Stochastic Decomposition Method for Two-Stage Distributionally Robust Optimization

Harsha Gangammanavar* and Manish Bansal†

First version: November 4, 2020; Current Version: February 15, 2022

Abstract. In this paper, we present a sequential sampling-based algorithm for the two-stage distributionally robust linear program (2-DRLP) with general ambiguity set. The algorithm is a distributionally robust version of the well-known stochastic decomposition algorithm of Higle and Sen (Math. of OR 16(3), 650-669, 1991) that was designed for risk-neutral two-stage stochastic linear programs. We refer to the algorithm as the distributionally robust stochastic decomposition (DRSD) method. The algorithm works with data-driven approximations of ambiguity sets that are constructed during the course of the algorithm using samples of increasing size. It constructs statistical approximations of the worst-case expectation function by solving subproblems corresponding to the latest observation(s) in every iteration. We show that the DRSD method asymptotically identifies an optimal solution, with probability one, for a family of ambiguity sets that includes the moment-based and Wasserstein distance-based ambiguity sets. We also computationally evaluate the performance of the DRSD method for solving distributionally robust variants of instances considered in the stochastic programming literature. The numerical results corroborate our analysis of the DRSD method and illustrate the computational advantage over an external sampling-based decomposition approach (distributionally robust L-shaped method).

1 Introduction

Stochastic programming (SP) is a well-known framework for decision-making under uncertainty that arises in applications such as finance, capacity expansion, manufacturing, wildfire planning, power systems, healthcare, and many more. The SP models with recourse, particularly in a two-stage setting, have gained wide acceptance across these application domains. In the two-stage SP models, the first-stage decision (referred to as the here-and-now decision) is taken before the realization of uncertainty. Following this, the second-stage decision (referred to as the wait-and-see decision) is taken in response to the first-stage decision and a realization of the uncertain data. In the classical setting of two-stage stochastic linear programs (2-SLPs), the decisions are solutions to linear programs in both stages [11].

The SP models are stated with an expectation-valued objective function. Therefore, stating an SP model either requires complete knowledge of the underlying probability distribution or the ability to simulate observations from this distribution. The latter leads to the construction of a sample average approximation (SAA) of the problem. In many practical applications, the distribution associated with random parameters in the optimization model is not precisely known. It either has to be estimated

*Department of Engineering Management, Information, and Systems, Southern Methodist University, Dallas, TX 75275 (harsha@smu.edu).

†Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061 (bansal@vt.edu).

from data or constructed by expert judgments, which tend to be subjective. In any case, identifying a distribution using available information may be cumbersome at best. Stochastic min-max programming that has gained significant attention in recent years under the name of *distributionally robust optimization* (DRO) is intended to address the ambiguity in distributional information.

In this paper, we study a particular manifestation of the DRO problem in the two-stage setting, viz., the two-stage distributionally robust linear program (2-DRLP). We state this problem as:

$$\min \ \{f(x) = c^\top x + \mathbb{Q}(x) \mid x \in \mathcal{X}\}. \tag{1}$$

Here, c is the coefficient vector of a linear cost function and $\mathcal{X}$ is the feasible set of the first-stage decision vector. The feasible region takes the form of a compact polyhedron, i.e., $\mathcal{X} = \{x \in \mathbb{R}^{d_x} \mid Ax \geq b, x \geq 0\}$, where $A \in \mathbb{R}^{m_1 \times d_x}$ and $b \in \mathbb{R}^{m_1}$. The function $\mathbb{Q}(x)$ is the worst-case expected recourse cost, that we define as:

$$\mathbb{Q}(x) = \max_{P \in \mathfrak{P}} \ \Big\{ \mathcal{Q}(x; P) := \mathbb{E}_P[Q(x, \tilde{\omega})] \Big\}. \tag{2}$$

We define the random vector $\tilde{\omega} \in \mathbb{R}^d$ on a measurable space $(\Omega, \mathcal{F})$, where Ω is a continuous or discrete sample space equipped with the sigma-algebra $\mathcal{F}$. $\mathfrak{P}$ is a set of probability distributions defined on the measurable space $(\Omega, \mathcal{F})$. The set of probability distributions $\mathfrak{P}$ is referred to as the *ambiguity set*. The expectation operation $\mathbb{E}_P[\cdot]$ is taken with respect to a probability distribution $P \in \mathfrak{P}$. For a given $x \in \mathcal{X}$, we refer to the optimization problem in (2) as the *distribution separation problem*. For a given realization ω of the random vector $\tilde{\omega}$ and a first-stage solution x, the recourse value in (2) is the optimal value of the following second-stage linear program:

$$\begin{aligned} Q(x, \omega) := \min \quad & g(\omega)^\top y \\ \text{s.t.} \quad & y \in \mathcal{Y}(x, \omega) := \big\{ W(\omega) y = r(\omega) - T(\omega) x, \ y \in \mathbb{R}^{d_y}_+ \big\}. \end{aligned} \tag{3}$$

Here, for each $\omega \in \Omega$, we have uncertain second-stage parameters: $g(\omega)$, the recourse matrix $W(\omega)$, the right-hand side vector $r(\omega)$, and the technology matrix $T(\omega)$ of appropriate dimensions. A special case of 2-DRLP is the 2-SLP where $\mathfrak{P}$ is a singleton, i.e., $\mathfrak{P} = \{P^\star\}$, resulting in the following optimization problem:

$$\min \ \{c^\top x + \mathbb{E}_{P^\star}[Q(x, \tilde{\omega})] \mid x \in \mathcal{X}\}. \tag{4}$$

Most data-driven and SAA-based approaches to solve 2-SLPs tackle the problem in two steps. In the first simulation/sampling step, an uncertainty representation is generated using a finite set of observations that serves as an approximation of Ω and the corresponding empirical distribution serves as an approximation of $P^\star$. For a given uncertainty representation, one obtains a deterministic approximation of (4). In the second optimization step, the approximate problem is solved using deterministic optimization methods. Such a two-step approach may lead to poor out-of-sample performance, forcing the entire process to be repeated from scratch with an improved uncertainty representation. Since sampling is performed prior to the optimization step, this two-step approach is also referred to as the *external sampling procedure*. This procedure has also been utilized for solving 2-DRLPs where in the first step, an approximation of the ambiguity set $\mathfrak{P}$ is obtained using a finite set of observations. Then, in the second step, a deterministic min-max problem, i.e., Problem (1) where expectation operator is replaced by summation over the finite sample, is solved. Once again, using a finite sample to approximate the original sample space may result in similar out-of-sample performance as in the case for 2-SLP.

1.1 Contributions

In light of the above observations regarding the two-step external sampling procedure, the main contributions of this manuscript are as follows.

1. *A Sequential Sampling Algorithm*: We present a sequential sampling approach for solving a 2-DRLP. We refer to this algorithm as the *distributionally robust stochastic decomposition* (DRSD) algorithm following its risk-neutral predecessor, the two-stage stochastic decomposition (SD) method [23] that was designed for 2-SLPs. The DRSD algorithm concurrently performs the simulation and optimization steps in every iteration. In the simulation step, new observation(s) are included to improve the representation of the ambiguity set. The sequential inclusion of observations results in approximate ambiguity sets that evolve over the course of the algorithm. In the optimization step, the solution is updated in an online manner by solving second-stage programs for only the new observation(s) in each iteration. In this sense, the DRSD method is an *internal sampling procedure*. Moreover, the algorithmic design of the DRSD does not depend on any specific ambiguity set description. Hence, this method is suitable for any (general) ambiguity set for which the distribution separation problem (2) can be solved efficiently.

2. *Convergence Analysis*: The DRSD method is an inexact bundle method that creates outer linearization for the dynamically evolving approximation of the first-stage problem. We provide the asymptotic analysis of DRSD and identify conditions on ambiguity sets under which the sequential sampling approach identifies an optimal solution to the 2-DRLP in (1) with probability one.

3. *Computational Evidence of Performance*: We provide the first set of experiments that illustrates the advantages of a sequential sampling approach to solving 2-DRLPs. We demonstrate these advantages through computational experiments conducted on well-known problems in the SP literature. These problems are modified to create distributionally robust variants with moment-based, ℓ_1-type Wasserstein, and ℓ_∞-type Wasserstein ambiguity sets.

1.2 Related work

For 2-SLPs with finite support, including the SAA problems, the L-shaped method due to Van Slyke and Wets [45] has proven to be very effective. Other algorithms for 2-SLPs such as the Dantzig-Wolfe decomposition [12] and the progressive hedging (PH) algorithm [35] also operate on problems with finite support. The well-established theory of SAA (see Chapter 5 in [42]) supports the external sampling procedure for 2-SLP. The quality of the solution obtained by solving an SAA problem is assessed using the procedures developed, e.g., in [4]. When the quality of the SAA solution is not acceptable, a new SAA is constructed with a larger number of observations. Prior works, such as [5] and [37], provide rules on how to choose the sample sizes in a sequential SAA procedure.

In contrast to the above, SD incorporates one new observation in every iteration to create approximations of the dynamically updating SAAs of (4). First proposed in [23], this method has seen significant development in the past three decades with the introduction of the quadratic proximal term [24], statistical optimality rules [26], and extensions to multistage stochastic linear programs [20]. The DRSD method extends the notion of sequential sampling of SD to DRO problems.

The concept of DRO dates back to the work of Scarf [39], and has gained significant attention in recent years. Readers can refer to [33] for a comprehensive treatment on various aspects of the DRO. The algorithmic works on DRO are either decomposition-based or reformulation-based approaches. The

decomposition-based methods for 2-DRLP mimic the two-stage SP approach of using a deterministic representation of the sample space using a finite number of observations. As a consequence, the SP solution methods with suitable adaptation can be applied to solve the 2-DRLP problems. For instance, Breton and El Hachem [10] applied the PH algorithm for a 2-DRLP model with a moment-based ambiguity set. Riis and Anderson [34] extended the L-shaped method for 2-DRLP with continuous recourse and moment-based ambiguity set. Bansal et al. [1] extended the algorithm in [34], which they refer to as the distributionally robust (DR) L-shaped method, to solve 2-DRLPs, with an ambiguity set defined by a polytope. Further extensions of this decomposition approach are presented in [1] and [2] for two-stage DRO problems with mixed-binary recourse and disjunctive programs, respectively. We discuss key differences of DRSD with SD and DR L-shaped method in Remark 3.2 at the end of §3.

Another predominant approach to solve 2-DRLP problems is to reformulate the distribution separation problem in (2) as a minimization problem, pose the problem in (1) as a single deterministic optimization problem, and use off-the-shelf deterministic optimization tools to solve the reformulation. For example, Shapiro and Kleywegt [43] and Shapiro and Ahmed [41] used such an approach for a 2-DRLP with moment matching set. They derived an equivalent stochastic program defined with a reference distribution. Bertsimas et al. [7] provided tight semidefinite programming reformulations for 2-DRLP where the ambiguity set is defined using multivariate distributions with known first and second moments. Likewise, Hanasusanto and Kuhn [21] provided a conic programming reformulation for 2-DRLP where the ambiguity set comprises of a ℓ_2-type Wasserstein ball centered at a discrete distribution. Xie [46] provided similar reformulations to tractable convex programs for 2-DRLP problems with ambiguity set defined using ℓ_∞ Wasserstein metric. By taking the dual of the inner maximization problem, Love and Bayraksan [3] demonstrated that a 2-DRLP with the ambiguity set defined using ϕ-divergence and finite sample space is equivalent to 2-SLP with a coherent risk measure. A similar reformulation approach is employed in [15] for ambiguity sets defined using Wasserstein and quadratic transport function on unbounded and hyper-rectangle support. Jiang and Guan [28] reduced the worst-case expectation in 2-DRLP, where the ambiguity set is defined using the ℓ_1-norm on the space of all (continuous and discrete) probability distributions, to a convex combination of CVaR and an essential supremum. Under the assumption of finite support, [27] showed that a 2-DRLP with CVaR objective can be reformulated into a linear program. On the other hand, the two-stage DRO problem with a linear recourse was reformulated as a conic optimization problem under an assumption that second-stage decisions are affine functions of the random vector in [29]. When reformulations result in equivalent stochastic programs (as in [3, 27, 28, 41], for instance), an SAA of the reformulation is used to obtain an approximate problem. This approximate problem is amenable to standard cutting plane or bundle type methods prevalent in SP.

Data-driven approaches for DRO have been presented for specific ambiguity sets. In [13], problems with ellipsoidal moment-based ambiguity set whose parameters are estimated using sampled data are addressed. Esfahani et al. [31] tackled data-driven problems with Wasserstein metric-based ambiguity sets with convex reformulations. In both these works, the authors provide finite-sample performance guarantees that probabilistically bound the gap between approximate and true DRO problems. Sun and Xu presented asymptotic convergence analysis of DRO problems with ambiguity sets that are based on moments and mixture distributions constructed using a finite set of observations in [44]. A practical approach to incorporate the results of these works to identify a high-quality DRO solution is similar to the sequential SAA procedure for SP in [5]. Such an approach involves the following steps performed in series – obtaining a deterministic representation of ambiguity set using sampled observations, applying appropriate reformulation, and solving the resulting deterministic optimization problem. If the quality of the solution is deemed insufficient, then the entire series of steps is repeated with an improved representation of the ambiguity set (possibly with a larger number of observations).

Organization

We organize the remainder of the paper as follows. In §2, we present the two key ideas of the DRSD-the sequential approximation of the ambiguity set and the recourse function. We provide a detailed description of the DRSD method in §3. We show the convergence of the value functions and solutions generated by the DRSD method in §4. We present results of our computational experiments in §5, and finally we conclude and discuss potential future directions in §6.

Notations and Definitions

We define the ambiguity sets over $\mathcal{M}$, the set of all finite signed measures on the measurable space $(\Omega, \mathcal{F})$. A nonnegative measure that satisfies $P(\Omega) = 1$ is a probability distribution. For probability distributions $P, P' \in \mathfrak{P}$, we define

$$\text{dist}(P, P') := \sup_{F \in \mathcal{F}} \left| \mathbb{E}_P[F(\tilde{\omega})] - \mathbb{E}_{P'}[F(\tilde{\omega})] \right| \tag{5}$$

as the uniform distance of expectation, where $\mathcal{F}$ is a class of measurable functions. The above is the distance with ζ-structure that is used for stability analysis in SP [36]. The distance between a single probability distribution P to a set of distributions $\mathfrak{P}$ is given as $\text{dist}(P, \mathfrak{P}) = \inf_{P' \in \mathfrak{P}} \text{dist}(P, P')$. The distance between two sets of probability distributions $\mathfrak{P}$ and $\widehat{\mathfrak{P}}$ is given as

$$\mathbb{D}(\mathfrak{P}, \widehat{\mathfrak{P}}) := \sup_{P \in \widehat{\mathfrak{P}}} \text{dist}(P, \mathfrak{P}). \tag{6}$$

Finally, the Hausdorff distance between $\mathfrak{P}$ and $\widehat{\mathfrak{P}}$ is defined as

$$\mathbb{H}(\mathfrak{P}, \widehat{\mathfrak{P}}) := \max\{\mathbb{D}(\mathfrak{P}, \widehat{\mathfrak{P}}),\ \mathbb{D}(\widehat{\mathfrak{P}}, \mathfrak{P})\}. \tag{7}$$

With suitable definitions for the set $\mathcal{F}$, the distance in (5) accepts the bounded Lipschitz, the Kantorovich and the p-th order Fortet-Mourier metrics (see [36]).

