin{array}{l} \min\limits_{x,y,\gamma,\phi,\psi} \quad c^\top x + \varepsilon\gamma + \dfrac{1}{n} \displaystyle\sum_{\sigma \in [n]} f^\top y^\sigma \\ \quad \text{s.t. } x \in \mathcal{X}; \; \gamma \in \mathbb{R}_+; \; y^\sigma \in R_+^{d_y} \; \forall \sigma \in [n]; \; \phi^k \in \mathbb{R}^{d_y}, \psi^k \in \mathbb{R}^{d_y} \; \forall k \in [m], \\ \qquad Dy^\sigma = Bx + \xi^\sigma \quad \forall \sigma \in [n], \\ \qquad \gamma \geq f^\top \phi^k \quad \forall k \in [m], \\ \qquad \gamma \geq f^\top \psi^k \quad \forall k \in [m], \\ \qquad D\phi^k = e_k \quad \forall k \in [m], \\ \qquad D\psi^k = -e_k \quad \forall k \in [m]. \end{array} \tag{14}$$

Remark 1. *We assume $T = I$ in our second-stage problem* (2)*, but when $T(x)$ depends on x, the unit vectors e_k in problem* (14) *are multiplied by $T(x)$. In this case, the linear program* (14) *does not decouple as it does in Theorem 1.*

Remark 2. *When $p > 1$ in Definition 1, the maximum-norm problem required to compute K is NP-hard as shown in [11, Theorem 7].*

Theorem 1 shows that the formulation of a DRO problem with $V(x, \xi)$ as in (2), with unbounded conic support, and with the Wassersten distance is not helpful if the goal is to provide a first-stage solution that hedges against distributional uncertainty or is risk averse. In what follows we discuss alternative DRO formulations that are meaningful in this regard.

3.2 Ambiguity sets with box-type support

We turn to the case in which Ξ is defined as a hyper-rectangle in $\mathbb{R}^m$, i.e.,

$$\Xi = \{\xi \in \mathbb{R}^m : l_k \leq \xi_k \leq u_k \ \forall k \in [m]\}, \tag{15}$$

where all l_k's and u_k's are finite. Problem (7) suggests a delayed constraint generation algorithm in which constraints in (7b) are added iteratively; i.e., instead of considering infinitely many constraints corresponding to every $\xi \in \Xi$, we restrict consideration to a subset $\Xi' \subset \Xi$, which we enlarge as the algorithm proceeds. For a given $(x, \gamma) \in \mathcal{X} \times \mathbb{R}_+$ and $\sigma \in [n]$, the separation problem to enlarge Ξ' is:

$$\mathcal{V}(x, \gamma, \xi^\sigma) = \max_{\xi \in \Xi} \{V(x, \xi) - \gamma \|\xi - \xi^\sigma\|\}. \tag{16}$$

Problem (16) is non-convex because $V(x, \cdot)$ is a convex function on Ξ. More specifically, problem (16) can be recast as a bilinear program by expressing $V(x, \xi)$ using the dual (3). We can reformulate (16) as a mixed-integer program (MIP) under the assumption that Ξ is specified as in (15). The following theorem formalizes this result. Our formulation builds upon the development in [20, Proposition 5.2] for the unit commitment problem, but avoids several big-M type terms.

Theorem 3. *Let $x \in \mathcal{X}$ and $\gamma \geq 0$ be given, and consider a particular realization ξ^σ for some $\sigma \in [n]$. Assume (A.2), with known simple bounds $\underline{\pi}_k \leq \pi_k \leq \bar{\pi}_k$ for all $k \in [m]$, and further assume (A.3)–(A.4) with $p = 1$ in Definition 1. If Ξ is specified as in* (15) *then the separation problem* (16) *is equivalent to the following MIP:*

$$\begin{aligned}
\mathcal{V}(x, \gamma, \xi^\sigma) = \\
&\max_{\pi, b, w} \ \pi^\top B x + \sum_{k \in [m]} \left[u_k w_k^+ + (\xi_k^\sigma - u_k)\gamma b_k^+ + l_k w_k^- + (l_k - \xi_k^\sigma)\gamma b_k^- + \xi_k^\sigma w_k^0 \right] && \text{(17a)} \\
&\text{s.t. } \pi \in \Pi, b^+ \in \{0,1\}^m, b^- \in \{0,1\}^m, b^0 \in \{0,1\}^m, w^+ \in \mathbb{R}^m, w^- \in \mathbb{R}^m, w^0 \in \mathbb{R}^m && \text{(17b)} \\
&\quad b_k^+ + b_k^- + b_k^0 = 1 \ \forall k \in [m], && \text{(17c)}
\end{aligned}$$

$$w_k^- \leq -\gamma b_k^- \ \forall k \in [m], \tag{17d}$$
$$w_k^+ \geq \gamma b_k^+ \ \forall k \in [m], \tag{17e}$$
$$w_k^- \leq \bar{\pi}_k b_k^- \ \forall k \in [m], \tag{17f}$$
$$w_k^- \leq \pi_k + \underline{\pi}_k(b_k^- - 1) \ \forall k \in [m], \tag{17g}$$
$$w_k^- \geq \underline{\pi}_k b_k^- \ \forall k \in [m], \tag{17h}$$
$$w_k^- \geq \pi_k - \bar{\pi}_k(1 - b_k^-) \ \forall k \in [m], \tag{17i}$$
$$w_k^+ \leq \bar{\pi}_k b_k^+ \ \forall k \in [m], \tag{17j}$$
$$w_k^+ \leq \pi_k + \underline{\pi}_k(b_k^+ - 1) \ \forall k \in [m], \tag{17k}$$
$$w_k^+ \geq \underline{\pi}_k b_k^+ \ \forall k \in [m], \tag{17l}$$
$$w_k^+ \geq \pi_k - \bar{\pi}_k(1 - b_k^+) \ \forall k \in [m], \tag{17m}$$
$$w_k^0 \leq \bar{\pi}_k b_k^0 \ \forall k \in [m], \tag{17n}$$
$$w_k^0 \leq \pi_k + \underline{\pi}_k(b_k^0 - 1) \ \forall k \in [m], \tag{17o}$$
$$w_k^0 \geq \underline{\pi}_k b_k^0 \ \forall k \in [m], \tag{17p}$$
$$w_k^0 \geq \pi_k - \bar{\pi}_k(1 - b_k^0) \ \forall k \in [m]. \tag{17q}$$

Given (b^+, b^-, b^0) *in an optimal solution to model* (17)*, an optimal solution,* ξ*, to* (16) *is given by:*

$$\begin{aligned} \xi_k &= u_k \qquad \text{if } b_k^+ = 1, \\ \xi_k &= l_k \qquad \text{if } b_k^- = 1, \\ \xi_k &\in [l_k, u_k] \qquad \text{if } b_k^0 = 1. \end{aligned} \tag{18}$$

Proof. From equation (16) we can reformulate the problem so that the optimization over ξ is carried out analytically. In particular we have:

$$\begin{aligned} &\mathcal{V}(x, \gamma, \xi^\sigma) \\ &= \max_{\pi \in \Pi} \max_{\xi \in \Xi} \left\{ \pi^\top (Bx + \xi) - \gamma \|\xi - \xi^\sigma\| \right\} \\ &= \max_{\pi \in \Pi} \left\{ \pi^\top Bx + \max_{\xi \in \Xi} \min_{\|z\|_* \leq \gamma} \left(\pi^\top \xi - z^\top \xi + z^\top \xi^\sigma \right) \right\} \\ &= \max_{\pi \in \Pi} \left\{ \pi^\top Bx + \min_{\|z\|_* \leq \gamma} \max_{\xi \in \Xi} \left(\pi^\top \xi - z^\top \xi + z^\top \xi^\sigma \right) \right\}, \end{aligned}$$

where the first equality holds by (A.2) and duality, the second equality follows from the dual norm, and the third equality follows from the min-max theorem, analogous to its application in the proof of Theorem 1.

