Two-Stage Distributionally Robust Optimization: Intuitive Understanding and Algorithm Development from the Primal Perspective

Zhengsong Lu and Bo Zeng

Department of Industrial Engineering, University of Pittsburgh

Abstract

In this paper, we study the two-stage distributionally robust optimization (DRO) problem from the primal perspective. Unlike existing approaches, this perspective allows us to build a deeper and more intuitive understanding on DRO, to leverage classical and well-established solution methods and to develop a general and fast decomposition algorithm (and its variants), and to address a couple of unsolved issues that are critical for modeling and computation. Theoretical analyses regarding the strength, convergence, and iteration complexity of the developed algorithm are also presented. A numerical study on different types of instances of the distributionally robust facility location problem demonstrates that the proposed solution algorithm (and its variants) significantly outperforms existing methods. It solves instances up to several orders of magnitude faster, and successfully addresses new types of practical instances that previously could not be handled. We believe these results will significantly enhance the accessibility of DRO, break down barriers, and unleash its potential to solve real-world challenges.

1 Introduction

Distributionally robust optimization (DRO) has emerged as a powerful optimization paradigm to address uncertainties in decision making. As a flexible unification of stochastic programming (SP) and robust optimization (RO), it adopts an *ambiguity set*, which leverages available distributional information while accounting for unforeseen perturbations on top of that distributional information, to describe randomness. Hence, rather than assuming a single fixed probability distribution, which is the central idea behind SP, DRO takes an RO perspective to hedge against all (often infinitely many) possible distributions within that ambiguity set to ensure robustness. Obviously, the derived solution is more robust than that of SP. Also, by restricting the perturbations in parameters of the underlying distribution, the solution is less conservative compared to that of RO, which just focuses on worst-case scenarios without considering any distributional information. Actually, if those parameters are exact and complete that defines a single distribution, DRO reduces to SP. And if the perturbations are allowed to be arbitrarily large, DRO reduces to RO.

Because of its flexibility in handling data and strength in taking advantage of the associated information, DRO is believed to be an ideal data-driven decision making tool. On one hand, it attracts a large amount of methodological studies in the literature, including those on the design of ambiguity sets and their mathematical and statistical properties, new variants and customizations for different modeling purposes, and strong solution methods to handle complex data/structures, including exact and approximate mixed integer reformulations and specific algorithms. On the other hand, it has found applications in various fields, including finance, energy, supply chain management, and machine learning, where some data on random parameters are available but the true distribution may not be known precisely.

Regardless of the fact that the first recognized study on DRO [1] appeared in 1958, a long time ago, DRO is notably less popular compared to SP or RO, particularly in the context of two-stage decision making. We observe that a couple of issues contribute to this situation. The first one is that present solution methods, which are reviewed in the next section, are not sufficiently strong or lack scalability to handle practical-scale problems. Another one is that some fundamental issues have not been sufficiently addressed, e.g., the feasibility requirement of the recourse problem and the non-convex ambiguity sets, and hence restrictive assumptions are often imposed to circumvent those situations. Actually, existing studies tend to be theoretically sophisticated, providing less intuitive understanding to appreciate DRO. This high theoretical threshold renders DRO less accessible to practitioners, further hindering its applications in the real world.

In this paper, instead of following the mainstream that adopts the dual perspective to study DRO and to develop solution methods, we take the primal perspective to analyze its structure. This "new" angle offers prominent advantages that help us address aforementioned barriers. The remainder of this paper will demonstrate those advantages, with an emphasis on intuitive understanding and developing general and fast algorithms that drastically outperform the state-of-the-art. We first review the relevant literature in Section 2. Then, Section 3 analyzes and computes the worst-case expected value from the primal perspective that is more accessible and general. In Section 4, we present the whole computational algorithm in its basic form for two-stage DRO, with a couple of variants described in Section 5 to achieve a stronger solution capacity. Section 6 reports and analyzes the results of numerical experiments, while Section 7 provides the conclusions of this paper.

Notation: We generally denote matrices in upper-case, vectors in bold lower-case and scalars in lower-case, unless explicitly noted otherwise. Special notations include that: an upper-case letter in calligraphic denotes a major set, $\hat{\ }$ generally denotes a subset and $\frown$ denotes a union of subsets, $\mathbf{M}$ represents a sufficiently large positive number, and $[n]$ indicates the set of positive integers from 1 to n.

2 Related Research in the Literature

In this section, we review the literature that is most relevant to the work presented in this paper. Note that over the last 10 years, many studies on two-stage DRO (and those with more stages) have been published, which involve making recourse decisions after the randomness is cleared [2]. Consider the following tri-level two-stage DRO formulation

$$\mathbf{2-Stg\ DRO}: \qquad w^* = \min_{\mathbf{x}\in\mathcal{X}} f_1(\mathbf{x}) + \sup_{P\in\mathcal{P}} E_P[Q(\mathbf{x},\xi)], \tag{1}$$

where $\mathcal{X}$ denotes the feasible set of the first-stage (also known as "here-and-now") decision variables $\mathbf{x}$, and ξ the random vector representing the uncertain parameter, and $Q(\mathbf{x},\xi)$ the value function capturing the optimal value of the second-stage recourse problem

$$\mathbf{Reco\ Prob}: \ \min\Big\{f_2(\mathbf{x},\boldsymbol{\xi},\mathbf{y}) : \mathbf{y}\in\mathcal{Y}(\mathbf{x},\xi)\Big\}. \tag{2}$$