2 Approximating Ambigiuty Set and Recourse Function

In this section, we present the building blocks that we embed within a sequential sampling setting of the DRSD method. Specifically, we present procedures to approximate ambiguity set $\mathfrak{P}$ and recourse function $Q(x, \omega)$ in an iteration of the DRSD. Going forward we make the following assumptions on the 2-DRLP models:

(A1) The first-stage feasible region $\mathcal{X}$ is a non-empty and compact set.

(A2) $Q(\cdot)$ satisfies relatively complete recourse. The dual feasible region of the recourse problem is a nonempty compact polyhedral set. The transfer (or technology) matrix satisfies $\sup_{P \in \mathfrak{P}} \mathbb{E}_P[T(\tilde{\omega})] < \infty$.

(A3) The randomness only affects the right-hand sides of constraints in (3).

(A4) The sample space Ω is a compact metric space and the ambiguity set $\mathfrak{P}$ is nonempty.

As a consequence of (A2), the recourse function satisfies $Q(x,\tilde{\omega}) < \infty$ with probability one for all $x \in \mathcal{X}$. It also implies that the second-stage feasible region, i.e., $\{y \ : \ Wy = r(\omega) - T(\omega)x, \ y \geq 0\}$, is non-empty for all $x \in \mathcal{X}$ and every $\omega \in \Omega$. The non-empty dual feasible region $\mathcal{D}$ implies that there exists a constant $L > -\infty$ such that $Q(x,\tilde{\omega}) > L$, almost surely. Without loss of generality, we assume that $L = 0$. As a consequence of (A3), the cost coefficient vector g and the recourse matrix W are not affected by uncertainty. Problems that satisfy (A3) are said to have a fixed recourse. Finally, the compactness of the support Ω guarantees that every probability measure $P \in \mathfrak{P}$ is tight.

2.1 Approximating the Ambiguity Set

The DRO approach assumes only partial knowledge about the underlying uncertainty that is captured by a suitable description of the ambiguity set. An ambiguity set must capture the true distribution with an absolute or high degree of certainty and must be computationally manageable. The description of the ambiguity set involves parameters that are determined based on a practitioner's risk preferences. The ambiguity set descriptions that are prevalent in the literature include moment-based ambiguity sets with linear constraints (e.g., [14]) or conic constraints (e.g., [13]); Kantorovich distance or Wasserstein metric-based ambiguity sets [30]; ζ-structure metrics [47], ϕ-divergences such as χ^2 distance and Kullback-Leibler divergence [6]; Prokhorov metrics [16], among others. In this section, we present steps to construct approximate ambiguity sets in a data-driven manner. We use moment-based and Wasserstein distance-based ambiguity sets to illustrate these steps.

In a data-driven setting, the parameters used in the description of ambiguity sets are estimated using a finite set of independent observations which can either be past realizations of the random variable $\tilde{\omega}$ or generated using computer simulations. We will denote such a sample by $\Omega^k \subseteq \Omega$. When one observation is added to the sample in every iteration, we obtain $\Omega^k = \{\omega^j\}_{j=1}^k$. Naturally, we can view Ω^k as a random sample and define the empirical frequency

$$\hat{p}^k(\omega) = \frac{\kappa(\omega)}{k} \qquad \text{for all } \omega \in \Omega^k, \tag{8}$$

where $\kappa(\omega)$ denotes the number of times observation ω is observed in the sample. Since in the sequential sampling setting, the sample set is updated within the optimization algorithm, it is worthwhile to note that the empirical frequency can be updated using the following recursive equations:

$$\hat{p}^k(\omega) = \begin{cases} \theta^k \hat{p}^{k-1}(\omega) & \text{if } \omega \in \Omega^{k-1}, \omega \neq \omega^k \\ \theta^k \hat{p}^{k-1}(\omega) + (1-\theta^k) & \text{if } \omega \in \Omega^{k-1}, \omega = \omega^k \\ (1-\theta^k) & \text{if } \omega \notin \Omega^{k-1}, \omega = \omega^k. \end{cases} \tag{9}$$

where $\theta^k = \frac{k-1}{k}$. In general, when more than one observation is added to the sample in every iteration, we have $\theta^k \in (0,1)$. We will succinctly denote the above using the operator $\Theta^k : \mathbb{R}^{|\Omega^{k-1}|} \to \mathbb{R}^{|\Omega^k|}$.

In this paper, we focus on a setting where the ambiguity set $\mathfrak{P}$ is replaced by a sequence of *approximate ambiguity sets* $\{\widehat{\mathfrak{P}}^k\}_{k>0}$ such that the following properties are satisfied: $(B1)$ for any $P \in \widehat{\mathfrak{P}}^{k-1}$, there exists $\theta^k \in (0,1)$ such that $\Theta^k(P) \in \widehat{\mathfrak{P}}^k$ and $(B2)$ $\mathbb{H}(\widehat{\mathfrak{P}}^k, \mathfrak{P}) \to 0$ as $k \to \infty$, almost surely. We show that approximate ambiguity sets for the moment-based ambiguity set $\mathfrak{P}_{\text{mom}}$ and Wasserstein distance-based ambiguity set $\mathfrak{P}_{\text{w}}$ can be constructed such that these properties are satisfied (Propositions 2.1 and 2.2, respectively).

Let $\mathcal{F}^k = \sigma(\omega^j \mid j \leq k)$ be the σ-algebra generated by the observations in the sample Ω^k. Notice that $\mathcal{F}^{k-1} \subseteq \mathcal{F}^k$, and hence, $\{\mathcal{F}^k\}_{k\geq 1}$ is a filtration. We will define the approximate ambiguity sets over

the measurable space $(\Omega^k, \mathcal{F}^k)$. These sets should be interpreted to include all distributions that could have been generated using the sample Ω^k, which share a certain relationship with sample statistics. We will use $\mathcal{M}^k$ to denote the finite signed measures on $(\Omega^k, \mathcal{F}^k)$.

2.1.1 Moment-based Ambiguity Sets

Given the first q moments associated with the random variable $\tilde{\omega}$, the moment-based ambiguity set can be defined as

$$\mathfrak{P}_{\text{mom}} = \left\{ P \in \mathcal{M} \;\middle|\; \begin{array}{ll} \int_\Omega dP(\tilde{\omega}) = 1, & \\ \int_\Omega \psi_i(\tilde{\omega}) dP(\tilde{\omega}) = b_i & i = 1, \ldots, q \end{array} \right\}. \tag{10}$$

While the first constraint ensures the definition of a probability measure, the moment requirements are guaranteed by the second constraints. Here, $\psi_i(\tilde{\omega})$ denotes a real valued measurable function on $(\Omega, \mathcal{F})$ and $b_i \in \mathbb{R}$ is a scalar for $i = 1, \ldots, q$. Existence of moments ensures that $b_i < \infty$ for all $i = 1, \ldots, q$. Notice that the description of the ambiguity set requires explicit knowledge of the following statistics: the support Ω and the moments b_i for $i = 1, \ldots, q$. In the data-driven setting, the support is approximated by Ω^k and the sample moments $\hat{b}_i^k = (1/k) \sum_{j=1}^k \psi_i(\omega^j)$ are used to define the following approximate ambiguity set

$$\widehat{\mathfrak{P}}^k_{\text{mom}} = \left\{ P \in \mathcal{M}^k \;\middle|\; \begin{array}{ll} \sum_{\omega \in \Omega^k} p(\omega) = 1, & \\ \sum_{\omega \in \Omega^k} p(\omega) \psi_i(\omega) = \hat{b}_i^k & i = 1, \ldots, q \end{array} \right\}. \tag{11}$$

The following result characterizes the relationship between distributions drawn from the above approximate ambiguity set, as well as asymptotic behavior of the sequence $\{\widehat{\mathfrak{P}}^k_{\text{mom}}\}_{k \geq 1}$.

Proposition 2.1. *For any $P \in \widehat{\mathfrak{P}}^{k-1}_{\text{mom}}$, we have $\Theta^k(P) \in \widehat{\mathfrak{P}}^k_{\text{mom}}$. Further, suppose $\widehat{\mathfrak{P}}^k_{\text{mom}} \neq \emptyset$ for all $k \geq 1$, $\mathbb{H}(\widehat{\mathfrak{P}}^k_{\text{mom}}, \mathfrak{P}_{\text{mom}}) \to 0$ as $k \to \infty$, almost surely.*

Proof. See Appendix §A. □

In the context of DRO, similar ambiguity sets have been studied in [8, 14] where only the first moment (i.e., $q = 1$) is considered. The above form of ambiguity set also relates to those used in [13, 34, 39, 44] where constraints were imposed only on the mean and covariance. In the data-driven setting of [13] and [44], the statistical estimates are used in constructing the approximate ambiguity set as in the case of (11). However, the ambiguity sets in these previous works are defined over the original sample space Ω, as opposed to Ω^k that is used in (11). This marks a critical deviation in the way the approximate ambiguity sets are constructed.

Remark 2.1. When moment information is available about the underlying distribution $P^\star$, an approximate moment-based ambiguity set with constant parameters in (11) (i.e., with $\hat{b}_i^k = b_i$ for all k) can be constructed. Such an approximate ambiguity set defined over Ω^k is studied in [34]. Notice that these approximate ambiguity sets satisfy $\cup_{k \geq 1} \widehat{\mathfrak{P}}^k \subseteq \mathfrak{P}$ and $\widehat{\mathfrak{P}}^k \subseteq \widehat{\mathfrak{P}}^{k+1}$, for all $k \geq 1$. Therefore, they satisfy the properties (i) and (ii) necessary for approximate ambiguity sets.

2.1.2 Wasserstein distance-based Ambiguity Sets

We next present approximations of another class of ambiguity sets that has gained significant attention in the DRO literature, viz., the Wasserstein distance-based ambiguity sets. Consider probability distri-

butions $\mu_1, \mu_2 \in \mathcal{M}$, and a function $\nu : \Omega \times \Omega \to \mathbb{R}_+ \cup \{\infty\}$ such that ν is symmetric, $\nu^{\frac{1}{r}}(\cdot)$ satisfies triangle inequality for $1 \leq r < \infty$, and $\nu(\omega_1, \omega_2) = 0$ whenever $\omega_1 = \omega_2$. If $\mathcal{J}(\mu_1, \mu_2)$ denotes the joint distribution of random vectors ω_1 and ω_2 with marginals μ_1 and μ_2, respectively, then the Wasserstein metric of order r is given by

$$d_{\mathrm{w}}(\mu_1, \mu_2) = \left[\inf_{\eta \in \mathcal{J}(\mu_1, \mu_2)} \left\{ \int_{\Omega \times \Omega} \nu^r(\omega_1, \omega_2) \eta(d\omega_1, d\omega_2) \right\} \right]^{1/r}. \tag{12}$$

In the above definition, the decision variable $\eta \in \mathcal{J}$ can be viewed as a plan to transport goods/mass from an entity whose spatial distribution is given by the measure μ_1 to another entity with spatial distribution μ_2. Therefore, the $d_{\mathrm{w}}(\mu_1, \mu_2)$ measures the optimal transport cost between the measures. Notice that an arbitrary norm $\| \bullet \|^r$ on $\mathbb{R}^d$ satisfies the requirement of the function $\nu(\cdot)$. In our presentation, we will use the ℓ_1 Wasserstein metric. However, the definition of the approximate ambiguity sets and their use within the solution method are applicable to ambiguity sets defined using Wasserstein metric of higher orders. Using the ℓ_1 Wasserstein metric, we define an ambiguity set as follows:

$$\mathfrak{P}_{\mathrm{w}} = \{P \in \mathcal{M} \mid d_{\mathrm{w}}(P, P^*) \leq \epsilon\} \tag{13}$$

for a given $\epsilon > 0$ and a reference distribution P^*. In practice, the value of ϵ is chosen based on user's risk preferences; a smaller value indicates lower risk aversion. As done in §2.1.1, we present approximate Wasserstein distance-based ambiguity sets defined over the measurable space $(\Omega^k, \mathcal{F}^k)$ as follows:

$$\widehat{\mathfrak{P}}^k_{\mathrm{w}} = \{P \in \mathcal{M}^k \mid d_{\mathrm{w}}(P, \widehat{P}^k) \leq \epsilon\}, \tag{14}$$

where $\widehat{P}^k = (\hat{p}^k(\omega))_{\omega \in \Omega^k}$. For this approximate ambiguity set, the distribution separation problem in (2) is a finite dimensional linear program:

$$\max \sum_{\omega \in \Omega^k} p(\omega) Q(x, \omega) \tag{15a}$$

$$\text{subject to } P \in \widehat{\mathfrak{P}}^k_{\mathrm{w}} = \left\{ P \in \mathcal{M}^k \;\middle|\; \begin{array}{ll} \sum_{\omega \in \Omega^k} p(\omega) = 1 & \\ \sum_{\omega' \in \Omega^k} \eta(\omega, \omega') = p(\omega) & \forall \omega \in \Omega^k, \\ \sum_{\omega \in \Omega^k} \eta(\omega, \omega') = \hat{p}^k(\omega') & \forall \omega' \in \Omega^k, \\ \sum_{(\omega, \omega') \in \Omega^k \times \Omega^k} \|\omega - \omega'\| \eta(\omega, \omega') \leq \epsilon & \\ \eta(\omega, \omega') \geq 0 \quad \forall \omega, \omega' \in \Omega^k & \end{array} \right\}. \tag{15b}$$

Note that when Wasserstein metric of order $r > 1$ is used in the definition of the ambiguity sets, the foregoing optimization problem remains a finite dimensional linear program. In this case, the coefficients $\|\omega - \omega'\|$ and right-hand side ϵ in the fourth set of constraints in (15b) must be replaced by $\|\omega - \omega'\|^r$ and ϵ^r, respectively. The following result characterizes the distributions drawn from the approximate ambiguity sets of the form in (14), or equivalently (15b).