The inner-most maximization over the hyper-rectangle $\xi \in \Xi$ means:

$$\xi_k^* \in \begin{cases} \{u_k\} & \text{if } \pi_k - z_k > 0 \\ \{l_k\} & \text{if } \pi_k - z_k < 0 \\ [l_k, u_k] & \text{if } \pi_k - z_k = 0 \end{cases} \qquad \forall k \in [m].$$

Hence, we can re-write the separation problem as:

$$\mathcal{V}(x,\gamma,\xi^\sigma) = \max_{\pi\in\Pi}\left\{\pi^\top Bx + \min_{\|z\|_*\le\gamma}\sum_{k\in[m]}\left[u_k(\pi_k - z_k)^+ - l_k(z_k-\pi_k)^+ + z_k\xi_k^\sigma\right]\right\}.$$

By hypothesis $\|\cdot\|_* = \|\cdot\|_\infty$, and so the inner minimization over z separates component-wise with optimal values:

$$z_k^* = \begin{cases} -\gamma & \text{if } \pi_k \le -\gamma \\ \pi_k & \text{if } -\gamma \le \pi_k \le \gamma \\ \gamma & \text{if } \pi_k \ge \gamma \end{cases} \qquad \forall k \in [m].$$

Introducing binary variables b_k^+, b_k^-, and b_k^0 we can model each case of z_k with the following convention:

$$\begin{aligned} b_k^- = 1 &\implies \pi_k \le -\gamma, \\ b_k^0 = 1 &\implies -\gamma \le \pi_k \le \gamma, \\ b_k^+ = 1 &\implies \pi_k \ge \gamma, \end{aligned}$$

and the requirement that $b_k^+ + b_k^- + b_k^0 = 1$ for all $k \in [m]$. Hence, with $M > 0$ sufficiently large we can write:

$$\begin{aligned} \mathcal{V}(x,\gamma,\xi^\sigma) = \\ \max\ & \left\{\pi^\top Bx + \sum_{k\in[m]}\left[u_k\pi_k b_k^+ + (\xi_k^\sigma - u_k)\gamma b_k^+ + l_k\pi_k b_k^- + (l_k - \xi_k^\sigma)\gamma b_k^- + \xi_k^\sigma \pi_k b_k^0\right]\right\} && \text{(19a)} \\ \text{s.t. } & \pi \in \Pi, b^+ \in \{0,1\}^m, b^- \in \{0,1\}^m, b^0 \in \{0,1\}^m, && \text{(19b)} \\ & b_k^+ + b_k^- + b_k^0 = 1\ \forall k \in [m], && \text{(19c)} \\ & \pi_k \le -\gamma + M(1 - b_k^-)\ \forall k \in [m], && \text{(19d)} \\ & \pi_k \ge \gamma - M(1 - b_k^+)\ \forall k \in [m]. && \text{(19e)} \end{aligned}$$

Multiplying both sides of constraints (19d) and (19e) by b_k^- and b_k^+, respectively, allows us to replace the "big M" term and obtain:

$$\begin{aligned} \pi_k b_k^- &\le -\gamma b_k^-\ \forall k \in [m], \\ \pi_k b_k^+ &\ge \gamma b_k^+\ \forall k \in [m]. \end{aligned}$$

Introducing variables $w_k^- = \pi_k b_k^-$, $w_k^+ = \pi_k b_k^+$, and $w_k^0 = \pi_k b_k^0$, for all $k \in [m]$, allows us to replace the bilinear terms in these constraints and the objective function with the corresponding equivalent McCormick linearization. In particular, the McCormick envelope for $w_k^- = \pi_k b_k^-$ is given by:

$$w_k^- \le \bar{\pi}_k b_k^-\ \forall k \in [m],$$

$$
\begin{aligned}
&w_k^- \leq \pi_k + \underline{\pi}_k(b_k^- - 1) \ \forall k \in [m], \\
&w_k^- \geq \underline{\pi}_k b_k^- \ \forall k \in [m], \\
&w_k^- \geq \pi_k - \bar{\pi}_k(1 - b_k^-) \ \forall k \in [m],
\end{aligned}
$$

which mirrors (17f)-(17i). Applying the same argument for $w_k^+ = \pi_k b_k^+$ and $w_k^0 = \pi_k b_k^0$ yields, in turn, constraints (17j)-(17m) and (17n)-(17q) in model (17).

We can map from a solution to model (17) to that of model (19) and vice versa. Let (π, b^+, b^-, b^0) be an optimal solution to (19), and construct, for each $k \in [m]$, w_k^+, w_k^- and w_k^0 with the solution (π, b^+, b^-, b^0). Now, consider the vector $(\pi, b^+, b^-, b^0, w^+, w^-, w^0)$, constructed from the solution to (19). Constraints (17b) and (17c) hold trivially for the constructed solution, as well as constraints (17f)–(17q) from the definition of the McCormick linearization. If $b_k^- = 0$ for $k \in [m]$, constraint (17d) holds since $b_k^- = 0$ implies that $w_k^- = 0$ by construction. Similarly, if $b_k^- = 1$ for $k \in [m]$, constraint (17d) holds since $\pi_k \leq -\gamma$ from (19d) and $w_k^- = \pi_k$ by construction. A parallel argument holds for constraint (17e). The objective function values of models (17) and (19) are identical for the solutions just considered by construction of w_k^+, w_k^-, and w_k^0. Now consider an optimal solution to (17), and note that binary nature of (b^+, b^-, b^0) and constraints (17f)–(17q) directly imply $w_k^+ = \pi_k b_k^+$, $w_k^- = \pi_k b_k^-$, and $w_k^0 = \pi_k b_k^0$, for all $k \in [m]$, and therefore both (19d) and (19e) are satisfied since $b_k^+ + b_k^- + b_k^0 = 1$, for all $k \in [m]$ and M is sufficiently large. □

Although model (17) provides a mechanism to generate violated inequalities when only a subset of the constraints (7b) are included, we do not have access to the cost function, $V(x, \xi)$, for each $\xi \in \Xi$, in closed form. Nonetheless, we can leverage convexity of $V(\cdot, \xi)$ for every ξ, to write a suitable master problem for a cutting-plane algorithm. Let $\mathcal{J}$ index all the extreme points of Π, and consider a subset $\mathcal{J}^\omega \subseteq \mathcal{J}$ that has been generated as optimal solutions to $\max_{j \in \mathcal{J}}\{[\pi^j]^\top (B\hat{x} + \xi^\omega)\}$, where $\hat{x}$ is a first-stage solution at some iteration of the algorithm. Model (20) states the master problem of our cutting-plane algorithm. Algorithm 1 formalizes the procedure to solve model (1) upon reformulation as problem (7) under hyper-rectangular support, Ξ.

$$
\begin{aligned}
\min_{x \in \mathcal{X}, \gamma \geq 0, \nu} \quad & c^\top x + \varepsilon\gamma + \frac{1}{n}\sum_{\sigma \in [n]} \nu^\sigma && && \text{(20a)} \\
\text{s.t.} \quad & \nu^\sigma \geq \theta^\omega - \gamma\|\xi^\omega - \xi^\sigma\| && \forall \sigma \in [n], \xi^\omega \in \Xi' && \text{(20b)} \\
& \theta^\omega \geq \max_{j \in \mathcal{J}^\omega}\{[\pi^j]^\top (Bx + \xi^\omega)\} && \forall \omega : \xi^\omega \in \Xi' && \text{(20c)}
\end{aligned}
$$

Algorithm 1 Cutting-plane algorithm under hyper-rectangular Ξ

Require: Instance of model (1) with Ξ as in equation (15) under (A.1)–(A.4) and $p = 1$ in Definition 1
Ensure: $\hat{x}$ solves model (1)