Proposition 2.2. *Under compactness of the support set $\Omega \subset \mathbb{R}^d$, i.e., (A4), with $d > 2$, the sequence of Wasserstein distance-based approximate ambiguity sets satisfies the following properties (i) for any $P \in \widehat{\mathfrak{P}}^{k-1}_{\mathrm{w}}$, we have $\Theta^k(P) \in \widehat{\mathfrak{P}}^k_{\mathrm{w}}$, and (ii) $\mathbb{H}(\widehat{\mathfrak{P}}^k_{\mathrm{w}}, \mathfrak{P}_{\mathrm{w}}) \to 0$ as $k \to \infty$, almost surely.*

Proof. See appendix §A. □

Note that, as in the case of moment-based ambiguity set, we also define Wasserstein distance-based

approximate ambiguity set over an approximation of the true sample space, i.e., Ω^k. This approach precludes the need for exact knowledge of the sample space and allows us to depend only on what is known until iteration k.

Remark 2.2. In [31], an approach that involves solving a sequence of DRO problems is used to tackle the risk-neutral 2-SLP problem (4). They use approximate ambiguity set to be a ball constructed in the space of probability distributions that are defined over the sample space Ω and whose radius reduces with an increase in the number of observations. Using Wasserstein balls of shrinking radii, the authors of [31] show that the optimal value of the sequence of DRO problems converges to the optimal value of the expectation-valued objective in (4) associated with the true distribution $P^\star$. A similar approach of involving a sequence of DRO problems is used in [47] to solve (4), albeit using ambiguity sets with $\zeta-$structure. In contrast to these works, our goal is to solve the DRO problem in (1). Therefore, we use a constant radius for all $k \geq 1$ to define the approximate ambiguity set in (14).

2.2 Approximating the Recourse Problem

Cutting plane methods for the 2-SLPs use an outer linearization-based approximation of the first-stage objective function in (4). In such algorithms, the challenging aspect of computing the expectation is addressed by taking advantage of the structure of the recourse problem (3). Specifically, for a given ω, the recourse value $Q(\cdot, \omega)$ is known to be convex in the right-hand side parameters that includes the first-stage decision vector x. Additionally, if (A2) holds, then the function $Q(\cdot, \omega)$ is polyhedral. Under assumptions (A2) and (A4), this structural property of convexity extends to the expected recourse value $\mathcal{Q}(x)$.

Due to the strong duality of linear programs, the recourse value is also equal to the optimal value of the dual of (3), i.e.,

$$\begin{aligned} Q(x, \omega) = \max \ & \pi^\top [r(\omega) - T(\omega)x] \\ & \text{subject to } \pi \in \mathcal{D} := \{\pi \mid W^\top \pi \leq g\}. \end{aligned} \tag{16}$$

Due to (A2) and (A4), the dual feasible region $\mathcal{D}$ is a polytope that is not impacted by the uncertainty. If $\Pi \subseteq \mathcal{D}$ denotes the set of all extreme points of the polytope $\mathcal{D}$, then the recourse value can also be expressed as the pointwise maximum of affine functions computed using elements of set Π:

$$Q(x, \omega) = \max_{\pi \in \mathcal{D}} \pi^\top [r(\omega) - T(\omega)x]. \tag{17}$$

The outer linearization approaches tend to approximate the above form of recourse function by identifying the extreme points (optimal solutions to (16)) at a sequence of candidate (or trial) solutions $\{x^k\}$, and generating the corresponding affine functions. If $\pi(x^k, \omega)$ is an optimal dual obtained by solving (16) with x^k as input, then the affine function $\alpha^k(\omega) + (\beta^k(\omega))^\top x$ is obtained by computing the coefficients $\alpha^k(\omega) = (\pi(x^k, \omega))^\top r(\omega)$ and $\beta^k(\omega) = T(\omega)^\top \pi(x^k, \omega)$. Following linear programming duality, notice that this affine function is a supporting hyperplane to $Q(x, \omega)$ at x^k, and lower bounds the function at every other $x \in \mathcal{X}$.

If the support Ω is finite, then one can solve a dual subproblem for all $\omega \in \Omega$ with the candidate solution as input, generate the affine functions, and collate them together to obtain an approximate first-stage objective function. This is the essence of the L-shaped method applied to 2-SLP in (4). In each iteration of the L-shaped method, the affine functions generated using a candidate solution x^k and information gathered from individual observations are weighed by the probability density of the

observation to update the approximation of the first-stage objective function. The L-shaped method can also be applied to the SAA of the 2-SLP with continuous sample space Ω that uses a sample $\Omega_N \subset \Omega$ of finite size N. A similar approximation strategy is used in the DR L-shaped method for 2-DRLP problems [1, 34].

Alternatively, we can consider the following approximation of the recourse function expressed in the form given in (17):

$$Q^k(x,\omega) = \max_{\pi \in \Pi^k} \pi^\top [r(\omega) - C(\omega)x]. \tag{18}$$

Notice that the above approximation is built using only a subset $\Pi^k \subset \Pi$ of extreme points, and therefore, satisfies $Q^k(x,\omega) \leq Q(x,\omega)$. Since $Q(x,\omega) \geq 0$, we begin with $\Pi^0 = \{0\}$. Subsequently, we construct a sequence of sets $\{\Pi^k\}$ such that $\Pi^0 \subseteq \ldots \Pi^k \subseteq \Pi^{k+1} \subseteq \ldots \subset \Pi$ that ensures $Q^k(x,\omega) \geq 0$ for all k. The following result from [23] captures the behavior of the sequence of approximation $\{Q^k\}$.

Proposition 2.3. *The sequence $\{Q^k(x,\omega)\}_{k\geq 1}$ converges uniformly to a continuous function on $\mathcal{X}$ for any $\omega \in \Omega$.*

Proof. See Appendix A. □

The approximation of the form in (18) is one of the principal features of the SD algorithm (see [23, 24]). While the L-shaped and DR L-shaped methods require finite support for $\tilde{\omega}$, SD is applicable even for problems with continuous support. The algorithm uses an "incremental" SAA for the first-stage objective function by adding one new observation in each iteration. Therefore, the first-stage objective function approximation used in SD is built using the recourse problem approximation in (18) and the incremental SAA. This approximation is given by:

$$\mathcal{Q}^k(x) = c^\top x + \frac{1}{k}\sum_{j=1}^{k} Q^k(x,\omega^j). \tag{19}$$

The affine functions generated in SD provide an outer linearization for the approximation in (19). The sequence of sets that grow monotonically in size, viz. $\{\Pi^k\}$, is generated by adding one new vertex to the previous set Π^{k-1} to obtain the updated set Π^k. The newly added vertex is an optimal dual solution obtained by solving (16) with the most recent observation ω^k and candidate solution x^k as input.

We refer the reader to [9], [1, 34], and [23, 25] for the a detailed exposition of the L-shaped, the DR L-Shaped, and the SD methods, respectively. Here, we only note the key differences between these methods. Firstly, the sample used in the (DR) L-shaped method is fixed before the optimization. In SD, this sample is updated dynamically throughout the course of the algorithm. Secondly, in the (DR) L-shaped method, subproblems corresponding to the current iterate and all observations in the sample are solved exactly. The resulting optimal dual solutions are used to compute the affine lower bounding functions (cuts). On the other hand, in SD, only two subproblems corresponding to the latest observation are solved exactly, while the subproblems corresponding to other observations in the sample use the approximation in (18).

3 Distributionally Robust Stochastic Decomposition

In this section, we provide a detailed description of the DRSD algorithm. The pseudocode of the DRSD method is given in Algorithm 1. In the following, we discuss the main steps of the algorithm in iteration

Algorithm 1 Distributionally Robust Stochastic Decomposition

1: **Input:** Incumbent solution $\hat{x}^0 \in \mathcal{X}$; initial sample $\Omega^0 \subseteq \Omega$; stopping tolerance $\tau > 0$; $\gamma \in (0,1)$, $\theta^1 = 0$, and maximum and minimum iterations $k^{\max} > k^{\min} > 1$.
2: **Initialization:** Set iteration counter $k \leftarrow 1$; $\Pi^0 = \emptyset$; $\mathcal{L}^0 = \emptyset$, and $f^0(x) = 0$.
3: **while** $(k \leq k^{\max})$ **do**
4: Solve the master problem (20) to obtain a candidate solution x^k.
5: **if** $k > k^{\min}$ and $f^{k-1}(\hat{x}^{k-1}) - f^{k-1}(x^k) < \tau f^{k-1}(\hat{x}^{k-1})$ **then**, Go to Line 28.
6: **end if**
7: Generate a scenario $\omega^k \in \Omega$ to get sample $\Omega^k \leftarrow \Omega^{k-1} \cup \{\omega^k\}$.
8: Solve the second-stage linear program (3) with (x^k, ω^k) as input;
9: Obtain the optimal value $Q(x^k, \omega^k)$ and optimal dual solution $\pi(x^k, \omega^k)$;
10: Update dual vertex set $\Pi^k \leftarrow \Pi^{k-1} \cup \{\pi(x^k, \omega^k)\}$.
11: **for** $\omega \in \Omega^k \setminus \{\omega^k\}$ **do**
12: Use the argmax procedure (21) to identify dual vertex $\pi(x^k, \omega^k)$;
13: Store $Q^k(x^k, \omega) = (\pi(x^k, \omega))^\top [r(\omega) - T(\omega)x^k]$.
14: **end for**
15: Solve the distribution separation problem using the ambiguity set $\widehat{\mathfrak{P}}^k$ and $\{Q^k(x^k, \omega)\}_{\omega \in \Omega^k}$ to get an extremal distribution $P^k := (p^k(\omega))_{\omega \in \Omega^k}$.
16: Derive affine function $\ell_k^k(x) = \alpha_k^k + (\beta_k^k)^\top x$ using $\{\pi(x^k, \omega)\}_{\omega \in \Omega^k}$ and P^k to get lower bound approximation of $\mathbb{Q}^k(x)$ as in (24);
17: Perform Steps 8-16 with $\hat{x}^{k-1}$ (incumbent solution) to obtain $\hat{\ell}_k^k(\cdot)$.
18: **for** $\ell_j^{k-1} \in \mathcal{L}^{k-1}$ **do**
19: Update previously generated affine functions $\ell_j^{k-1}(x)$:

$$\alpha_j^k = \theta^k \alpha_j^{k-1} \text{ and } \beta_j^k = \theta^k \beta_j^{k-1};$$

20: Set $\ell_j^k(x) = \alpha_j^k + (\beta_j^k)^\top x$ that provides lower bound approx. of $\mathbb{Q}^k(x)$;
21: **end for**
22: Build a collection of these affine functions, denoted by $\mathcal{L}^k$;
23: Update approximation of the first-stage objective function:

$$c^\top x + \mathbb{Q}^k(x) \geq f^k(x) = c^\top x + \max_{j \in \mathcal{L}^k} \{\alpha_j^k + (\beta_j^k)^\top x\};$$

24: If incumbent update rule (28) is satisfied, then set $\hat{x}^k \leftarrow x^k$ and $\hat{x}^k \leftarrow \hat{x}^{k-1}$, otherwise.
25: Update the master problem (20) by replacing $f^{k-1}(x)$ with $f^k(x)$;
26: $k \leftarrow k+1$; $\theta^k \leftarrow (k-1)/k$
27: **end while**
28: **return** Incumbent solution $\hat{x}^k$ and objective function estimate $f^k(\hat{x}^k)$.

k (Steps 4-26 of Algorithm 1). At the beginning of iteration k, we have a certain approximation of the first-stage objective function that we denote as $f^{k-1}(x)$, a finite set of observations Ω^{k-1} and an incumbent solution $\hat{x}^{k-1}$. We use the term *incumbent solution* to refer to the best solution discovered by the algorithm until iteration k. The solution identified in the current iteration is referred to as the *candidate solution* and denoted as x^k (without $\hat{\bullet}$).

Iteration k begins by first identifying the candidate solution by solving the following the master problem (Step 4):

$$x^k \in \arg\min \{f^{k-1}(x) \mid x \in \mathcal{X}\}. \tag{20}$$

Following this, a new observation $\omega^k \in \Omega$ is obtained and added to the current sample of observations Ω^{k-1} to get $\Omega^k = \Omega^{k-1} \cup \{\omega^k\}$ (Step 7).

In order to build the first-stage objective function approximation, we rely upon the recourse function

approximation presented in Section 2.2. For the most recent observation ω^k and the candidate solution x^k, we evaluate the recourse function value $Q(x^k, \omega^k)$ by solving (3), and obtain the dual optimum solution $\pi(x^k, \omega^k)$ in Steps 8–10. These dual vectors are added to a set Π^{k-1} of previously discovered optimal dual vectors. In other words, we recursively update $\Pi^k \leftarrow \Pi^{k-1} \cup \{\pi(x^k, \omega^k)\}$. For all other observations ($\omega \in \Omega^k,\ \omega \neq \omega^k$), we identify a dual vector in Π^k that provides the best lower bounding approximation at $Q(x^k, \omega)$ using the following operation (Steps 12–13):

$$\pi(x^k, \omega) \in \arg\max \ \{\pi^\top [r(\omega) - T(\omega)x^k] \mid \pi \in \Pi^k\}. \tag{21}$$

Note that the calculations in (21) are carried out only for previous observations as $\pi(x^k, \omega^k)$ provides the best lower bound at $Q(x^k, \omega^k)$. Further, notice that

$$\pi(x^k, \omega)^\top [r(\omega) - T(\omega)x^k] = Q^k(x^k, \omega),$$

the approximate recourse function value at x^k defined in (18), for all $\omega \in \Omega^k$, and $Q^k(x^k, \omega^k) = Q(x^k, \omega^k)$.

Using $\{Q^k(x^k, \omega^j)\}_{j=1}^k$, we solve a *distribution separation problem* (in Step 15):

$$\mathbb{Q}^k(x^k) = \max \ \left\{ \sum_{\omega \in \Omega^k} p(\omega) Q^k(x^k, \omega) \mid p(\omega) \in \widehat{\mathfrak{P}}^k \right\}. \tag{22}$$

Let $P^k = (p^k(\omega))_{\omega \in \Omega^k}$ denote an optimal solution of the above problem which we identify as a maximal/extremal probability distribution. Since the problem is solved over measures $\mathcal{M}^k$ that are defined only over the observed set Ω^k, the maximal probability distribution has weights $p^k(\omega)$ for $\omega \in \Omega^k$, and $p^k(\omega) = 0$ for $\omega \in \Omega \setminus \Omega^k$. Notice that the problem in (2) differs from the distribution separation problem (22) as the latter uses the recourse function approximation $Q^k(\cdot)$ and approximate ambiguity set $\widehat{\mathfrak{P}}^k$ as opposed to the true recourse function $Q(\cdot)$ and ambiguity set $\mathfrak{P}$, respectively. For the moment-based and Wasserstein distance-based ambiguity sets (discussed in Section 2.1), the distribution separation problem is a deterministic linear program. In general, the distribution separation problems associated with well-known ambiguity sets remain deterministic convex optimization problems [33], and off-the-shelf solvers can be used to obtain the extremal distribution.