1: Initialize $\Xi' \leftarrow \Xi^n$, `termination` $\leftarrow$ `False`, and for all $\xi^\omega \in \Xi'$ let $\mathcal{J}^\omega \leftarrow \emptyset$ and $\theta^\omega \geq -M$
2: **while** `termination` = `False` **do**
3: Solve (20) and obtain $(\hat{x}, \hat{\gamma}, \hat{\theta})$
4: **for** $\xi^\omega \in \Xi'$ **do** ▷ Solve LP subproblems
5: Solve $\max_{\pi \in \Pi} \pi^\top (B\hat{x} + \xi^\omega)$ and obtain $\hat{\pi}^\omega$
6: Append $\hat{\pi}^\omega$ to $\mathcal{J}^\omega$ if $\hat{\theta}^\omega < [\hat{\pi}^\omega]^\top (B\hat{x} + \xi^\omega)$
7: **end for**
8: **if** for any $\xi^\omega \in \Xi'$ a new $\hat{\pi}^\omega$ was added to $\mathcal{J}^\omega$ **then** go to step 3
9: **for** $\xi^\sigma \in \Xi^n$ **do** ▷ Solve MIP subproblems
10: Solve (17) for fixed $(\hat{x}, \hat{\gamma}, \xi^\sigma)$ and obtain $(\hat{\pi}^\omega, \xi^\omega)$ with ξ^ω given by (18)
11: **if** $\xi^\omega \notin \Xi'$ **then** ▷ Does ω index a new scenario?
12: $\Xi' \leftarrow \Xi' \cup \{\xi^\omega\}$
13: $\mathcal{J}^\omega \leftarrow \{\hat{\pi}^\omega\}$
14: Add $\theta^\omega \geq \max_{j \in \mathcal{J}^\omega} \{[\pi^j]^\top (Bx + \xi^\omega)\}$ to (20) ▷ New cut *and* new variable, θ^ω
15: Add constraints $\nu^\sigma \geq \theta^\omega - \gamma \|\xi^\omega - \xi^\sigma\| \quad \forall \sigma \in [n]$ to (20)
16: **end if**
17: **end for**
18: **if** no new support points were added **then** `termination` = `True`
19: **end while**

Algorithm 1 first repeats steps 3–8 until it solves model (7) under the relaxation that $\Xi = \Xi^n$, and this is accomplished by only solving LP subproblems in step 5. Only after this relaxation has been solved does the algorithm begin to generate $\xi^\omega \in \Xi \setminus \Xi^n$ in steps 9–17, which requires solving a MIP separation problem in step 10. After a new ξ^ω is generated, the algorithm returns to steps 3–8 to complete solution under this expanded set of scenarios. The algorithm terminates when no new scenarios and cuts can be generated, which is guaranteed to happen in a finite number of iterations because there are a finite number of extreme points in both Π and the polytope generated from the intersection of Ξ and the linearization of $\|\cdot\|_1$ in (16), i.e., $\{(\xi, w^+, w^-) \in \mathbb{R}^{3m} : \xi \in \Xi,\ w^+ - w^- = \xi - \xi^\sigma\ \forall \sigma \in [n]\}$.

Termination based on upper and lower bounds on the optimal value of model (1) is also possible. The optimal value of the master problem (20) at each iteration provides a lower bound. This holds by examining the reformulation in problem (7) and noting that our master problem: (i) contains some but not all of the scenarios $\xi \in \Xi$, and (ii) replaces $V(x, \xi)$ with an outer-linearization. Using reformulation (8)-(9) we have an upper bound via $c^\top \hat{x} + \varepsilon\hat{\gamma} + \frac{1}{n}\sum_{\sigma \in [n]} \mathcal{V}(\hat{x}, \hat{\gamma}, \xi^\sigma)$, where we evaluate $\mathcal{V}(\hat{x}, \hat{\gamma}, \xi^\sigma)$ in step 10 by solving the MIP (17) for each $\sigma \in [n]$. Thus we can terminate with a near-optimal solution once these upper and lower bounds are within a specified tolerance.

For more general compact support sets, Ξ, Luo and Mehrotra [13] establish that such a cutting plane—or cutting surface—algorithm obtains an ε-optimal solution in a finite number of iterations. Naturally, the practicality of any such termination criterion hinges on tractability of the separation problem (16).

4 Optimal transport distance in two-stage stochastic programs

The optimal transport distance extends the Wasserstein distance by considering a general cost function, c, instead of the L_p-norm used in Definition 1, and this is formalized in Definition 2.

Definition 2 (Optimal transport distance)**.** *Let $\mathbb{Q} \in \mathcal{M}(\Xi^n)$ be the empirical distribution on $\Xi^n = \{\xi^1, \ldots, \xi^n\}$, let $\mathbb{P} \in \mathcal{M}(\Xi)$, and let $c : \Xi \times \Xi \to \mathbb{R}_+$. Then,*

$$\begin{aligned} d(\mathbb{Q}, \mathbb{P}) = \inf_{\mathbb{Z}} & \sum_{\sigma \in [n]} \int_{\Xi} c(\xi, \xi^\sigma) \mathbb{Z}(\xi^\sigma, \xi) d\xi \\ \text{s.t.} & \int_{\Xi} \mathbb{Z}(\xi^\sigma, \xi) d\xi = \frac{1}{n} \; \forall \sigma \in [n], \\ & \sum_{\sigma \in [n]} \mathbb{Z}(\xi^\sigma, \xi) = d\mathbb{P}(\xi) \; \forall \xi \in \Xi, \\ & \mathbb{Z}(\xi^\sigma, \xi) \geq 0 \qquad \sigma \in [n], \xi \in \Xi. \end{aligned}$$

We are interested in cost functions of the form,

$$c(\xi, \xi^\sigma) = \frac{1}{2}(\xi - \xi^\sigma)^\top C (\xi - \xi^\sigma) \tag{21}$$

where C is a symmetric positive definite matrix. This form of optimal transport has been adopted, e.g., in Blanchet et al. [7]. Within a factor of two, this is known as the square of the Mahalanobis distance measure when C is the inverse of an estimate of ξ's covariance matrix and serves to normalize with respect to standard deviations and account for correlations. Our interest in this form of cost function is further motivated by the fact that when we consider an unbounded support for the random parameters, e.g., $\Xi = \mathbb{R}^m$, in contrast to Theorem 1, it is no longer the case that solving the DRO problem is equivalent to the underlying SAA problem. Rather, the DRO problem can hedge against distributional uncertainty.

Theorem 4. *Consider problem* (1) *with $V(x, \xi)$ defined by equation* (2)*, and assume (A.1)–(A.3). Let the ambiguity set* (4) *be defined with the transport distance of Definition 2 with $c(\xi, \xi^\sigma)$ as in equation* (21)*, where C is symmetric positive definite, and let $\Xi = \mathbb{R}^m$. Then problem* (1) *is equivalent to problem* (8) *with:*

$$\mathcal{V}(x, \gamma, \xi^\sigma) = \max_{\xi \in \Xi} \{ V(x, \xi) - \gamma c(\xi, \xi^\sigma) \} \tag{22a}$$

$$= \max_{\pi \in \Pi} \left\{ \pi^\top (Bx + \xi^\sigma) + \frac{1}{2\gamma} \pi^\top C^{-1} \pi \right\}, \tag{22b}$$

where $\Pi = \{\pi : D^\top \pi \leq f\}$ and $\gamma > 0$. Given an optimal solution π to model (22b)*, an optimal solution, ξ, to* (22a) *is given by:*

$$\xi = \xi^\sigma + \frac{1}{\gamma} C^{-1} \pi. \tag{23}$$

Proof. Following the same line argument given in Section 2, we first arrive at problem (6) with $\|\xi - \xi^\sigma\|$ replaced by $c(\xi, \xi^\sigma)$ and then to problem (8) with $\mathcal{V}(x, \gamma, \xi^\sigma)$ given by (22a). Then,

$$\begin{aligned}\mathcal{V}(x, \gamma, \xi^\sigma) &= \max_{\xi \in \Xi} \{V(x, \xi) - \gamma c(\xi, \xi^\sigma)\} \\ &= \max_{\xi \in \mathbb{R}^m} \left\{ \max_{\pi \in \Pi} \pi^\top (Bx + \xi) - \frac{\gamma}{2}(\xi - \xi^\sigma)^\top C (\xi - \xi^\sigma) \right\} \\ &= \max_{\pi \in \Pi} \left\{ \pi^\top Bx + \max_{\xi \in \mathbb{R}^m} \left[\pi^\top \xi - \frac{\gamma}{2}(\xi - \xi^\sigma)^\top C (\xi - \xi^\sigma) \right] \right\}.\end{aligned}$$

Given that $\Pi \neq \{0\}$ from (A.2) and that $\Xi = \mathbb{R}^m$, we must have $\gamma > 0$ because otherwise $\mathcal{V}(x, \gamma, \xi^\sigma) = \infty$. Optimizing over ξ for a given π yields $\xi = \xi^\sigma + \frac{1}{\gamma} C^{-1} \pi$, and substitution yields:

$$\mathcal{V}(x, \gamma, \xi^\sigma) = \max_{\pi \in \Pi} \left\{ \pi^\top (Bx + \xi^\sigma) + \frac{1}{2\gamma} \pi^\top C^{-1} \pi \right\}.$$

□

Remark 3. *Suppose the right-hand side of* (2b) *is replaced by* $Bx + T(x)\xi$*, where* $T(x)$ *is an* $m \times d_\xi$ *dimensional matrix whose components are affine in* x*, and* ξ *is a* d_ξ *dimensional random vector. Then the result in Theorem 4 extends to:*

$$\mathcal{V}(x, \gamma, \xi^\sigma) = \max_{\pi \in \Pi} \left\{ \pi^\top (Bx + T(x)\xi^\sigma) + \frac{1}{2\gamma} \pi^\top T(x) C^{-1} T(x)^\top \pi \right\}, \tag{25}$$

and given an optimal solution π *to model* (25)*, an optimal solution,* ξ*, to* (22a) *is given by:*

$$\xi = \xi^\sigma + \frac{1}{\gamma} C^{-1} T(x)^\top \pi. \tag{26}$$

Under the hypotheses of Theorem 4 we can reformulate model (1) as:

$$\min_{x \in \mathcal{X}, \gamma \geq 0} \left[c^\top x + \varepsilon \gamma + \frac{1}{n} \sum_{\sigma \in [n]} \max_{\pi \in \Pi} \left\{ \pi^\top (Bx + \xi^\sigma) + \frac{1}{2\gamma} \pi^\top C^{-1} \pi \right\} \right]. \tag{27}$$

Because the inner maximization in model (27) is non-convex, we cannot employ quadratic programming duality to reformulate the model in an equivalent extensive form. That said, the objective function of (27) is convex in (x, γ). Equivalently, $\mathcal{V}(x, \gamma, \xi^\sigma)$, as given in equation (22), is a convex function in (x, γ) on the convex hull of $\mathcal{X} \times \mathbb{R}_+$. Thus a cutting-plane method is again possible, even if it requires solving the non-convex quadratic separation problem given in problem (27). Due to the $\frac{1}{\gamma}$ term, it is computationally useful to bound $\gamma \geq 0$ away from zero prior to applying the cutting-plane procedure. The following result helps in this regard.

Theorem 5. *Let the hypotheses of Theorem 4 hold, let* (x^*, γ^*) *be an optimal solution to problem* (8)*, and let* $\|\cdot\|$ *denotes the two-norm. Then, there exists* $\underline{\gamma}$ *such that*

$$\gamma^* \geq \underline{\gamma} \geq \frac{1}{\sqrt{2n\varepsilon \lambda_{\max}(C)}} \|\pi^*\| > 0,$$

where $\pi^* \in \arg\min_{\pi \in ext(\Pi)\setminus\{0\}} \|\pi\|$. *Moreover, if* $\pi^\sigma \neq 0$ *for all* $\sigma \in [n]$ *denote optimal solutions to the inner maximization of problem* (27) *under* (x^*, γ^*) *then*

$$\gamma^* \geq \underline{\gamma} \geq \frac{1}{\sqrt{2\varepsilon\lambda_{\max}(C)}}\|\pi^*\| > 0, \tag{28}$$

where $\|\pi^*\| = \min_{\sigma\in[n]} \|\pi^\sigma\|$.

Proof. Let $f(x,\gamma) = c^\top x + \varepsilon\gamma + \frac{1}{n}\sum_{\sigma\in[n]} \mathcal{V}(x,\gamma,\xi^\sigma)$. By Theorem 4 we have:

$$f(x,\gamma) = c^\top x + \varepsilon\gamma + \frac{1}{n}\sum_{\sigma\in[n]} \max_{\pi\in\Pi}\left\{\pi^\top(Bx+\xi^\sigma) + \frac{1}{2\gamma}\pi^\top C^{-1}\,\pi\right\}. \tag{29}$$

For any $x \in \mathcal{X}$ we have $\lim_{\gamma\to 0^+} f(x,\gamma) = \infty$ because $\Pi \neq \{0\}$ and C^{-1} is positive definite. This shows existence of $\underline{\gamma} > 0$ with $\gamma^* \geq \underline{\gamma}$. We know that for each $\sigma \in [n]$ the inner maximization is achieved at an extreme point of Π, which we denote π^σ. Given these maximizers, γ^* minimizes

$$\varepsilon\gamma + \frac{1}{\gamma}\frac{1}{2n}\sum_{\sigma\in[n]} [\pi^\sigma]^\top C^{-1}\,\pi^\sigma.$$

Thus

$$\gamma^* = \left(\frac{1}{2n\varepsilon}\sum_{\sigma\in[n]} [\pi^\sigma]^\top C^{-1}\,\pi^\sigma\right)^{1/2}.$$

We know $\gamma^* > 0$, and so for at least one $\sigma \in [n]$, $\pi^\sigma \neq 0$. Using the inequality $[\pi^\sigma]^\top C^{-1}\,\pi^\sigma \geq \frac{1}{\lambda_{\max}(C)}\|\pi^\sigma\|^2$, the results follow. □

Algorithm 2 describes a cutting-plane algorithm, which makes use of Theorems 4 and 5, and master problem:

$$\min_{x\in\mathcal{X},\gamma\geq\underline{\gamma},\nu} \; c^\top x + \varepsilon\gamma + \frac{1}{n}\sum_{\sigma\in[n]}\nu^\sigma \tag{30a}$$

$$\text{s.t.} \quad \nu^\sigma \geq \theta^\omega - \gamma\, c(\xi^\omega,\xi^\sigma) \qquad \forall\sigma\in[n], \xi^\omega \in \Xi' \tag{30b}$$

$$\theta^\omega \geq \max_{j\in\mathcal{J}^\omega}\{[\pi^j]^\top(Bx+\xi^\omega)\} \quad \forall\omega : \xi^\omega \in \Xi', \tag{30c}$$

where $c(\xi^\omega,\xi^\sigma)$ is defined via equation (21). The cutting-plane algorithm includes a search for $\underline{\gamma}$. We use Theorem 5 to guide our initial value of $\underline{\gamma} = 1/\sqrt{2\varepsilon\lambda_{\max}(C)}$. This assumes $\pi^\sigma \neq 0$ for all $\sigma \in [n]$ and ignores $\|\pi^*\|$ in inequality (28), but is perhaps reasonable if the second-stage subproblem is well scaled. If $\gamma^* > \underline{\gamma}$ at optimality, then the choice of $\underline{\gamma}$ is valid. Otherwise, we reduce $\underline{\gamma}$ in the fashion of a backtracking line search.