In Step 16 of Algorithm 1, we use the dual vectors $\{\pi(x^k, \omega^j)\}_{j \leq k}$ and the maximal probability distribution P^k to generate a lower bounding affine function:

$$\mathbb{Q}^k(x) = \max_{P \in \widehat{\mathfrak{P}}^k} \mathbb{E}_P[Q^k(x, \tilde{\omega})] \geq \sum_{\omega \in \Omega^k} p^k(\omega) \cdot (\pi(x^k, \omega))^\top [r(\omega) - T(\omega)x], \tag{23}$$

for the worst-case expected recourse function measured with respect to the maximal probability distribution $P^k \in \widehat{\mathfrak{P}}^k$. We denote the coefficients of the affine function on the right-hand side of (23) by

$$\alpha_k^k = \sum_{\omega \in \Omega^k} p^k(\omega) \pi(x^k, \omega)^\top r(\omega) \text{ and } \beta_k^k = -\sum_{\omega \in \Omega^k} p^k(\omega) T(\omega)^\top \pi(x^k, \omega), \tag{24}$$

and succinctly write the affine function as $\ell_k^k(x) = \alpha_k^k + (\beta_k^k)^\top x$. Similar calculations are carried out using the incumbent solution $\hat{x}^{k-1}$ to identify a maximal probability distribution and a lower bounding affine function resulting in the affine function $\hat{\ell}_k^k(x) = \hat{\alpha}_k^k + (\hat{\beta}_k^k)^\top x$ (Step 17). Note that we use two indices for the cut coefficients (α, β) and the affine function ℓ. The superscript indicates the current iteration, while the subscript indicates the iteration when the quantities were first computed. Since, one observation is added to Ω^k in every iteration, the subscript also indicates the number of observations used in computing the quantities.

While the latest affine functions provide a lower bound for $\mathbb{Q}^k$, the affine functions generated in previous iteration are not guaranteed to lower bound $\mathbb{Q}^k$. To see this, let us consider the moment-based approximate ambiguity sets $\{\widehat{\mathfrak{P}}^k_{\text{mom}}\}_{k\geq 1}$. Let $P^j_{\text{mom}} \in \widehat{\mathfrak{P}}^j_{\text{mom}}$ be the maximal distribution identified in an iteration $j < k$ which was used to compute the affine function $\ell^j_j(x)$. By assigning $p^j(\omega) = 0$ for all new observations encountered after iteration j, i.e., $\omega \in \Omega^k \setminus \Omega^j$, we can construct a probability distribution $\bar{P} = ((p^j(\omega))_{\omega\in\Omega^j}, (0)_{\omega\in\Omega^k\setminus\Omega^j}) \in \mathbb{R}_+^{|\Omega^k|}$. This reconstructed distribution satisfies $\sum_{\omega\in\Omega^k} \bar{p}(\omega) = 1$. However, it is easy to see that $\sum_{\omega\in\Omega^k} \psi_i(\omega)\bar{p}(\omega) = \hat{b}^j_i \neq \hat{b}^k_i$ for all $i = 1, \ldots, q$. Therefore, $\bar{P} \notin \mathfrak{P}^k$. In other words, while the coefficients (α^j_j, β^j_j) are $\mathcal{F}^j$-measurable, the corresponding measure is not feasible to the approximate ambiguity set $\mathfrak{P}^k$. Therefore, $\ell^j_j(x)$ is not a valid lower bound to $\mathbb{Q}^k$. The arguments for the Wasserstein-based approximate ambiguity set are more involved, but persistence of a similar issue can be demonstrated.

To address this, we recursively update the previously generated affine functions $\ell^{k-1}_j(x) = \alpha^{k-1}_j + (\beta^{k-1}_j)^\top x$ for $j < k$ as follows (Steps 18 - 21):

$$\alpha^k_j = \theta^k \alpha^{k-1}_j, \quad \beta^k_j = \theta^k \beta^{k-1}_j, \text{ and } \ell^k_j(x) = \alpha^k_j + (\beta^k_j)^\top x \qquad \text{for all } j < k, \tag{25}$$

such that $\ell^k_j(x)$ provides lower bound approximation of $\mathbb{Q}^k(x)$ for all $j \in \{1, \ldots, k-1\}$. Similarly, we update the affine functions $\hat{\ell}^k_j(x)$, $j < k$, associated with incumbent solution. The candidate and the incumbent affine functions ($\ell^k_k(x)$ and $\hat{\ell}^k_k(x)$, respectively), as well as the updated collection of previously generated affine functions are used to build the set of affine functions which we denote by $\mathcal{L}^k$ (Step 22). Using this collection of affine functions $\mathcal{L}^k$, we update the approximation of the first-stage objective function in Step 23, as follows:

$$f^k(x) = c^\top x + \max_{\ell\in\mathcal{L}^k} \{\ell(x)\}. \tag{26}$$

The lower bounding property of this first-stage objective function approximation is captured in the following result.

Theorem 3.1. *Under assumption (A2), the first-stage objective function approximation in* (26) *satisfies*

$$f^k(x) \leq c^\top x + \mathbb{Q}^k(x) \text{ for all } x \in \mathcal{X} \text{ and } k \geq 1.$$

Proof. For the non-empty approximate ambiguity set $\widehat{\mathfrak{P}}^1$ of ambiguity set $\mathfrak{P}$, the construction of the affine function ensures that $\ell^1_1(x) \leq \mathbb{Q}^1(x)$. Now assume that $\ell(x) \leq \mathbb{Q}^{k-1}(x)$ for all $\ell \in \mathcal{L}^{k-1}$ and $k > 1$. The maximal nature of the probability distribution P^k satisfies:

$$\sum_{\omega\in\Omega^k} p^k(\omega)Q^k(x,\omega) \geq \sum_{\omega\in\Omega^k} p(\omega)Q^k(x,\omega) \qquad \forall P \in \widehat{\mathfrak{P}}^k.$$

Using above and the monotone property of the approximate recourse function, we have

$$\begin{aligned}\sum_{\omega\in\Omega^k} p^k(\omega)Q^k(x,\omega) &\geq \sum_{\omega\in\Omega^k} p(\omega)Q^{k-1}(x,\omega) \\ &= \sum_{\omega\in\Omega^k\setminus\{\omega^k\}} p(\omega)Q^{k-1}(x,\omega) + p(\omega^k)Q^{k-1}(x,\omega^k),\end{aligned} \tag{27}$$

for all $\{p(\omega)\}_{\omega\in\Omega^k} \in \widehat{\mathfrak{P}}^k$. Based on the properties of $\mathfrak{P}$ and $\{\widehat{\mathfrak{P}}^k\}_{k\geq 1}$ (similar to Propositions 2.1 and 2.2), we know that for every $P \in \widehat{\mathfrak{P}}^{k-1}$ we can construct a probability distribution in $\mathfrak{P}^k$ using the mapping Θ^k defined by (9). Considering a probability distribution $P' = \{p'(\omega)\}_{\omega\in\Omega^{k-1}} \in \widehat{\mathfrak{P}}^{k-1}$ we have

$\Theta^k(P') \in \widehat{\mathfrak{P}}^k$ and the inequality (27) reduces to

$$\begin{aligned}\sum_{\omega\in\Omega^k} p^k(\omega)Q^k(x,\omega) &\geq \sum_{\omega\in\Omega^k\setminus\{\omega^k\}} [\theta^k p'(\omega)Q^{k-1}(x,\omega)] + [\theta^k p'(\omega^k) + (1-\theta^k)]Q^{k-1}(x,\omega^k)\\ &= \theta^k\Bigg[\sum_{\omega\in\Omega^{k-1}} p'(\omega)Q^{k-1}(x,\omega)\Bigg] + (1-\theta^k)Q^{k-1}(x,\omega^k)\\ &\geq \theta^k\Bigg[\sum_{\omega\in\Omega^{k-1}} p'(\omega)Q^{k-1}(x,\omega)\Bigg].\end{aligned}$$

The last inequality is due to assumption $(A2)$, i.e., $Q(x,\omega^k) \geq 0$ and the construction of recourse function approximation Q^k described in §2.2. Since $\ell(x)$ lower bounds the term in bracket, we have

$$\sum_{\omega\in\Omega^k} p^k(\omega)Q^k(x,\omega) \geq \theta^k\ell(x).$$

Using the same arguments for all $\ell \in \mathcal{L}^{k-1}$, and the fact that the $\ell_k^k(x)$ and $\hat{\ell}_k^k(x)$ are constructed as lower bounds to the $\mathbb{Q}^k$, we have $f^k(x) \leq c^\top x + \mathbb{Q}^k(x)$. This completes the proof by induction. □

Once the approximation (26) is updated, the performance of the candidate solution is compared relative to the incumbent solution (Step 24). This comparison is performed by verifying if the following inequality is satisfied:

$$f^k(x^k) - f^k(\hat{x}^{k-1}) < \gamma[f^{k-1}(x^k) - f^{k-1}(\hat{x}^{k-1})], \tag{28}$$

where parameter $\gamma \in (0,1)$. If so, the candidate solution is designated to be the next incumbent solution, i.e., $\hat{x}^k = x^k$. If the inequality is not satisfied, the previous incumbent solution is retained as $\hat{x}^k = \hat{x}^{k-1}$. This completes a DRSD iteration.

Remark 3.1. We can extend the algorithm design for 2-DRLPs where the relatively complete recourse assumption of (A2) and/or assumption (A3) is not satisfied. For problems where relatively complete recourse condition is not met, a candidate solution may lead to one or more subproblems to be infeasible. In this case, the dual extreme rays can be used to compute a feasibility cut that is included in the first-stage approximation. The argmax procedure in (21) is only valid when assumption (A3) is satisfied. In problems where the uncertainty also affects the cost coefficients, the argmax procedure presented in [19] can be utilized. These algorithmic enhancements can be incorporated without affecting the convergence properties of DRSD.

Remark 3.2 (Relation between DRSD, SD, and DR L-shaped Method). We close this section by identifying the key differences in the DRSD algorithm design when compared to SD and DR L-shaped methods.

- There are two main differences between DRSD and the DR L-Shaped method. Firstly, the DR L-shaped method operates with a deterministic representation of the ambiguity set computed using a fixed sample of observations, an input to the algorithm. In contrast, a new observation is added (Line 7) in every iteration of DRSD to improve the approximation of the ambiguity set. Secondly, every iteration of the DR L-shaped method involves solving a subproblem corresponding to each observation used in the ambiguity set representation. On the other hand, in DRSD, only two subproblems corresponding to latest observation ω^k are solved to optimality, and the argmax procedure is used for the other observations.
- While DRSD is designed to address the 2-DRLP problem (1), the SD and its variants [23, 24, 40] are for risk-neutral 2-SLP. This generalization introduces another layer of approximation to SD, viz.,

the approximation of ambiguity sets. The algorithmic enhancements necessary to address this new layer of approximations make the DRSD significantly different from its risk-neutral predecessors. For instance, we need to solve an approximate distribution separation problem in every iteration (Line 15). The cut coefficients are computed and updated (in Lines 18-21) in a manner that is consistent with the updates carried out to approximate the ambiguity sets (see propositions 2.1 and 2.2, and coefficient updates in (25)). The cut updates that are undertaken in SD only need to be consistent with the updates in empirical distribution. This critical difference in cut computations also introduces significant differences in the convergence analysis of DRSD that we present next.

4 Convergence Analysis

In this section we provide the convergence result of the sequential sampling-based approach to solve DRO problems. In order to facilitate the exposition of our theoretical results, we will define certain quantities for notational convenience that are not necessarily computed during the course of the algorithm. Our convergence results are built upon stability analyses presented in [44] and convergence analysis of the SD algorithm in [23].

We define a function over the approximate ambiguity set using the recourse function $Q(\cdot,\cdot)$, that is

$$g^k(x) := c^\top x + \max_{P\in\widehat{\mathfrak{P}}^k} \mathbb{E}_P[Q(x,\tilde{\omega})]. \tag{29}$$

We begin by analyzing the behavior of the sequence $\{g^k\}_{k\geq 1}$ as $k\to\infty$. In particular, we will assess the sequence of function evaluations at a converging subsequence of first-stage solutions. The result is captured in the following proposition.

Proposition 4.1. *Suppose $\{\hat{x}^{k_n}\}$ denotes a subsequence of $\{\hat{x}^k\}$ such that $\hat{x}^{k_n}\to\bar{x}$, then $\lim_{n\to\infty}|g^{k_n}(\hat{x}^{k_n})-f(\bar{x})|=0$, with probability one.*

Proof. Consider an approximate ambiguity set $\widehat{\mathfrak{P}}^k$. For $i=1,2$ and $x_i\in\mathcal{X}$, let $P(x_i)\in\arg\max_{P\in\widehat{\mathfrak{P}}^k}\{\mathbb{E}_P[Q(x_i,\tilde{\omega})]\}$. Then,

$$\begin{aligned} g^k(x_1) =\ & c^\top x_1+\mathbb{E}_{P(x_1)}[Q(x_1,\tilde{\omega})] \geq\ c^\top x_1+\mathbb{E}_{P(x_2)}[Q(x_1,\tilde{\omega})] \\ =\ & c^\top x_2+\mathbb{E}_{P(x_2)}[Q(x_2,\tilde{\omega})]+c^\top(x_1-x_2)+ \\ & \qquad \mathbb{E}_{P(x_2)}[Q(x_1,\tilde{\omega})]-\mathbb{E}_{P(x_2)}[Q(x_2,\tilde{\omega})] \\ =\ & g^k(x_2)+c^\top(x_1-x_2)+\mathbb{E}_{P(x_2)}[Q(x_1,\tilde{\omega})]-\mathbb{E}_{P(x_2)}[Q(x_2,\tilde{\omega})]. \end{aligned}$$

The inequality in the above follows from optimality of $P(x_1)$. The above implies that

$$\begin{aligned} g^k(x_2)-g^k(x_1) \leq\ & c^\top(x_2-x_1)+\mathbb{E}_{P(x_2)}[Q(x_2,\tilde{\omega})]-\mathbb{E}_{P(x_2)}[Q(x_1,\tilde{\omega})] \\ \leq\ & |c^\top(x_2-x_1)|+\left|\mathbb{E}_{P(x_2)}[Q(x_2,\tilde{\omega})]-\mathbb{E}_{P(x_2)}[Q(x_1,\tilde{\omega})]\right| \end{aligned}$$