Algorithm 2 Cutting-plane algorithm under $\Xi = \mathbb{R}^m$

Require: Instance of model (1) with $\Xi = \mathbb{R}^m$ under (A.1)–(A.3) and where (4) uses Definition 2 with (21).
Ensure: $\hat{x}$ solves model (1)

1: Initialize $\Xi' \leftarrow \Xi^n$, $\underline{\gamma} \leftarrow \frac{1}{\sqrt{2\varepsilon\lambda_{\max}(C)}}$, `termination` $\leftarrow$ `False`, and for all $\xi^\omega \in \Xi'$ let $\mathcal{J}^\omega \leftarrow \emptyset$ and $\theta^\omega \geq -M$
2: **while** `termination` = `False` **do**
3: Solve (30) and obtain $(\hat{x}, \hat{\gamma}, \hat{\theta})$
4: **for** $\xi^\omega \in \Xi'$ **do** ▷ Solve LP subproblems
5: Solve $\max_{\pi \in \Pi} \pi^\top (B\hat{x} + \xi^\omega)$ and obtain $\hat{\pi}^\omega$
6: Append $\hat{\pi}^\omega$ to $\mathcal{J}^\omega$ if $\hat{\theta}^\omega < [\hat{\pi}^\omega]^\top (B\hat{x} + \xi^\omega)$
7: **end for**
8: **if** for any $\xi^\omega \in \Xi'$ a new $\hat{\pi}^\omega$ was added to $\mathcal{J}^\omega$ **then** go to step 3
9: **for** $\xi^\sigma \in \Xi^n$ **do** ▷ Solve non-convex subproblems
10: Solve (22b) for fixed $(\hat{x}, \hat{\gamma}, \xi^\sigma)$ and obtain $(\hat{\pi}^\omega, \xi^\omega)$ with $\xi^\omega = \xi^\sigma + \frac{1}{\hat{\gamma}} C^{-1} \hat{\pi}^\omega$ via (23)
11: **if** $\xi^\omega \notin \Xi'$ **then**
12: $\Xi' \leftarrow \Xi' \cup \{\xi^\omega\}$
13: $\mathcal{J}^\omega \leftarrow \{\hat{\pi}^\omega\}$
14: Add $\theta^\omega \geq \max_{j \in \mathcal{J}^\omega} \{[\pi^j]^\top (Bx + \xi^\omega)\}$ to (30) ▷ New cut *and* new variable, θ^ω
15: Add constraints $\nu^\sigma \geq \theta^\omega - \gamma \|\xi^\omega - \xi^\sigma\| \quad \forall \sigma \in [n]$ to (30)
16: **end if**
17: **end for**
18: **if** no new support points were added and $\hat{\gamma} > \underline{\gamma}$ **then** `termination` = `True`
19: **if** $\hat{\gamma} = \underline{\gamma}$ **then**
20: Let $\underline{\gamma} \leftarrow \frac{1}{2}\underline{\gamma}$ ▷ Backtracking line search on $\underline{\gamma}$
21: **end if**
22: **end while**

Algorithm 2 follows the same basic steps as in Algorithm 1, with a few modifications. Step 3 now solves a master problem with $c(\cdot, \cdot)$ in place of the L_p-norm and with $\gamma \geq \underline{\gamma}$ in place of $\gamma \geq 0$. Step 10 solves the separation problem and computes a potentially new support point in closed form. Finally, step 20 updates $\underline{\gamma}$ whenever the constraint $\gamma \geq \underline{\gamma}$ is binding. Termination of the algorithm requires $\hat{\gamma} > \underline{\gamma}$ in addition to the conditions described in Algorithm 1, i.e., γ needs to be constrained only by (20b) for our procedure to be correct. Parallel to the discussion following Algorithm 1, we can terminate Algorithm 2 based on upper and lower bounds on problem (1)'s optimal value. The upper bound is computed in an analogous manner to that discussed at the end of Section 3.2. We obtain a valid lower bound from the optimal value of the master problem if $\hat{\gamma} > \underline{\gamma}$, or if we resolve the master (30) with $\underline{\gamma}$ replaced by 0.

5 Numerical experiments

We conduct numerical experiments to illustrate the relative merits of our DRO approaches in a relatively simple setting. In particular, we consider a two-stage stochastic program in which the first-stage decision allocates supplies of a single commodity to facilities. Given each facility's allocation, in the second stage we satisfy demand, which has now been realized, at minimum cost, where the cost varies for each facility-demand pair. If supply is inadequate we incur a large penalty

cost for every unit of subcontracted demand. Excess supply at a facility incurs a holding cost. Table 1 summarizes the model's notation.

Sets	
$g \in \mathcal{G}$	set of facilities
$d \in \mathcal{D}$	set of demand sites
Parameters	
$\ell_g > 0$	upper limit on supply installed at facility g
$c_{gd} > 0$	unit cost of satisfying demand at site d from facility g
$\rho > 0$	unit cost for subcontracted demand
$h > 0$	unit cost for holding inventory
ξ_d	random demand at site d
Decision variables	
x_g	supply allocated to facility g
y_{gd}	amount supplied by facility g to site d
u_d	shortfall subcontracted at site d
v_g	excess supply held at facility g

Table 1: Notation for supply-allocation problem.

Model (31) formulates the DRO problem with the recourse function defined in (32):

$$\min_{0 \le x \le \ell} \max_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[V(x, \xi) \right], \tag{31}$$

where

$$V(x,\xi) = \min_{y,u,v} \sum_{g \in \mathcal{G}} \sum_{d \in \mathcal{D}} c_{gd} y_{gd} + \sum_{d \in \mathcal{D}} \rho u_d + \sum_{g \in \mathcal{G}} h v_g \tag{32a}$$

$$\text{s.t.} \sum_{d \in \mathcal{D}} y_{gd} + v_g = x_g \ \forall g \in \mathcal{G}, \tag{32b}$$

$$\sum_{g \in G} y_{gd} + u_d \ge \xi_d \ \forall d \in \mathcal{D}, \tag{32c}$$

$$y_{gd}, u_d, v_g \ge 0 \ \forall g \in \mathcal{G}, d \in \mathcal{D}. \tag{32d}$$

We provide the dual of model (32) as it forms the basis for the separation problems that we have proposed. Let α_g be the dual variable associated with constraints (32b) and let β_d be the dual variable associated with constraints (32c). Then, we can represent the value function as:

$$\begin{aligned} V(x,\xi) = \max_{\alpha,\beta} & \sum_{g \in \mathcal{G}} x_g \alpha_g + \sum_{d \in \mathcal{D}} \xi_d \beta_d \\ \text{s.t. } & \alpha_g + \beta_d \le c_{gd} \ \forall g \in \mathcal{G}, \ d \in \mathcal{D}, \\ & \alpha_g \le h \ \forall g \in \mathcal{G}, \\ & 0 \le \beta_d \le \rho \ \forall d \in \mathcal{D}. \end{aligned}$$

Both α and β are naturally bounded from above. Unlike β, variable α is not explicitly bounded from below in the dual. However, given $0 < c_{gd} < \rho$ for all $g \in \mathcal{G}, d \in \mathcal{D}$, we can append $\alpha_g \ge -\rho$

for all $g \in \mathcal{G}$ while maintaining that $V(x, \xi)$ takes the optimal value of the dual, and hence (A.2) is satisfied.

In our computation we use $|\mathcal{G}| \in \{5, 10, 20\}$, $|\mathcal{D}| \in \{20, 30, 50\}$, $\rho = 10$, and $h = 1$. The unit cost of satisfying demand, c_{gd}, is the Euclidean distance between facility g and demand site d, and locations of both facilities and demand sites are randomly generated on the unit square. The limit on installed capacity, ℓ, is sufficiently large so that it is nonbinding. For the random demand, ξ_d, we consider two separate cases: a lognormal distribution with parameters $\mu = 1$ and $\sigma = 1$, and a uniform distribution on $[0, \xi_{\max}]$. In both cases, we assume demands are independent and identically distributed across sites. That said, we assume that we only have access to n realizations of the vector $\xi = (\xi_d)_{d \in \mathcal{D}}$ when computing the first-stage decision x. After generating lognormal realizations, we specify $\xi_{\max} = \max_{d \in \mathcal{D}} \max_{\sigma \in [n]} \xi_d^\sigma$ for the uniformly distributed case. All randomly generate demands are rounded to one decimal.