The second relationship is due to the triangular inequality. Under assumption (A2), the recourse function $Q(x,\tilde{\omega})$ is a uniformly Lipschitz continuous function, with probability one (see Chapter 2 in [42] for details). This implies that there exists a constant C such that $|\mathbb{E}_P[Q(x_1,\tilde{\omega})]-\mathbb{E}_P[Q(x_2,\tilde{\omega})]|\leq C\|x_1-x_2\|$ for any probability distribution P. As a result,

$$g^k(x_2)-g^k(x_1)\leq\ (\|c\|+C)\|x_2-x_1\|. \tag{30}$$

Starting with x_2 and using the same arguments, we have

$$g^k(x_1) - g^k(x_2) \leq (\|c\| + C)\|x_1 - x_2\|. \tag{31}$$

Therefore, the function $g^k(x)$ is equi-continuous on $x \in \mathcal{X}$. Now consider ambiguity sets $\mathfrak{P}$ and $\widehat{\mathfrak{P}}^k$. Note that for all $x \in \mathcal{X}$,

$$\begin{aligned} |f(x) - g^k(x)| &= \left| \max_{P \in \mathfrak{P}} \mathbb{E}_P[Q(x,\tilde{\omega})] - \max_{P' \in \widehat{\mathfrak{P}}^k} \mathbb{E}_{P'}[Q(x,\tilde{\omega})] \right| \\ &\leq \max_{P \in \mathfrak{P}} \min_{P' \in \widehat{\mathfrak{P}}^k} \left| \mathbb{E}_P[Q(x,\tilde{\omega})] - \mathbb{E}_{P'}[Q(x,\tilde{\omega})] \right| \\ &\leq \max_{P \in \mathfrak{P}} \min_{P' \in \widehat{\mathfrak{P}}^k} \sup_{x \in \mathcal{X}} \left| \mathbb{E}_P[Q(x,\tilde{\omega})] - \mathbb{E}_{P'}[Q(x,\tilde{\omega})] \right|. \end{aligned}$$

Using the definition of deviation (6) and Hausdorff distance (7) between ambiguity sets $\mathfrak{P}$ and $\widehat{\mathfrak{P}}^k$, we have

$$|f(x) - g^k(x)| \leq \mathbb{D}(\mathfrak{P}, \widehat{\mathfrak{P}}^k) \leq \mathbb{H}(\mathfrak{P}, \widehat{\mathfrak{P}}^k). \tag{32}$$

For $\hat{x}^{k_n}$ and $\bar{x}$, using the triangle inequality we have

$$\begin{aligned} |f(\bar{x}) - g^{k_n}(\hat{x}^{k_n})| &\leq |f(\bar{x}) - g^{k_n}(\bar{x})| + |g^{k_n}(\bar{x}) - g^{k_n}(\hat{x}^{k_n})| \\ &\leq \mathbb{H}(\mathfrak{P}, \widehat{\mathfrak{P}}^{k_n}) + (\|c\| + C)\|\bar{x} - \hat{x}^{k_n}\|. \end{aligned}$$

The second inequality is justified by combining (30), (31), and (32). As $n \to \infty$, $\mathbb{H}(\mathfrak{P}, \widehat{\mathfrak{P}}^{k_n}) \to 0$ due to property $(B2)$ of the considered family of ambiguity sets. Furthermore, since $\hat{x}^{k_n} \to \bar{x}$, the right-hand side of the above inequality vanishes. Therefore, we conclude that $g^{k_n}(\hat{x}^{k_n}) \to f(\bar{x})$ as $n \to \infty$. □

Notice that the behavior of the approximate ambiguity sets defined in §2.1, in particular, the condition $\mathbb{H}(\mathfrak{P}, \widehat{\mathfrak{P}}^k) \to 0$ as $k \to \infty$ plays a central role in the above proof. Recall that for the moment and Wasserstein distance-based ambiguity sets, the condition is established in propositions 2.1 and 2.2, respectively. It is also worthwhile to note that under the foregoing conditions, (32) also implies uniform convergence of the sequence $\{g^k\}$ to $f(x)$, with probability one.

The above result applies to any algorithm that generates a converging sequence of iterates $\{x^k\}$ and a corresponding sequence of extremal distributions. Such an algorithm is guaranteed to exhibit convergence to the optimal distributionally robust objective function value. Therefore, this result is applicable to the sequence of instances constructed using external sampling and solved, for example, using reformulation-based methods. Such an approach was adopted in [34] and [44]. The analysis in [34] relies upon two rather restrictive assumptions. The first assumption is that for all $P \in \mathfrak{P}$, there exists a sequence of measures $\{P^k\}$ such that $P^k \in \widehat{\mathfrak{P}}^k$ and converges weakly to P. The second assumption requires the approximate ambiguity sets to be strict subsets of the true ambiguity set, i.e., $\widehat{\mathfrak{P}}^k \subset \mathfrak{P}$. Both of these assumptions are very difficult to satisfy in a data-driven setting (also see Remark 2.1).

The analysis in [44], on the other hand, does not make the above assumptions. Therefore, their analysis is more broadly applicable in settings where external sampling is used to generate Ω^k. DRO instances are constructed based on statistics estimated using Ω^k and solved to optimality for each $k \geq 1$. They show the convergence of optimal objective function values and optimal solution sets of approximate problems to the optimal objective function value and solutions of the true DRO problem, respectively. In this regard, the result in Proposition 4.1 can alternatively be derived using Theorem 1(i) in [44]. While the above function is not computed during the course of the sequential sampling algorithm, it provides

the necessary benchmark for our convergence analysis.

One of the main point of deviation in our analysis stems from the fact that we use the objective function approximations that are built based on the approximate recourse function in (18). In order to study the piecewise affine approximation of the first-stage objective function, we introduce another benchmark function

$$\phi^k(x) := c^\top x + \max_{P\in\widehat{\mathfrak{P}}^k} \mathbb{E}_P[Q^k(x,\tilde{\omega})]. \tag{33}$$

Notice that the above function uses the approximations for the ambiguity set (as in the case of (29)) as well as the approximation of the recourse function. This construction ensures that $\phi^k(x) \leq g^k(x)$ for all $x \in \mathcal{X}$ and $k \geq 1$, which follows from the fact that $Q^k(x,\tilde{\omega}) \leq Q(x,\tilde{\omega})$, almost surely. Further, the result in Theorem 3.1 ensures that $f^k(x) \leq \phi^k(x)$. Putting these together, we obtain the following relationship:

$$f^k(x) \leq \phi^k(x) \leq g^k(x) \qquad \forall x \in \mathcal{X}, k \geq 1. \tag{34}$$

While the previous proposition was focused on the upper limit in the above relationship, we present the asymptotic behavior of the $\{f^k\}$ sequence in the following results.

Lemma 4.2. *Suppose $\{\hat{x}^{k_n}\}$ denotes a subsequence of $\{\hat{x}^k\}$ such that $\hat{x}^{k_n} \to \bar{x}$. Then, $\lim_{n\to\infty} f^{k_n}(\hat{x}^{k_n}) - f(\bar{x}) = 0$, with probability one.*

Proof. From Proposition 4.1, we have $\lim_{n\to\infty} |f(\bar{x}) - g^{k_n}(\hat{x}^{k_n})| = 0$. Therefore, there exists $N_1 < \infty$ and $\epsilon_1 > 0$ such that

$$\left| \max_{P\in\mathfrak{P}} \mathbb{E}_P[Q(\bar{x},\tilde{\omega})] - \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[Q(\hat{x}^{k_n},\tilde{\omega})] \right| < \epsilon_1/2 \quad \forall n > N_1. \tag{35}$$

Now consider,

$$\begin{aligned} \left| \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[Q(\hat{x}^{k_n},\tilde{\omega})] - \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[Q^{k_n}(\hat{x}^{k_n},\tilde{\omega})] \right| & \\ \leq \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \left| \mathbb{E}_P[Q(\hat{x}^{k_n},\tilde{\omega})] - \mathbb{E}_P[Q^{k_n}(\hat{x}^{k_n},\tilde{\omega})] \right| & \\ = \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[|Q(\hat{x}^{k_n},\tilde{\omega}) - Q^{k_n}(\hat{x}^{k_n},\tilde{\omega})|]. & \end{aligned}$$

The last equality follows from the fact that $Q(x,\tilde{\omega}) \geq Q^k(x,\tilde{\omega})$ for all $x \in \mathcal{X}$ and $k \geq 1$, almost surely. Moreover, because of the uniform convergence of $\{Q^k\}$ (Proposition 2.3), the sequence of approximate functions $\{\phi^k\}$ also convergences uniformly. This implies that, there exists $N_2 < \infty$ such that for all $n > N_2$,

$$\left| \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[Q(\hat{x}^{k_n},\tilde{\omega})] - \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[Q^{k_n}(\hat{x}^{k_n},\tilde{\omega})] \right| < \epsilon_1/2. \tag{36}$$

Let $N = \max\{N_1, N_2\}$. Using (35) and (36), we have for all $n > N$,

$$\left| \max_{P\in\mathfrak{P}} \mathbb{E}_P[Q(\bar{x},\tilde{\omega})] - \max_{P\in\widehat{\mathfrak{P}}^{k_n}} \mathbb{E}_P[Q^{k_n}(x^{k_n},\tilde{\omega})] \right| < \epsilon_1.$$

This implies that $|f(\bar{x}) - \phi^{k_n}(\hat{x}^{k_n})| \to 0$ as $n \to \infty$. Based on (21), we have $Q^{k_n}(\hat{x}^{k_n},\omega) =$

$(\pi(\hat{x}^{k_n},\omega))^\top[r(\omega)-T(\omega)\hat{x}^{k_n}] \geq (\pi(\hat{x}^{k_n},\omega))^\top[r(\omega)-T(\omega)x]$ for all $x \in \mathcal{X}$ and $\omega \in \Omega^{k_n}$. Let

$$\alpha_{k_n}^{k_n} = \sum_{\omega\in\Omega^{k_n}} p^{k_n}(\omega)(\pi(\hat{x}^{k_n},\omega))^\top r(\omega) \text{ and } \beta_{k_n}^{k_n} = -\sum_{\omega\in\Omega^{k_n}} p^{k_n}(\omega)T(\omega)^\top\pi(\hat{x}^{k_n},\omega),$$

where $\{p^{k_n}(\omega)\}_{\omega\in\Omega^{k_n}}$ is an optimal solution of the distributional separation problem (22) where index k is replaced by k_n. Then, the affine function $\alpha_{k_n}^{k_n} + (c+\beta_{k_n}^{k_n})^\top x$ provides a lower bound approximation for function $\phi^{k_n}(x)$, i.e.,

$$\phi^{k_n}(x) \geq \alpha_{k_n}^{k_n} + (c+\beta_{k_n}^{k_n})^\top x \qquad \text{for all } x \in \mathcal{X},$$

with strict equality holding only at $\hat{x}^{k_n}$. Therefore, using the definition of $f^k(x)$ we have $\lim_{n\to\infty}\alpha_{k_n}^{k_n} + (c+\beta_{k_n}^{k_n})^\top\hat{x}^{k_n} = \lim_{n\to\infty} f^{k_n}(\hat{x}^{k_n}) = \lim_{n\to\infty}\phi^{k_n}(\hat{x}^{k_n}) = f(\bar{x})$, almost surely. This completes the proof. □

The above result characterizes the behavior of the sequence of affine functions generated during the course of the algorithm. In particular, the sequence $\{f^k(\hat{x}^k)\}_{k\geq 1}$ accumulates at the objective value of the original DRO problem (1). Recall that the candidate solution x^k is a minimizer of $f^{k-1}(x)$ and an affine function is generated at this point such that $f^k(x^k) = \phi^k(x^k)$ in all iterations $k \geq 1$. However, due to the update procedure in (25) the quality of the estimates at x^k gradually diminishes leading to a large value for $(\phi^k(x^k) - f^k(x^k))$ as k increases. This emphasizes the role of the incumbent solution and computing the incumbent affine function $\hat{\ell}(x)$ during the course of the algorithm. By updating the incumbent solution and frequently reevaluating the affine functions at the incumbent solution, we can ensure that the approximation is "sufficiently good" in the neighborhood of the incumbent solution. In order to assess the improvement of approximation quality, we define

$$\delta^k := f^{k-1}(x^k) - f^{k-1}(\hat{x}^{k-1}) \leq 0 \qquad \forall k \geq 1. \tag{37}$$

The inequality follows from the optimality of x^k with respect to the objective function f^{k-1}. The quantity δ^k measures the error in objective function estimate at the candidate solution with respect to the estimate at the current incumbent solution. The following result captures the asymptotic behavior of this error term.

Lemma 4.3. *Let $\mathcal{K}$ denote a sequence of iterations where the incumbent solution changes. There exists a subsequence of iterations, denoted as $\mathcal{K}^* \subseteq \mathcal{K}$, such that* $\lim_{k\in\mathcal{K}^*}\delta^k = 0$.

Proof. We will consider two cases depending on whether the set $\mathcal{K}$ is finite or not. First, suppose that $|\mathcal{K}|$ is not finite. By the incumbent update rule and (37),

$$f^{k_n}(x^{k_n}) - f^{k_n}(\hat{x}^{k_n-1}) < \gamma[f^{k_n-1}(x^{k_n}) - f^{k_n-1}(\hat{x}^{k_n-1})] = \gamma\delta^{k_n} \leq 0 \qquad \forall k_n \in \mathcal{K}.$$

Subsequently, we have $\limsup_{n\to\infty}\delta^{k_n} \leq 0$. Since $x^{k_n} = \hat{x}^{k_n}$ and $\hat{x}^{k_n-1} = \hat{x}^{k_{n-1}}$, we have

$$f^{k_n}(\hat{x}^{k_n}) - f^{k_n}(\hat{x}^{k_{n-1}}) \leq \gamma\delta^{k_n} \leq 0.$$

The left-hand side of the above inequality captures the improvement in the objective function value at the current incumbent solution over the previous incumbent solution. Using the above, we can write the average improvement attained over n incumbent changes as

$$\frac{1}{n}\sum_{j=1}^{n}\left[f^{k_j}(\hat{x}^{k_j}) - f^{k_j}(\hat{x}^{k_{j-1}})\right] \leq \frac{1}{n}\sum_{j=1}^{n}\gamma\delta^{k_j} \leq 0 \qquad \text{for all } n.$$

This implies that

$$\frac{1}{n}\underbrace{\left(f^{k_n}(\hat{x}^{k_n}) - f^{k_1}(\hat{x}^{k_0})\right)}_{(a)} + \frac{1}{n}\left[\sum_{j=1}^{n-1}\underbrace{\left(f^{k_j}(\hat{x}^{k_j}) - f^{k_{j+1}}(\hat{x}^{k_j})\right)}_{(b)}\right] \leq \frac{1}{n}\sum_{j=1}^{n}\gamma\delta^{k_j} \leq 0,$$

for all n. Under the assumption that the dual feasible region is non-empty and bounded (this is ensured by relatively complete recourse, (A2)), $\{f^k\}$ is a sequence of Lipschitz continuous functions. This, along with compactness of $\mathcal{X}$ (A1), implies that $f^{k_n}(\hat{x}^{k_n}) - f^{k_1}(\hat{x}^{k_0})$ is bounded from above. Hence, the term (a) reduces to zero as $n \to \infty$. The term (b) converges to zero, with probability one, due to uniform convergence of $\{f^k\}$. Since $\gamma \in (0,1)$, we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\delta^{k_j} = 0,$$

with probability one. Further,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\delta^{k_j} \leq \limsup_{n\to\infty}\delta^{k_n} \leq 0.$$

Thus, there exists a subsequence indexed by the set $\mathcal{K}^*$ such that $\lim_{k\in\mathcal{K}^*}\delta^k = 0$, with probability one.