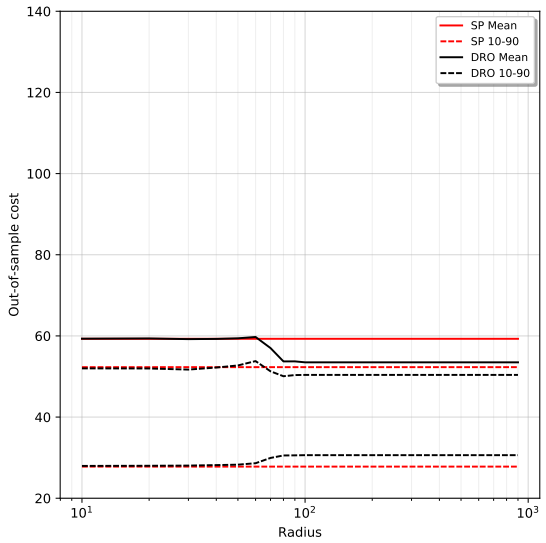
We test the solution obtained with limited data in an out-of-sample simulation that draws additional realizations of the random parameters from the true distribution, either lognormal or uniform. In what follows, we first compare the out-of-sample performance of different DRO formulations in the case of both the lognormal, i.e., a skewed distribution, and the uniform, i.e., a symmetric distribution. Next, we analyze the relative merit of using Algorithms 1 and 2 versus a standard cutting-plane algorithm and discuss implementation details. We close with a geometric analysis of how the worst-case probability distribution changes, both in terms of support points and probability mass, as the radius, ε, increases.

5.1 Out-of-sample performance

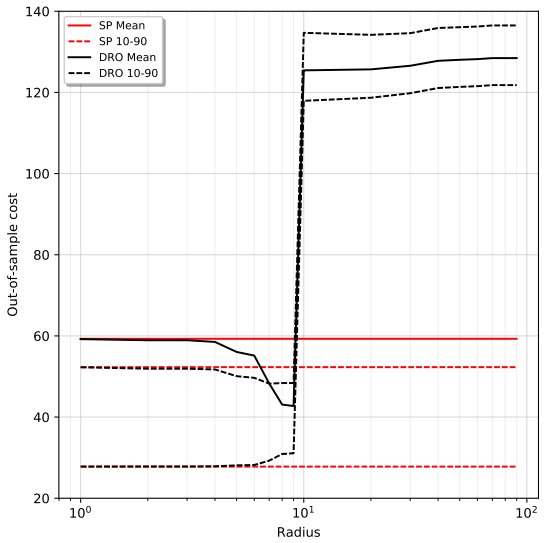
We evaluate three DRO models in terms of their out-of-sample performance. The first assumes $\Xi = \Xi^n$ and uses Definition 1 to define the ambiguity set under the L_p-norm with $p = 1$. In this case, we restrict attention to the empirical support, Ξ^n, and so the subproblems are LPs since ξ is not a decision variable in the separation problem. Second, we consider a DRO model in which the support, Ξ, is a hyper-rectangle per equation (15), where we let $l_k = 0$ and $u_k = \max_{d \in \mathcal{D}} \max_{\sigma \in [n]} \xi_d^\sigma$ for all $k \in [m]$. Here, we again use Definition 1 with $p = 1$ to construct the ambiguity set. To solve this DRO formulation we use Algorithm 1. In our third DRO model, we assume $\Xi = \mathbb{R}^m$ and use Definition 2 with equation (21) and $C = I$ to define the ambiguity set, and we solve the problem via Algorithm 2.

We use $n = 10$ scenarios to center the ambiguity set, and share the same 10 scenarios across all three DRO models. Figure 1 shows the results using lognormal random variables, with Figures 1a, 1b, and 1c respectively corresponding to empirical support, rectangular support, and unrestricted support with quadratic transport cost. We solve each DRO model for a range of ε. We then simulate the cost of the optimal first-stage solution over 1,000 realizations of the random parameters, using common random numbers across the three formulations. We show in black the results of the expected cost as well as the 10th percentile and 90th percentile of the cost over the 1,000 realizations.

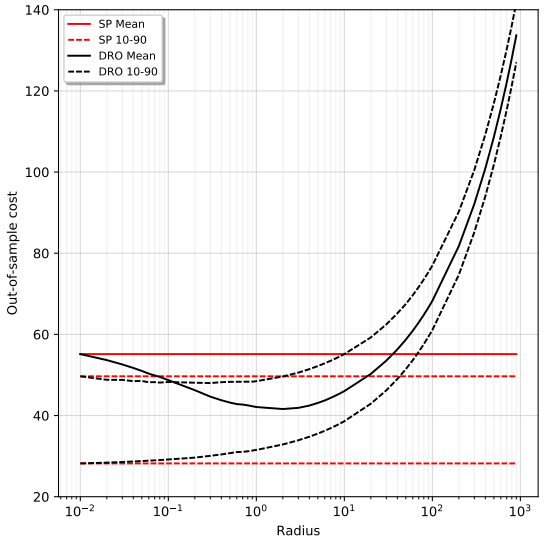
As a benchmark, we plot in red the same metrics for the solution obtained from solving the nominal problem, i.e., the SAA problem or equivalently, the DRO problem with $\varepsilon = 0$.



(a) $\Xi = \Xi^n$, Wasserstein distance, $p = 1$.



(b) Ξ defined by (15), Wasserstein distance, $p = 1$.



(c) $\Xi = \mathbb{R}^m$, quadratic optimal transport distance.

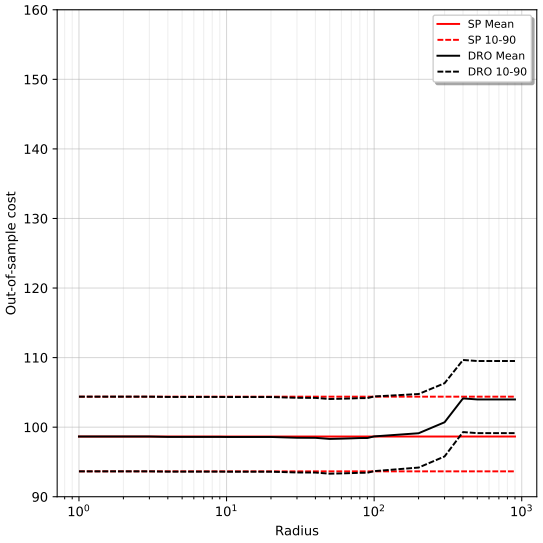
Figure 1: Out-of-sample assessment of policies obtained for three different DRO formulations. Here, we use $|\mathcal{G}| = 10$ and $|\mathcal{D}| = 30$, a nominal distribution with $n = 10$ scenarios, and 1,000 out-of-sample scenarios, generated from a lognormal distribution.

Due to the lognormal demand and the cost structure, the solution, $\hat{x}$, to the nominal problem has such a strongly skewed distribution of costs, $V(\hat{x}, \xi)$, that the mean exceeds the 90th percentile. In this type of setting, DRO can improve out-of-sample performance as we now discuss. A common result across the three formulations is that there is a range of values of the radius, ε, for which the expected out-of-sample cost is better for the DRO solution than that of the nominal problem. A similar observation holds for the 90th percentile of the cost, suggesting that our goal to find a first-

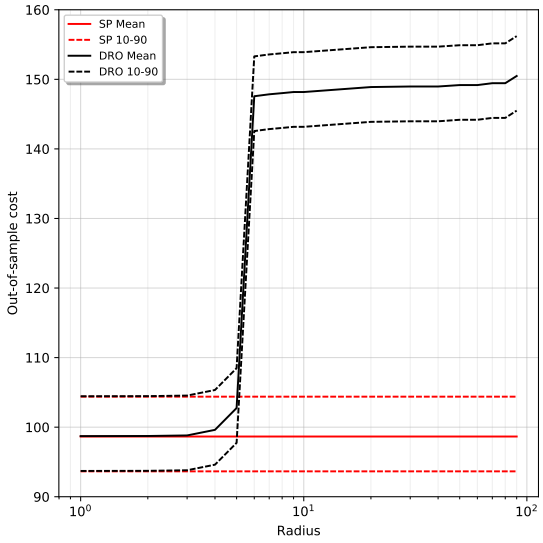
stage solution, which hedges against costly realizations is achieved to some extent for appropriate values of ε.

The range of interesting values of ε differs, along with the ambiguity set for each model. In Figure 1a, when the radius exceeds 90 we obtain the same first-stage decision, which outperforms the nominal solution in mean. In Figure 1b, the solution obtained with $\varepsilon = 8$ achieves a significant improvement in out-of-sample mean. The case of the quadratic transport cost in Figure 1c again offers out-of-sample improvements, with solutions for $\varepsilon \in (1, 4)$ significantly reducing the out-of-sample mean. For a wide range of values of ε, the results of Figures 1b and 1c show that the distribution of second-stage costs is reshaped so that the mean falls below the 90th percentile. The results of Figure 1a differ from those of Figures 1b and 1c in that new support points are generated in the latter two, which help hedge against realizations not contained in Ξ^n. However, as ε grows large the additional support points yield overly conservative and costly solutions.

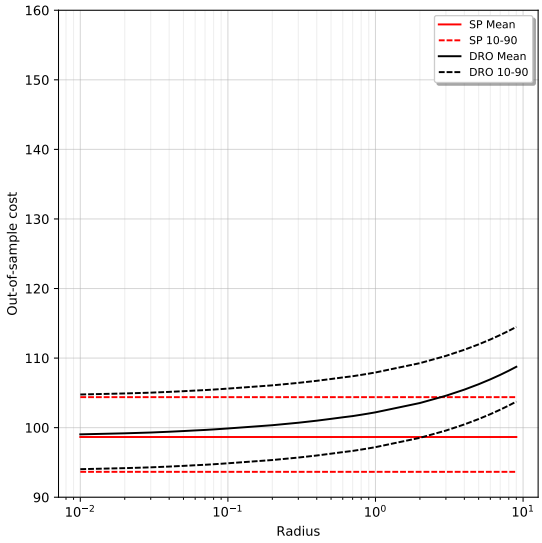
Figure 2 mirrors Figure 1, but in this cases the random demand is assumed to be uniformly distributed. In contrast to the lognormal case, there is little to no benefit in the out-of-sample mean or 90th percentile of the cost. Instead, as the radius increases, the cost deteriorates relative to that of the nominal solution. In this case, the distribution of the costs, $V(\hat{x}, \xi)$, is not as skewed as in the lognormal case and both the mean and median are centered between the 10th and 90th percentiles. The qualitatively different results associated with skewed (Figure 1) versus symmetric (Figure 2) distributions that we observe are consistent with the analysis of Anderson and Philpott [1].



(a) $\Xi = \Xi^n$, Wasserstein distance, $p = 1$.



(b) Ξ defined by (15), Wasserstein distance, $p = 1$.



(c) $\Xi = \mathbb{R}^m$, quadratic optimal transport distance.

Figure 2: Out-of-sample assessment of policies obtained for three different DRO formulations. Here, we use $|\mathcal{G}| = 10$ and $|\mathcal{D}| = 30$, a nominal distribution with $n = 10$ scenarios, and 1,000 out-of-sample scenarios, generated from a uniform distribution.

5.2 Algorithmic considerations and implementation details

Algorithm 1 is an enhanced version of a standard cutting-plane procedure for solving problem (1) via reformulation (7) under the assumption that ξ's support is a hyper-rectangle as specified by equation (15). A standard version of the algorithm would solve subproblem (16), or rather its MIP reformulation (17), for every $\xi^\sigma \in \Xi^n$ and add cuts accordingly. Instead, Algorithm 1 first solves problem (7) with Ξ replaced by $\Xi' = \Xi^n$, which it can accomplish by solving LP subproblems (3). After enlarging Ξ' with new scenarios by solving the MIP (17) for each $\sigma \in [n]$, the algorithm again reverts to solving solving LP subproblems to solve problem (7) with the enlarged Ξ'.

The simple enhancement just sketched can help in two important ways. First, we reduce the number of computationally expensive MIP subproblems (17) that we must solve by solving LPs instead. Second, to achieve convergence we can require fewer iterations, solving significantly fewer total subproblems of type (3) and (17). To understand the latter issue, we observe that like other cutting-plane algorithms, in early iterations the first-stage solutions $(\hat{x}, \hat{\gamma})$ tend to oscillate over $\mathcal{X} \times \mathbb{R}_+$, visiting solutions far from the neighborhood of an optimal solution. Each $(\hat{x}, \hat{\gamma})$ and ξ^σ, $\sigma \in [n]$, gives rise to a potentially new $\xi \in \Xi \setminus \Xi^n$. Each such new ξ means that we need to solve additional LPs at subsequent iterations. Algorithm 1 helps limit such behavior by repeatedly optimizing problem (7) over the existing Ξ' prior to expanding the set, and hence guides master problem solutions, $(\hat{x}, \hat{\gamma})$, to promising parts of the feasible region. Of course, we could remove elements from Ξ', and they would be regenerated if need be. However, regeneration is computationally expensive, and so we do not pursue this. When $\mathcal{X}$ is convex, level-set methods or regularization could also help, but we show results without these enhancements.

Table 2 shows results for various instances of model (31) using the ambiguity set defined with the Wasserstein distance, the one-norm, and hyper-rectangular support (15). In particular, we fix $n = 10$ and $\varepsilon = 8$ based on Figure 1b, and we vary the number of facilities and demand sites ($\varepsilon = 8$ is also reasonable for these instances). Columns 1 and 2 describe the instance in terms of $|\mathcal{G}|$ and $|\mathcal{D}|$. Columns 3–6 show the number of iterations of the cutting-plane algorithm; the number of LP subproblems (3) solved; the number of MIP subproblems (17) solved; and the total computational time, respectively. Columns 7–10 present the same information but for Algorithm 1.

Facilities	Demand Sites	Standard Cutting Plane				Algorithm 1			
		Iterations	LP subs	MIP subs	Time (s)	Iterations	LP subs	MIP subs	Time (s)
5	20	48	0	480	5.8	27	337	50	1.6
10	20	102	0	1020	18.1	39	534	50	2.7
20	20	115	0	1150	23.8	42	633	50	2.8
5	30	48	0	480	9.8	29	304	50	1.4
10	30	116	0	1160	27.7	44	450	20	1.5
20	30	191	0	1910	57.2	67	816	50	4.6
5	50	52	0	520	24.2	30	438	60	5.1
10	50	337	0	3370	127.9	67	883	60	13.8
20	50	5027	0	50270	1707.8	156	1988	60	21.9

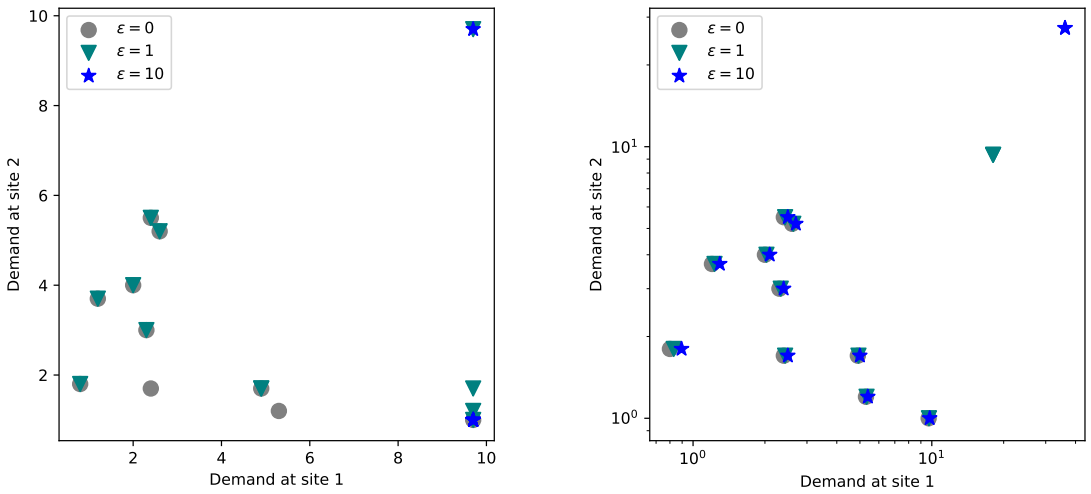
Table 2: Profiling two variants of the cutting-plane method from Section 3.2.