Now if $|\mathcal{K}|$ is finite, then there exists $\hat{x}$ and $K < \infty$ such that for all $k \geq K$, we have $\hat{x}^k = \hat{x}$. Notice that, if $\lim_{k\in\mathcal{K}^*} x^k = \bar{x}$, uniform convergence of the sequence $\{f^k\}$ and Lemma 4.2 ensure that

$$\lim_{k\in\mathcal{K}^*} f^k(x^k) = \lim_{k\in\mathcal{K}^*} f^{k-1}(x^k) = f(\bar{x}) \tag{38a}$$

$$\lim_{k\in\mathcal{K}^*} f^k(\hat{x}) = \lim_{k\in\mathcal{K}^*} f^{k-1}(\hat{x}) = f(\hat{x}). \tag{38b}$$

Further, since the incumbent is not updated in iterations $k \geq K$, we must have from the update rule in (28) that

$$f^k(x^k) - f^k(\hat{x}) \geq \gamma[f^{k-1}(x^k) - f^{k-1}(\hat{x})] = \gamma\delta^k \quad \text{for all } k \geq K.$$

Using (38), we have

$$\begin{aligned}\lim_{k\in\mathcal{K}^*}\left(f^k(x^k) - f^k(\hat{x})\right) &\geq \gamma \lim_{k\in\mathcal{K}^*}\left(f^{k-1}(x^k) - f^{k-1}(\hat{x})\right),\\ \Rightarrow \qquad f(\bar{x}) - f(\hat{x}) &\geq \gamma(f(\bar{x}) - f(\hat{x})).\end{aligned}$$

Noting that $\gamma \in (0,1)$, the above inequality reduces to $f(\bar{x}) - f(\hat{x}) \geq 0$. Further, using (37) in the limit as $k \to \infty$ and the fact that $\hat{x}^k = \hat{x}$ for all $k \geq K$, we have $f(\bar{x}) - f(\hat{x}) \leq 0$. Therefore, we have $f(\bar{x}) - f(\hat{x}) = 0$. Hence, $\lim_{k\in\mathcal{K}^*}\delta^k = f(\bar{x}) - f(\hat{x}) = 0$, with probability one. □

Equipped with the results in lemmas 4.2 and 4.3, we state the main theorem which establishes the existence of a subsequence of the incumbent sequence generated by the algorithm for which every accumulation point is an optimal solution to (1).

Theorem 4.4. *Let $\{x^k\}_{k=1}^{\infty}$ and $\{\hat{x}^k\}_{k=1}^{\infty}$ be the sequence candidate and incumbent solutions generated by the DRSD algorithm. There exists a subsequence $\{\hat{x}^k\}_{k\in\mathcal{K}}$ for which every accumulation point is an optimal solution of 2-DRLP* (1)*, with probability one.*

Proof. Let $x^* \in \mathcal{X}$ be an optimal solution of (1). Consider a subsequence indexed by $\mathcal{K}$ such that $\lim_{k\in\mathcal{K}} \hat{x}^k = \bar{x}$. Compactness of $\mathcal{X}$ ensures the existence of accumulation point $\bar{x} \in \mathcal{X}$ and therefore,

$$f(x^*) \leq f(\bar{x}). \tag{39}$$

From Theorem 3.1, we have for all $k, x \in \mathcal{X}$

$$f^k(x) \leq \ c^\top x + \mathbb{Q}^k(x) \leq \ c^\top x + \max_{P\in\widehat{\mathfrak{P}}^k} \mathbb{E}_P[Q(x,\tilde{\omega})] = g^k(x).$$

Thus, using the uniform convergence of $\{g^k\}$ (Proposition 4.1) we have

$$\limsup_{k\in\mathcal{K}'} f^k(x^*) \leq \lim_{k\in\mathcal{K}'} g^k(x^*) = f(x^*) \tag{40}$$

for all subsequences indexed by $\mathcal{K}' \subseteq \{1, 2, \ldots\}$, with probability one. Recall that,

$$\delta^k = f^{k-1}(x^k) - f^{k-1}(\hat{x}^{k-1}) \leq f^{k-1}(x^*) - f^{k-1}(\hat{x}^{k-1}) \qquad \text{for all } k \geq 1.$$

The inequality in the above follows from the optimality of x^k with respect to $f^{k-1}(x)$. Taking limit over $\mathcal{K}$, we have

$$\begin{aligned}\lim_{k\in\mathcal{K}} \delta^k \leq & \lim_{k\in\mathcal{K}} \left(f^{k-1}(x^*) - f^{k-1}(\hat{x}^{k-1})\right) \\ \leq & \limsup_{k\in\mathcal{K}} f^{k-1}(x^*) - \liminf_{k\in\mathcal{K}} f^{k-1}(\hat{x}^{k-1}) \leq \ f(x^*) - f(\bar{x}).\end{aligned}$$

The last inequality follows from (40) and $\lim_{k\in\mathcal{K}} f^{k-1}(\hat{x}^{k-1}) = f(\bar{x})$ (Lemma 4.2). From Lemma 4.3, there exists a subsequence indexed by $\mathcal{K}^* \subseteq \mathcal{K}$ such that $\lim_{k\in\mathcal{K}^*} \delta^k = 0$. Therefore, if $\{\hat{x}^k\}_{k\in\mathcal{K}^*} \to \bar{x}$, we have

$$f(x^*) - f(\bar{x}) \geq 0.$$

Using (39) and the above inequality, we conclude that $\bar{x}$ is an optimal solution with probability one. □

5 Computational Experiment

In this section, we evaluate the effectiveness and efficiency of the DRSD method in solving 2-DRLPs. For our preliminary experiments, we consider 2-DRLPs with moment-based ambiguity set $\mathfrak{P}_{\text{mom}}$ for the first two moments ($q = 2$). We also consider 2-DRLPs with Wasserstein ambiguity set $\mathfrak{P}_{\text{w}}$ with ℓ_1 and ℓ_∞ distance metrics.

We report results from the computational experiments conducted on four well-known SP test problems: the capacity expansion planning (`CEP`) [25], the power generation planning (`PGP`) [25], multilocation transshipment (`RETAIL`) [22], and cargo flight scheduling (`STORM`) [32]. In Table 5, we provide the number of variables (#Var) and constraints (#Cons) in the first- and second-stage of the test problems. Notice that the `PGP` and `CEP` have relatively smaller supports (216 and 576, respectively), while `RETAIL` and `STORM` have a support size of 10^{11} and 10^{81}, respectively. In the table, we also provide computational results from solving the risk-neutral versions of these problems using the SD algorithm [40]. For these results, we report the number of iterations (#Iter), objective function estimate (ObjEst) at termination, and total time (in seconds) taken by the SD algorithm. We refer the readers interested in a computational comparison between SD and an external sampling-based approach for risk-neutral 2-SLPs to [40]

Table 1: Details of `CEP`, `PGP`, `RETAIL`, and `STORM` Test Problems, and Computational Results for the SD Algorithm

Problem	Stage I		Stage II				SD Results		
	#Var	#Cons	#Var	#Cons	#RV	$\lvert\Omega\rvert$	#Iter	ObjEst	Time
`PGP`	4	2	16	7	3	576	215($\pm$8)	446($\pm$2.4)	0.43($\pm$0.04)
`CEP`	8	5	15	7	3	216	153($\pm$7)	343886($\pm$12783)	0.18($\pm$0.02)
`RETAIL`	7	0	70	22	7	10^{11}	721($\pm$44)	154($\pm$1.92)	4.20($\pm$0.76)
`STORM`	121	185	1259	528	117	10^{81}	238($\pm$17)	15173494($\pm$657272)	2.83($\pm$0.21)

and [19].

Following the rule of thumb adopted in experiments involving sampling-based SP, we conduct 30 independent replications for each problem instance. The choice of 30 replications is the same as in previous experiments with SD (see [19] and [40], for example). Each replication uses a different seed for the random number generator. The algorithms are implemented in the C programming language, and the experiments are conducted on a 64-bit Intel core i7 - 4770 CPU at 3.4GHz $\times$ 8 machine with 32 GB memory. All linear programs, i.e., master problem, subproblems, and distribution separation problem, are solved using CPLEX 12.10 callable subroutines. For DRSD, we use $\tau = 0.001$ and $\gamma = 0.2$ in our experiments. We add one new observation to the sample in every iteration and therefore, $\theta^k = \frac{k-1}{k}$ is used for the updates in (25). The source code for the DR L-shaped, DRSD algorithms, and the reformulation techniques are available under the GNU general public license at `https://github.com/SMU-SODA/distributionallyRobust.git`. The repository also includes the test problems in SMPS file format.

5.1 Results for 2-DRLPs with Moment-based Ambiguity Set

The first set of experiments concerns the 2-DRLP problems with a moment-based ambiguity set $\mathfrak{P}_{\text{mom}}$ for which we use an external sampling-based approach as a benchmark for comparison with DRSD. The external sampling-based instances involve constructing approximate problems of the form (29) with a pre-determined number of observations $N \in \{100, 250, 500, 1000\}$. The resulting instances are solved using the DR L-Shaped method. For a fair comparison, the DRSD method is run for a maximum of N iterations to have an estimate based upon a sample of size not greater than N. Specifically, it terminates when conditions in Step 5 of Algorithm 1 are satisfied. We compare the solution quality provided by these methods along with the computational time. The results from the experiments are presented in Table 5.1.

Table 5.1 shows the number of iterations, objective function estimate $f^k(\hat{x}^k)$ at termination, and solution time (in seconds) averaged over 30 replications. The values in the parenthesis are the half-widths of the corresponding confidence intervals. Similar to SD, the number of iterations for DRSD is also equal to the number of observations used to approximate the ambiguity set. To begin, observe the increase in the objective function estimates of distributionally robust variants when compared to the risk-neutral results from SD in Table 5.

The objective function estimate obtained using the DRSD is comparable to the objective function estimate obtained using the DR L-shaped method. Notice that for instances with $N = 1000$, DRSD took less than 1000 iterations because the termination conditions were satisfied. The same is true for `CEP` instances with $N = 500$. This shows the potential ability of DRSD to dynamically determine the number of observations by assessing the progress made during the algorithm. For instance, the DRSD

Table 2: Computational Results for 2-DRLP Instances with Moment-based Ambiguity Set

N	DRSD Algorithm			DR L-Shaped Algorithm		
	#Iter	ObjEst	Time	#Iter	ObjEst	Time
			PGP			
100	100 (±0)	460.89 (±3.76)	0.04 (±0.00)	18 (±0.9)	457.61 (±3.28)	0.052 (±0.00)
250	250 (±0)	466.91 (±2.52)	0.13 (±0.00)	20 (±0.7)	462.92 (±2.28)	0.077 (±0.00)
500	500 (±0)	471.40 (±3.49)	0.32 (±0.00)	20 (±0.6)	464.70 (±1.95)	0.096 (±0.00)
1000	504 (±687)	463.19 (±16.28)	0.35 (±0.70)	20 (±0.8)	466.10 (±1.78)	0.121 (±0.00)
			CEP			
100	100 (±0)	658831 (±14453)	0.04 (±0.00)	3 (±0.2)	658817 (±14457)	0.015 (±0.00)
250	250 (±0)	680795 (±10524)	0.12 (±0.00)	2 (±0.2)	680736 (±10511)	0.024 (±0.00)
500	256 (±0)	683300 (±5955)	0.30 (±0.00)	20 (±0.6)	683252 (±5949)	0.028 (±0.00)
1000	256 (±0)	683300 (±5955)	0.30 (±0.00)	2 (±0)	679665 (±4926)	0.028 (±0.00)
			RETAIL			
100	100 (±0)	326.21 (±15.35)	0.07 (±0.00)	46 (±1)	327.26 (±14.79)	0.370 (±0.01)
250	250 (±0)	365.00 (±19.03)	0.27 (±0.01)	45 (±2)	365.54 (±19.45)	0.839 (±0.03)
500	500 (±0)	387.98 (±17.10)	0.86 (±0.02)	45 (±1)	388.84 (±17.48)	1.587 (±0.05)
1000	625 (±31)	396.67 (±15.99)	1.13 (±0.12)	45 (±1)	401.71 (±14.38)	3.176 (±0.09)
			STORM			
100	100 (±0)	15755337 (±12314)	0.74 (±0.02)	12 (±0.51)	15742456 (±12192)	0.434 (±0.02)
250	250 (±0)	15795815 (±8493)	4.66 (±0.13)	11 (±0.52)	15781725 (±8754)	1.008 (±0.05)
500	500 (±0)	15811923 (±5233)	20.54 (±0.51)	12 (±0.59)	15797020 (±5346)	2.117 (±0.10)
1000	516 (±108)	15786865 (±9155)	30.44 (±15.28)	12 (±0.52)	15806575 (±3772)	4.318 (±0.19)

objective function estimate for STORM that is based upon a sample of size 516 (on average) is within 0.1% and 0.12% of the objective function value estimate provided by the DR L-shaped method for $N = 500$ and $N = 1000$, respectively. These results show that the optimal objective function estimate obtained from DRSD are comparable to those obtained using an external sampling-based approach.

The results for small scale instances (PGP and CEP) show that both DRSD and the DR L-shaped method take a fraction of a second, but the computational time for DRSD is higher than the DR L-shaped method for all N. We attribute this behavior to two reasons. (i) The computational effort to solve all the subproblems in each iteration does not increase significantly with N as they are easy to solve. This observation is in-line with our computational experience with the SD method for 2-SLPs (see [19]). (ii) The DRSD takes a larger number of iterations, resulting in an increased number of master and distribution separation problems solved. It is important to note that, while the computational time for the DR L-shaped method on an individual instance may be lower, the iterative procedure necessary to identify a sufficient sample size may require solving several instances with increasing sample size. This may result in a significantly higher cumulative computational time.

On the other hand, for large-scale problems (RETAIL and STORM), we observe a noticeable increase in the computational time for the DR L-shaped method with an increase in N. A significant portion of this time is spent solving the subproblems. Since the DRSD solves only two subproblems in each iteration, the time taken to solve the subproblems is significantly less in comparison to the DR L-shaped method where all subproblems corresponding to unique observations are solved in each iteration. Notice that for RETAIL, the average number of iterations taken by DRSD is at least 8.2 times the average number of iterations taken by DR L-shaped for any N. This increases the computational time spent for solving

the master and distributional separation problems. However, the reduction in the overall computational time is a direct consequence of solving only two subproblems in each iteration. The results for `STORM` also show similar behavior in terms of computational time associated with solving master and subproblems. However, the overall increase in the computational time is due to a significant computational expense ($\sim$ 78%) in naively solving the distribution separation problem. This computational time associated with solving the distribution separation problem can be reduced by using column-generation procedures that take advantage of the problem structure. Such an implementation is not undertaken for our current experiments and is a fruitful future research avenue.

5.2 Results for 2-DRLPs with ℓ_1-type Wasserstein Ambiguity Set

For the Wasserstein distance-based ambiguity sets, we benchmark against the reformation techniques proposed by [48]. Specifically, in [48], it has been shown that a 2-DRLP (1) with Wasserstein ambiguity set can be reformulated as a two-stage robust optimization problem. This reformulation is given by

$$\min_{x\in X,\eta\geq 0}\left\{c^\top x+\eta\epsilon+\frac{1}{N_s}\sum_{n=1}^{N_s}\max_{\omega\in\Omega}\left\{Q(x,\omega)-\eta||\omega-\overline{\omega}_n||\right\}\right\}, \tag{41}$$

where $\{\overline{\omega}_1,\ldots,\overline{\omega}_{N_s}\}$ is a finite set of observations obtained using true distribution. Notice that the reformulation (41) can be written as the following semi-infinite program:

$$\begin{aligned}\min_{x\in X,\eta\geq 0}\ & c^\top x+\eta\epsilon+\frac{1}{N_s}\sum_{n=1}^{N_s}\nu_n:\\ &\text{s.t. } Q(x,\omega)-\eta||\omega-\overline{\omega}_n||\leq\nu_n,\quad n\in\{1,\ldots,N_s\},\omega\in\Omega.\end{aligned}$$

For problem instances with ℓ_1-type Wasserstein ambiguity set, we solve the foregoing program using a Benders decomposition approach.

For ℓ_1-norm, the reformulation in (41) admits the application of Benders decomposition algorithm. To address the semi-infinite nature of the linear program, [18] consider a special case where the sample space Ω is defined by a bounded hyper-rectangle and derive a finite subset of the sample space (without loss of optimality) using extreme points of the hyper-rectangle. Since the test problems used in our experiments do not impose any restrictions on Ω, we adopt a sampling-based discretization of the ambiguity set to tackle (41). Such a discretization satisfies the result in Proposition 2.2 and therefore, provides a suitable benchmark for DRSD. We use the reformulation corresponding to ambiguity set defined by the finite set of observations, i.e. $\Omega:=\{\omega_1,\ldots,\omega_N\}$, $N_s=N$, and $\overline{\omega}_i=\omega_i$ for $i=1,\ldots,N$.

In this second set of experiments, we consider $N=100, 150$, and 500 observations. We use an external sampling approach to construct the instances of reformulation and solve these instances using the Benders decomposition method. We run the DRSD algorithm for the same number of iterations (N) to have the same set of observations for approximating the ambiguity set (recall that we run replications of both algorithms with the same seed for random number generation). The results of this experiment are shown in Table 5.2 for $\epsilon=0.05$. The table shows the average objective function estimates and computational time (in seconds) computed across 30 replications along with half widths of the corresponding confidence interval.

The results indicate that the estimates of the objective function obtained from the DRSD algorithm and the reformulation approach are comparable. For all the test problems, the computational time for both approaches increases with N. We attribute this to the increase in the size of the master problem.

Table 3: Computational Results for 2-DRLP Instances with Wasserstein-1 Ambiguity Set

Problem	N	DRSD Algorithm		Reformulation Approach [48]	
		ObjEst	Time	ObjEst	Time
PGP	100	447.04 (±3.34)	0.05 (±0.00)	444.85 (±3.26)	0.94 (±0.07)
	250	454.06 (±2.64)	0.25 (±0.04)	449.85 (±2.23)	7.02 (±0.63)
	500	457.48 (±2.76)	1.79 (±0.26)	451.57 (±1.92)	27.73 (±1.81)
CEP	100	338295.71 (±14430.81)	0.25 (±0.01)	338295.71 (±14430.81)	0.32 (±0.02)
	250	355054.48 (±11823.37)	3.12 (±0.09)	355054.48 (±11823.37)	2.29 (±0.08)
	500	356757.34 (±6917.77)	13.24 (±0.26)	356757.34 (±6917.77)	10.90 (±0.15)
RETAIL	100	157.09 (±4.00)	0.53 (±0.04)	153.67 (±3.89)	9.41 (±0.67)
	250	155.32 (±3.39)	7.75 (±0.22)	154.06 (±3.40)	331.15 (±13.89)
	500	155.20 (±2.39)	72.02 (±2.05)	154.62 (±2.38)	2189.66 (±65.97)
STORM	100	15504501.91 (±11397)	0.64 (±0.04)	15498236.10 (±11445)	21.51 (±1.09)
	250	15508623.20 (±7481)	8.86 (±0.19)	15501074.50 (±7571)	333.22 (±11.80)
	500	15507815.12 (±5059)	83.33 (±2.55)	-	-

While the additional effort associated with solving distribution separation problems also contributes to the increased computation time in DRSD, the number of subproblems solved in each iteration of Benders decomposition increases with N. In any case, DRSD outperforms Benders decomposition applied to the reformulation across all test problems. Since we ran out of memory when solving the instances of STORM with $N = 500$ using Benders decomposition, we do not report its results.

5.3 Results for 2-DRLPs with ℓ_∞-type Wasserstein Ambiguity Set

In contrast to the case of problems with ℓ_1-type Wasserstein ambiguity sets, a problem with ℓ_∞-type Wasserstein ambiguity set (41) further reduces to a linear program (refer to Theorem 1 of [46]). We use this approach to benchmark the performance of DRSD for ℓ_∞-type Wasserstein ambiguity sets. As in the previous set of experiments, we use the empirical distribution with N observations as reference distribution for the ambiguity sets. We generate the observations using an external sampling approach and set up the linear programming reformulation. We solve this reformulation using an off-the-shelf solver (CPLEX 12.10). We summarize the results for $N = 100, 250$, and 500 in Table 5.3.

For all the problems, the estimates of the objective function obtained from DRSD and the reformulation linear program are comparable. The results show that for instances test problems PGP, CEP, and RETAIL, the linear programming reformulation outperforms the DRSD algorithm. However, for larger problem STORM, the advantages of sequential sampling become prevalent resulting in a nearly 3.5 times decrease in the overall computational time for $N = 500$, for instance.

Remark 5.1. Overall, the computational experiments with all three ambiguity sets illustrate the advantages of the sequential sampling approach of DRSD to tackle large-scale 2-DRLP problems. Before we end this section, we note that the external sampling-based benchmark instances are set up and solved for a given N. Since we are dealing with sampling-based approximations, identifying a suitable N a priori is not a trivial task. A procedure to tackle this task involves solving several instances with progressively increasing sample sizes (see for e.g., [5] for risk-neutral SP). The overall computational cost of identifying a high-quality solution is the cumulative cost associated with individual instances. The DRSD method, and the sequential sampling idea in general, mitigates the need for such an iterative process.

Table 4: Computational Results for 2-DRLP Instances with Wasserstein-∞ Ambiguity Set

Problem	N	DRSD Algorithm		Reformulation Approach [46]	
		ObjEst	Time	ObjEst	Time
PGP	100	448.94 (±3.64)	0.05 (±0.00)	447.10 (±3.26)	0.00 (±0.00)
	250	455.04 (±2.49)	0.24 (±0.04)	450.79 (±2.03)	0.06 (±0.00)
	500	458.40 (±2.94)	1.58 (±0.16)	451.31 (±1.23)	0.21 (±0.01)
CEP	100	338291.53 (±14434.11)	0.25 (±0.01)	338307.14 (±14431.53)	0.00 (±0.00)
	250	355061.07 (±11823.37)	3.13 (±0.08)	355066.83 (±11823.82)	0.06 (±0.01)
	500	356763.93 (±6917.77)	12.65 (±0.28)	356769.76 (±6918.07)	0.22 (±0.02)
RETAIL	100	156.40 (±4.24)	0.44 (±0.02)	153.49 (±3.89)	0.18 (±0.01)
	250	155.12 (±3.43)	7.89 (±0.16)	153.86 (±3.40)	0.59 (±0.02)
	500	155.12 (±2.39)	60.39 (±0.91)	154.42 (±2.38)	1.36 (±0.03)
STORM	100	15504413.45 (±11396.84)	0.56 (±0.01)	15502082.05 (±11445.84)	20.25 (±0.27)
	250	15508768.42 (±7477.52)	8.65 (±0.29)	15504919.06 (±7571.13)	81.22 (±1.30)
	500	15507936.99 (±5058.71)	63.10 (±0.87)	15503343.24 (±5140.36)	220.65 (±2.57)

6 Conclusions and Future Work

We presented a new decomposition approach for solving two-stage distributionally robust linear programs (2-DRLPs) with a general ambiguity set defined using probability distributions with continuous or discrete sample space. Since this approach extended the stochastic decomposition approach of Higle and Sen [23] for 2-DRLPs with a singleton ambiguity set, we referred to it as Distributionally Robust Stochastic Decomposition (DRSD) method. The DRSD method is a sequential sampling-based approach that allows sampling within the optimization step where we solved second-stage subproblem(s) associated with only the current observation in each iteration. While the design of DRSD accommodates general ambiguity sets, we provided its asymptotic convergence analysis for a family of ambiguity sets that includes the well-known moment-based and Wasserstein metric-based ambiguity sets. Furthermore, we performed computational experiments to evaluate the efficiency and effectiveness of solving distributionally robust variants of four well-known stochastic programming test problems that have supports of size ranging from 216 to 10^{81}. Based on our results, we observed that the objective function estimates obtained using the DRSD and the external sampling-based approaches are statistically comparable. These DRSD estimates are obtained while providing computational improvements on most problem instances. Such a computational edge will enable the application of DRO to critical applications that result in large-scale problem instances.

The preliminary computational experiments are encouraging. However, there are two components of the algorithm that require careful deliberation. Since DRSD is a randomized algorithm that simultaneously deals with the approximation of ambiguity sets and recourse function values, the deterministic stopping criteria are not applicable. Therefore, the development of reliable stopping criteria is a potential future research direction. Statistical approaches, similar to those developed in a series of papers for SD [25, 26, 40], could provide initial direction to address this issue. Another future research direction is to incorporate more efficient algorithms to solve the distribution separation problems. For example, instead of resolving distribution separation problem in every iteration, we can utilize a column generation procedure. Finally, we will explore a proximal point algorithm design to that will allow us to maintain a fixed-sized master problem.

A Proofs

In this appendix, we provide the proofs for the propositions related to the asymptotic behavior of the approximate ambiguity sets defined in §2.1 and the recourse function approximation presented in §2.2.

Proof. (Proposition 2.1) For $P = (p(\omega))_{\omega\in\Omega^{k-1}} \in \widehat{\mathfrak{P}}^{k-1}_{\text{mom}}$, it is easy to verify that $P' = (p'(\omega))_{\omega\in\Omega^k} = \Theta^k(P)$ satisfies the support constraint, viz., $\sum_{\omega\in\Omega^k} p'(\omega) = 1$. Now consider for $i = 1,\ldots,q$, we have

$$\begin{aligned}\sum_{\omega\in\Omega^k} p'(\omega)\psi_i(\omega) &= \sum_{\omega\in\Omega^{k-1},\omega\neq\omega^k} p'(\omega)\psi_i(\omega) + p'(\omega^k)\psi_i(\omega^k)\\ &= \theta^k \sum_{\omega\in\Omega^{k-1},\omega\neq\omega^k} p(\omega)\psi_i(\omega) + \theta^k p(\omega^k)\psi_i(\omega^k) + (1-\theta^k)\psi_i(\omega^k)\\ &= \theta^k \sum_{\omega\in\Omega^{k-1}} p(\omega)\psi_i(\omega) + (1-\theta^k)\psi_i(\omega^k) = \hat{b}_i^{k-1} + (1-\theta^k)\psi_i(\omega^k) = \hat{b}_i^k.\end{aligned}$$

This implies that $\Theta^k(P) \in \widehat{\mathfrak{P}}^k_{\text{mom}}$.

Using Proposition 4 in [44], there exists a positive constant χ such that

$$0 \le \mathbb{H}(\widehat{\mathfrak{P}}^k_{\text{mom}}, \mathfrak{P}_{\text{mom}}) \le \chi\|\hat{\mathbf{b}}^k - \mathbf{b}\|.$$

Here, $\mathbf{b} = (b_i)_{i=1}^q$ and $\hat{\mathbf{b}}^k = (\hat{b}_i^k)_{i=1}^q$, and $\|\cdot\|$ denotes the Euclidean norm. Since the approximate ambiguity sets are constructed using independent and identically distributed samples of $\tilde{\omega}$, using law of large numbers, we have $\hat{b}_i^k \to b_i$ for all $i = 1,\ldots,q$. This completes the proof. □

Proof. (Proposition 2.2) Consider approximate ambiguity sets $\widehat{\mathfrak{P}}^{k-1}_{\text{w}}$ and $\widehat{\mathfrak{P}}^k_{\text{w}}$ of the form given in (15b). Let $P = (p(\omega))_{\omega\in\Omega^{k-1}} \in \widehat{\mathfrak{P}}^{k-1}_{\text{w}}$, and let the reconstructed probability distribution be denoted by P'. We can easily check that $P' = \Theta^k(P)$ is indeed a probability distribution. With $P' = (p'(\omega))_{\omega\in\Omega^k}$ fixed, it suffices now to show that the polyhedron

$$\mathcal{E}(P', \widehat{P}^k) = \left\{ \eta' \in \mathbb{R}^{\Omega^k\times\Omega^k} \;\middle|\; \begin{array}{ll} \sum_{\omega'\in\Omega^k} \eta'(\omega,\omega') = p'(\omega) & \forall\omega\in\Omega^k,\\ \sum_{\omega\in\Omega^k} \eta'(\omega,\omega') = \hat{p}^k(\omega') & \forall\omega'\in\Omega^k,\\ \sum_{(\omega,\omega')\in\Omega^k\times\Omega^k} \|\omega-\omega'\|\eta'(\omega,\omega') \le \epsilon & \end{array} \right\}. \tag{42}$$

is non-empty. Since $P \in \widehat{\mathfrak{P}}^{k-1}_{\text{w}}$, there exist $\eta(\omega,\omega')$ for all $(\omega,\omega') \in \Omega^{k-1}\times\Omega^{k-1}$ such that the constraints in the description of the approximate ambiguity set in (15b) are satisfied. We show that $\mathcal{E}$ is non-empty by analyzing two possibilities,