We omit a detailed comparison between Algorithm 2 and its standard cutting-plane analog because their relative performance is more extreme in favor of Algorithm 2 than shown in Table 2. As with Algorithm 1, the exact separation problem in Algorithm 2 is computationally expensive to solve, and in general more expensive than the MIP used in Algorithm 1. The oscillating behavior described above more strongly degrades the performance of Algorithm 2 because even fractional changes in the first-stage solutions, $(\hat{x}, \hat{\gamma})$, tend to generate a new support point via (23). We use a relative tolerance of 10^{-6} to terminate the algorithm, and it takes about 300 seconds to solve instances of size $|\mathcal{G}| = 10$ and $|\mathcal{D}| = 30$, i.e., two orders of magnitude larger than for the

corresponding problem in Table 2.

5.3 Worst-case distribution

We close the discussion of numerical experiments with a geometric view of the worst-case probability distribution of our DRO formulations. Upon termination of the algorithm, we can retrieve the worst-case probability distribution from the dual variables of the master problems (20) and (30). The dual variable associated with each constraint in (20b) and (30b) corresponds to the amount of probability mass that is transferred from ξ^σ to ξ^ω. Hence, we can determine the worst-case probability distribution by considering all the support points, ξ^ω, for which there is at least one positive dual variable. With this in mind, our goal is to illustrate, in a small instance of model (31), how the ambiguity set shapes the support of the worst-case distribution. In this experiment we use a single facility and two demand sites, and assume that the demands are lognormal. Figure 3 shows the support of the worst-case probability distribution for various values of the radius, ε, for our two DRO formulations of interest.



(a) Hyper-rectangular support (b) Optimal quadratic transport

Figure 3: Support of the worst-case probability distribution for various values of the radius, ε. The nominal support points are identical in the two plots. Part (a): Under $\varepsilon = 1$ the probability mass values associated with $(2.4, 5.5)$, $(4.9, 1.7)$, $(9.7, 1.7)$, and $(9.7, 9.7)$ are 0.086, 0.17, 0.03, and 0.014 (and $\frac{1}{10}$ for the remaining seven support points, including the new point $(9.7, 1.2)$); when $\varepsilon = 10$ the probability mass on $(9.7, 1.0)$ is 0.985 and that on $(9.7, 9, 7)$ is 0.015. Part (b): The probability mass on all points of support is $\frac{1}{10}$ except for $(9.7, 1.0)$, which splits its mass between $(9.73, 1.0)$ and $(18.06, 9.33)$ with probabilities 0.086 and 0.014 when $\varepsilon = 1$; and between $(9.79, 1.0)$ and $(36.14, 27.35)$ with probabilities 0.086 and 0.014 when $\varepsilon = 10$.

When the ambiguity set uses hyper-rectangular support (see Figure 3a), new support points appear at the boundary of Ξ, including the upper-right corner, which correspond to pessimistic demand realizations. As the radius increases, the worst-case distribution is supported on fewer points as most of the probability mass from the original support, Ξ^n, moves to pessimistic support

points. In contrast, with an ambiguity set based on the optimal quadratic transport distance and $\Xi = \mathbb{R}^m$, the support points of the worst-case probability distribution drift away from Ξ^n in a direction that is determined by the optimal value of the dual variable, π^σ for each $\sigma \in [n]$, scaled by $\frac{1}{\gamma}$. Figure 3b plots the support points on a log scale. In this particular instance, one of the support point splits (i.e., the support point $(9.7, 1.0)$) and the rest of the points move more modestly away from their corresponding support with mass $\frac{1}{n}$.

6 Conclusions

We have studied several DRO formulations in the context of a two-stage stochastic program with recourse and right-hand side uncertainty. These models vary in how the ambiguity set of candidate distributions is defined. Among distance-based ambiguity sets, we focus on the Wasserstein metric and its extension, the optimal transport distance. For ambiguity sets that use the Wasserstein approach with an L_p norm, we show that the DRO model is equivalent to the SAA problem when the support of the random parameters is a convex cone (e.g., $\mathbb{R}^m$ or $\mathbb{R}^m_+$), and therefore there is no benefit to using this type of DRO model for the class of two-stage programs that we consider. In contrast to unbounded supports, when such a set is a hyper-rectangle, the DRO formulation is meaningful in the sense that it can provide a first-stage solution that hedges against distributional uncertainty. Solving a DRO problem with hyper-rectangular support requires solving bilinear subproblems, which can be computationally challenging. These subproblems can be further reformulated as a mixed-integer program that is significantly more tractable and suitable for a cutting-plane algorithm. The final DRO formulation that we consider revisits $\mathbb{R}^m$ as the support of the random parameters, but uses a quadratic cost function to define the optimal transport distance. Here, we show that the DRO problem does not simplify to the SAA problem as before. Intuitively, this result is driven by the faster rate of growth of the optimal transport distance (quadratic), relative to the rate of growth of the second-stage cost (linear). To solve this DRO variant, the requisite subproblem in a cutting-plane algorithm is a non-convex quadratic program.

We test our DRO models on a supply-allocation problem. The first-stage decision allocates supply of a single commodity to a set of facilities, while the second-stage recourse problem distributes supply from the facilities to a set of demand sites. Demand is model as a random vector, and we consider both lognormal and uniform distributions. We explore the out-of-sample cost of three DRO formulations, and show their advantage relative to solving the SAA problem. when the demand follows a lognormal distribution. For an appropriately sized ambiguity set, the DRO solution improves both the mean and 90th percentile of the out-of-sample cost distribution. In contrast to lognormal demands, when the demand is uniformly distributed, there is no improvement relative to the SAA solution. We also explore how the support of the worst-case probability distribution changes as the size of the ambiguity set grows. In particular, we look at the cases in which the ambiguity set uses a hyper-rectangular support equipped with a one-norm Wasserstein

distance and an unbounded support equipped with the quadratic optimal transport distance. In the former, the number of support points with probability mass reduces as the size of the ambiguity set grows. In the unbounded case, support points drift away from the given scenarios, with one of the scenarios splitting into two support points.

Acknowledgements

This work was supported, in part, by the U.S. Department of Homeland Security under Grant Award 2017-ST-061-QA0001 and by Northwestern University's Center for Optimization & Statistical Learning. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. The authors also thank Bernardo Pagnoncelli for helpful suggestions.

References

[1] E. J. Anderson and A. B. Philpott. Robust sample average approximation with small sample sizes. *Optimization Online,* `http://www.optimization-online.org/DB_HTML/2019/02/7092.html` , 2019.

[2] G. Bayraksan and D.K. Love. Data-driven stochastic programming using phi-divergences. In D. Aleman and A. Thiele, editors, *Tutorials in Operations Research: The Operations Research Revolution*, volume 45, pages 1–19. INFORMS, Catonsville, MD, 2015.

[3] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23:769–805, 1998.

[4] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

[5] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Mathematical Programming*, 171:217–282, 2018.

[6] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44:565–600, 2019.

[7] J. Blanchet, K. Murthy, and F. Zhang. Optimal transport based distributionally robust optimization: structural properties and iterative schemes. 2018.

[8] G.C. Calafiore and L. El-Ghaoui. On distributionally robust chance-constrainted linear programs. *Journal of Optimization Theory and Applications*, 130:1–22, 2006.

[9] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data–driven problems. *Operations Research*, 55(3):98–112, 2010.

[10] R. Gao and A.J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. 2016.

[11] G.A. Hanasusanto and D. Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, 66:849–869, 2018.

[12] R. Jiang and Y. Guan. Risk-averse two-stage stochastic program with distribution ambiguity. *Operations Research*, 66(5):1390–1405, 2018.

[13] F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *European Journal of Operational Research*, 167:20–35, 2019.

[14] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulation. *Mathematical Programming*, 171(1):115–166, 2018.

[15] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. 2019.

[16] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

[17] Z. Wang, P.W. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.

[18] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

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

[20] C. Zhao. *Data-driven risk-averse stochastic program and renewable energy integration.* PhD thesis, University of Florida, 2014.

[21] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46:262–267, 2018.