1. We encounter a previously seen observation, i.e., $\omega^k \in \Omega^{k-1}$ and $\Omega^k = \Omega^{k-1}$. Let $\eta'(\omega,\omega') = \theta^k\eta(\omega,\omega')$ for $\omega,\omega' \in \Omega^{k-1}$ and $\omega \neq \omega' \neq \omega^k$; and $\eta'(\omega^k,\omega^k) = \theta^k\eta(\omega^k,\omega^k) + (1-\theta^k)$. We verify the feasibility

of this choice by checking the three sets of constraints in (42). For all $\omega \in \Omega^k$

$$\begin{aligned}\sum_{\omega' \in \Omega^k} \eta'(\omega, \omega') &= \sum_{\omega' \in \Omega^k \setminus \{\omega^k\}} \eta'(\omega, \omega') + \eta'(\omega, \omega^k) \\ &= \sum_{\omega' \in \Omega^{k-1} \setminus \{\omega^k\}} \theta^k \eta(\omega, \omega') + \theta^k \eta(\omega, \omega^k) + \mathbf{1}_{\omega = \omega^k}(1 - \theta^k) \\ &= \theta^k \left(\sum_{\omega' \in \Omega^{k-1}} \eta(\omega, \omega') \right) + \mathbf{1}_{\omega = \omega^k}(1 - \theta^k) = \theta^k p(\omega) + \mathbf{1}_{\omega = \omega^k}(1 - \theta^k) = p'(\omega).\end{aligned}$$

For all $\omega' \in \Omega^k$, we have

$$\begin{aligned}\sum_{\omega \in \Omega^k} \eta'(\omega, \omega') &= \sum_{\omega \in \Omega^k \setminus \{\omega^k\}} \eta'(\omega, \omega') + \eta'(\omega^k, \omega') \\ &= \sum_{\omega \in \Omega^{k-1} \setminus \{\omega^k\}} \theta^k \eta(\omega, \omega') + \theta^k \eta(\omega^k, \omega') + \mathbf{1}_{\omega' = \omega^k}(1 - \theta^k) \\ &= \theta^k \sum_{\omega \in \Omega^{k-1}} \eta'(\omega, \omega') + \mathbf{1}_{\omega' = \omega^k}(1 - \theta^k) = \theta^k \hat{p}^{k-1}(\omega') + \mathbf{1}_{\omega' = \omega^k}(1 - \theta^k) = \hat{p}^k(\omega').\end{aligned}$$

And finally,

$$\begin{aligned}&\sum_{(\omega, \omega') \in \Omega^k \times \Omega^k} \|\omega - \omega'\| \eta'(\omega, \omega') \\ &\qquad = \sum_{\substack{(\omega, \omega') \in \Omega^{k-1} \times \Omega^{k-1} \\ \omega \neq \omega' \neq \omega^k}} \theta^k \|\omega - \omega'\| \eta(\omega, \omega') + \|\omega^k - \omega^k\| \eta'(\omega^k, \omega^k) \\ &\qquad = \theta^k \left(\sum_{(\omega, \omega') \in \Omega^{k-1} \times \Omega^{k-1}} \|\omega - \omega'\| \eta(\omega, \omega') \right) \leq \theta^k \epsilon \leq \epsilon.\end{aligned}$$

Since all the three constraints are satisfied, the chosen values for η is an element of the polyhedron $\mathcal{E}$, and therefore, $\mathcal{E} \neq \emptyset$.

2. We encounter a new observation, i.e., $\omega^k \notin \Omega^{k-1}$. Let $\eta'(\omega, \omega') = \theta^k \eta(\omega, \omega')$ for $\omega, \omega' \in \Omega^{k-1}$, $\eta'(\omega^k, \omega') = 0$ for $\omega' \in \Omega^{k-1}$, $\eta'(\omega, \omega^k) = 0$ for $\omega \in \Omega^{k-1}$, and $\eta'(\omega^k, \omega^k) = (1 - \theta^k)$. Let us again verify the three sets of constraints defining (42) with this choice for η'.

$$\begin{aligned}\sum_{\omega' \in \Omega^k} \eta'(\omega, \omega') &= \sum_{\omega' \in \Omega^k \setminus \{\omega^k\}} \eta'(\omega, \omega') + \eta'(\omega, \omega^k) \\ &= \sum_{\omega' \in \Omega^{k-1}} \theta^k \eta(\omega, \omega') + \mathbf{1}_{\omega = \omega^k}(1 - \theta^k) = \theta^k p(\omega) + \mathbf{1}_{\omega = \omega^k}(1 - \theta^k) = p'(\omega); \\ \sum_{\omega \in \Omega^k} \eta'(\omega, \omega') &= \sum_{\omega \in \Omega^k \setminus \{\omega^k\}} \eta'(\omega, \omega') + \eta'(\omega^k, \omega') \\ &= \sum_{\omega' \in \Omega^{k-1}} \theta^k \eta(\omega, \omega') + \mathbf{1}_{\omega' = \omega^k}(1 - \theta^k) = \theta^k \hat{p}^{k-1} + + \mathbf{1}_{\omega' = \omega^k}(1 - \theta^k) = \hat{p}^k(\omega');\end{aligned}$$

and finally,

$$\begin{aligned}
&\sum_{(\omega,\omega')\in\Omega^k\times\Omega^k} \|\omega-\omega'\|\eta'(\omega,\omega') \\
=\ &\sum_{(\omega,\omega')\in\Omega^{k-1}\times\Omega^{k-1}} \theta^k\|\omega-\omega'\|\eta(\omega,\omega') + \|\omega^k-\omega^k\|\eta'(\omega^k,\omega^k) \\
&\quad + \sum_{\omega\in\Omega^k} \|\omega-\omega^k\|\eta'(\omega,\omega^k) + \sum_{\omega'\in\Omega^k} \|\omega^k-\omega'\|\eta'(\omega^k,\omega') \leq\ \theta^k\epsilon \leq \epsilon.
\end{aligned}$$

Therefore, the value of η' variables satisfies the constraints and $\mathcal{E} \neq \emptyset$. This implies that $\Theta^k(P) \in \widehat{\mathfrak{P}}^k_{\mathrm{w}}$.

Next, let us consider a distribution $Q \in \widehat{\mathfrak{P}}^k_{\mathrm{w}}$. Then,

$$d_{\mathrm{w}}(Q,P^*) \leq\ d_{\mathrm{w}}(Q,\widehat{P}^k) + d_{\mathrm{w}}(\widehat{P}^k,P^*) \leq \epsilon + d_{\mathrm{w}}(\widehat{P}^k,P^*).$$

The above inequality is a consequence of the triangle inequality of Wasserstein distance. Since $Q \in \widehat{\mathfrak{P}}^k_{\mathrm{w}}$, we have $d_{\mathrm{w}}(Q,\widehat{P}^k) \leq \epsilon$. Under compactness assumption for Ω, we have $\mathbb{E}_{P^*}[\exp(\|\tilde{\omega}\|^a)] < \infty$. Therefore, for $d > 2$, Theorem 2 in [17] guarantees

$$\mathrm{Prob}\big[d_{\mathrm{w}}(\widehat{P}^k,P^*) \leq \delta\big] \leq \begin{cases} C\,\exp(-ck\delta^d) & \text{if } \delta > 1 \\ C\,\exp(-ck\delta^a) & \text{if } \delta \leq 1 \end{cases}$$

for all $k \geq 1$. This implies that the $\lim_{k\to\infty} d_{\mathrm{w}}(\widehat{P}^k,P^*) = 0$, almost surely. Consequently, we obtain that $d_{\mathrm{w}}(Q,P^*) \leq \epsilon$ (or equivalently $Q \in \mathfrak{P}_{\mathrm{w}}$) as $k \to \infty$, almost surely. This completes the proof. □

Proof. (Proposition 2.3) Recall that $\mathcal{X} \times \Omega$ is a compact set because of Assumptions (A1) and (A4), and $\{Q^k\}$ is a sequence of continuous (piecewise linear and convex) functions. Further, the construction of the set of dual vertices satisfies $\Pi^0 = \{\mathbf{0}\} \subseteq \ldots \subseteq \Pi^k \subseteq \Pi^{k+1} \subseteq \ldots \subseteq \mathcal{D}$ which ensures that $0 \leq Q^k(x,\omega) \leq Q^{k+1}(x,\omega) \leq Q(x,\omega)$ for all $(x,\omega) \in (\mathcal{X},\Omega)$ and $k \geq 1$. Since $\{Q^k\}$ increases monotonically and is bounded by a finite function Q (due to (A2)), this sequence pointwise converges to some function $\xi(x,\omega) \leq Q(x,\omega)$. Once again due to (A2), we know that the set of dual vertices $\mathcal{D}$ is finite and since $\Pi^k \subseteq \Pi^{k+1} \subseteq \mathcal{D}$, the set $\lim_{k\to\infty} \Pi^k := \overline{\Pi}$ ($\subseteq \mathcal{D}$) is also a finite set. Clearly,

$$\xi(x,\omega) = \lim_{k\to\infty} Q^k(x,\omega) = \max\ \{\pi^\top[r(\omega) - T(\omega)x] \mid \pi \in \overline{\Pi}\}$$

is the optimal value of a LP. Note that the right-hand side is a pointwise maximum of affine function and hence, is a continuous function. The compactness of $\mathcal{X} \times \Omega$, and continuity, monotonicity and pointwise convergence of $\{Q^k\}$ to ξ guarantees that the sequence uniformly converges to ξ (implied by a slight modification of Theorem 7.13 in [38]). □

References

[1] M. Bansal, K. Huang, and S. Mehrotra. Decomposition Algorithms for Two-Stage Distributionally Robust Mixed Binary Programs. *SIAM Journal on Optimization*, pages 2360–2383, 2018.

[2] M. Bansal and S. Mehrotra. On solving two-stage distributionally robust disjunctive programs with a general ambiguity set. *European Journal of Operational Research*, 279(2):296–307, 2019.

[3] G. Bayraksan and D. K. Love. Data-Driven Stochastic Programming Using Phi-Divergences. In *The Operations Research Revolution*, INFORMS TutORials in Operations Research, pages 1–19. INFORMS, 2015.

[4] G. Bayraksan and D.P. Morton. Assessing solution quality in stochastic programs. *Mathematical Programming*, 108(2):495–514, 2006.

[5] G. Bayraksan and D.P. Morton. A sequential sampling procedure for stochastic programming. *Operations Research*, 59(4):898–913, 2011.

[6] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, 2012.

[7] D. Bertsimas, X. V. Doan, K. Natarajan, and C. Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

[8] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

[9] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer, 2011.

[10] M. Breton and S. El Hachem. Algorithms for the solution of stochastic dynamic minimax problems. *Computational Optimization and Applications*, 4(4):317–345, 1995.

[11] G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3-4):197–206, 1955.

[12] G. B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8(1):101–111, 1960.

[13] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58:595–612, 2010.

[14] J. Dupacová. The minimax approach to stochastic programming and an illustrative application. *Stochastics*, 20:73–88, 1987.

[15] D. Duque, S. Mehrotra, and D. Morton. Distributionally robust two-stage stochastic programming. Available at `http://www.optimization-online.org/DB_FILE/2020/09/8042.pdf`, 2020.

[16] E. Erdogan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

[17] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[18] Carlos Andrés Gamboa, Davi Michel Valladão, Alexandre Street, and Tito Homem-de Mello. Decomposition methods for wasserstein-based data-driven distributionally robust problems. *Operations Research Letters*, 49(5):696–702, 2021.

[19] H. Gangammanavar, Y. Liu, and S. Sen. Stochastic decomposition for two-stage stochastic linear programs with random cost coefficients. *INFORMS Journal on Computing*, 33(1):51–71, 2021.

[20] H. Gangammanavar and S. Sen. Stochastic dynamic linear programming: A sequential sampling algorithm for multistage stochastic linear programming. *SIAM Journal on Optimization*, 31(3):2111–2140, 2021.

[21] G. A. Hanasusanto and D. Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

[22] Y. T. Herer, M. Tzur, and E. Yücesan. The multilocation transshipment problem. *IIE Transactions*, 38(3):185–200, 2006.

[23] J. L. Higle and S Sen. Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Mathematics of Operations Research*, 16(3):650–669, 1991.

[24] J. L. Higle and S Sen. Finite master programs in regularized stochastic decomposition. *Mathematical Programming*, 67(1-3):143–168, 1994.

[25] J. L. Higle and S Sen. *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming*. Kluwer Academic Publishers, Boston, MA., 1996.

[26] J. L. Higle and S Sen. Statistical approximations for stochastic linear programming problems. *Annals of Operations Research*, 85(0):173–193, 1999.

[27] R. Huang, S. Qu, Z. Gong, M. Goh, and Y. Ji. Data-driven two-stage distributionally robust optimization with risk aversion. *Applied Soft Computing*, 87:105978, 2020.

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

[29] B. Li, X. Qian, J. Sun, K. L. Teo, and C. Yu. A model of distributionally robust two-stage stochastic convex programming with linear recourse. *Applied Mathematical Modelling*, 58:86–97, 2018.

[30] S. Mehrotra and H. Zhang. Models and algorithms for distributionally robust least squares problems. *Mathematical Programming*, 148(1–2):123–141, 2014.

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

[32] J. M. Mulvey and A. Ruszczyński. A New Scenario Decomposition Method for Large-Scale Stochastic Optimization. *Operations Research*, 43(3):477–490, 1995. Publisher: INFORMS.

[33] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[34] M. Riis and K. A. Andersen. Applying the minimax criterion in stochastic recourse programs. *European Journal of Operational Research*, 165(3):569–584, 2005.

[35] R. T. Rockafellar and R. J. B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Math. Oper. Res.*, 16(1):119–147, 1991.

[36] W. Römisch. Stability of stochastic programming problems. *Handbooks in operations research and management science*, 10:483–554, 2003.

[37] J. O. Royset and R. Szechtman. Optimal budget allocation for sample average approximation. *Operations Research*, 61(3):762–776, 2013.

[38] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York, third edition, 1976. International Series in Pure and Applied Mathematics.

[39] H. Scarf. A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*, chapter 12, pages 201–209. RAND Corporation, Santa Monica CA, 1958.

[40] S. Sen and Y. Liu. Mitigating uncertainty via compromise decisions in two-stage stochastic linear programming: Variance reduction. *Operations Research*, 2016.

[41] A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249, 2004.

[42] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2014.

[43] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542, 2002.

[44] H. Sun and H. Xu. Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, 41(2):377–401, 2016.

[45] R. M. Van Slyke and R. J. B. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17(4):638–663, 1969.

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

[47] C. Zhao and Y. Guan. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Available at `http://www.optimization-online.org/DB_FILE/2015/07/5014`, 2015.

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