Variables $\mathbf{y}$ denote the recourse (also known as "wait-and-see") decision variables, which may contain discrete ones, and $\mathcal{Y}(\mathbf{x},\xi)$ is their feasible set. We highlight that $Q(\mathbf{x},\xi)$ returns the optimal value of the recourse problem for given $\mathbf{x}$ and ξ, not representing the whole recourse problem. When the recourse problem is infeasible, i.e., $\mathcal{Y}(\mathbf{x},\xi)=\emptyset$, $Q(\mathbf{x},\xi)$ is set to $+\infty$ by convention. Random vector ξ is defined on a measurable space $(\Xi,\mathcal{F})$ with $\Xi\subseteq\mathbb{R}^{n_\xi}$ being the support and $\mathcal{F}$ a σ-algebra that contains all singletons in Ξ. It follows probability distribution $P\in\mathcal{P}$ where $\mathcal{P}\subseteq\mathcal{M}(\Xi,\mathcal{F})$ is a family of probability distributions and $\mathcal{M}(\Xi,\mathcal{F})$ is the collection of all probability distributions on $(\Xi,\mathcal{F})$. Note that $\mathcal{P}$, assumed to be non-empty in this paper, is commonly referred to as the ambiguity set in the DRO literature.

It can be seen that the fundamental feature that distinguishes DRO from SP or RO is its ambiguity set. Based on the current literature, ambiguity sets generally can be grouped into two main categories: moment-based ones and statistical discrepancy-based ones. For the moment-based ambiguity, the set is usually defined by the moments of ξ across the entire support Ξ, e.g., [3, 4]. For discrepancy-based ambiguity sets, examples include the ϕ-divergence ambiguity sets [5, 6] and Wasserstein ambiguity sets [7, 8, 9]. The latter ones are becoming more popular as they are directly defined with respect to empirical data, a natural demonstration of the data-driven concept. A few less-investigated ambiguity sets also do exist in the literature ([2]). We note that, regardless of the structure of Ξ, $\mathcal{P}$ is always represented as a convex set in the existing literature. For sophisticated $\mathcal{P}$ that requires a mixed integer representation, we have not been aware of any study available yet.

Theoretically, two-stage DRO has been proven to be NP-hard in general [10, 11]. To solve this challenging problem exactly, there exist two types of popular solution strategies.

One is to reformulate it into single-level (nonlinear) mixed integer programs (MIPs) that can be directly handled by existing solvers or computational algorithms, e.g.,[8, 9, 12, 13]. Nevertheless, those reformulations are generally with very large (often infinite) number of variables and/or constraints. Directly computing them turns out to be practically infeasible. Hence, another strategy is to develop decomposition algorithms so that variables and/or constraints are introduced on the fly to reduce the computational burden, e.g., [14, 15, 16, 17, 18, 19]. It is worth highlighting that the fundamental idea underlying those two strategies is to leverage the duality results of the integration-based convex problems (i.e., the sup part in (1)) to reduce it to a monolithic formulation [7, 8, 9, 20, 21].

Regarding decomposition algorithms, they are generally designed with a master-subproblem framework that runs iteratively before convergence. Similar to the case of RO [22], those algorithms can be classified into two groups based on the representation of information feedback from the subproblem to cut the current solution of the master problem. One group is Benders-dual (BD) type of algorithms [15, 17, 18, 19, 23], which use the dual information of the recourse problem and define cutting planes in the form of Benders (optimality) cut. Another group uses the column-and-constraint (C&CG) method to generate cut represented by a new replicate of the whole recourse problem [14, 16, 17, 24]. Common to those decomposition algorithms is that their master problems are augmented over iterations to derive stronger relaxations and therefore tighter lower bounds for (1). On the other hand, the best feasible solution found so far provides an upper bound. Once those two bounds meet or are with a sufficiently small difference, the optimal value and an optimal first stage solution to $\mathbf{2-Stg\ DRO}$ in (1) are obtained.

While those decomposition algorithms are iterative, they, by making use of professional MIP solvers to compute master and subproblems, often demonstrate significantly better computational performance than directly computing DRO's monolithic reformulations. Also, the emerging features of those solvers, such as lazy constraints and the ability to handle bilinear constraints, can be leveraged to increase our solution capacity to address more complex practical problems. In addition to those general and exact methods, several algorithms that handle special structures (e.g., [25]) or compute approximate solutions (e.g., [26, 27, 28]) have also been investigated in the literature, which are beyond the scope of this paper. Regardless of the aforementioned many efforts on developing algorithms for general $\mathbf{2-Stg\ DRO}$, we note that the following fundamental issues remain unsolved or seriously underinvestigated. (i) No existing algorithm has addressed the feasibility (equivalently the infeasibility) issue of the recourse problem yet, noting that in general some $\mathbf{x}$ may render (2) infeasible under some scenario $\xi \in \Xi$. This challenge has always been circumvented by assuming that $Q(\mathbf{x}, \xi)$ is finite for any $\mathbf{x} \in \mathcal{X}$ and $\xi \in \Xi$ combination. Nevertheless, such a strategy is not realistic or practically acceptable for many real-world systems (e.g., critical infrastructure systems), as decision makers might be required to identify an $\mathbf{x}$ that guarantees the feasibility of (2) with respect to $\mathcal{P}$. Moreover, we note from our numerical studies that addressing the feasibility issue in the context of DRO could be computationally much more demanding.

(*ii*) No study has been performed on complex ambiguity sets that are non-convex. Such type of ambiguity sets may happen if data come from different sources and they are rather not consistent, which nevertheless can be represented by employing mixed integer sets. Actually, due to their non-convexity, we believe that simply taking the dual perspective, which is behind current reformulations or solution algorithms, to handle this type of ambiguity sets is not a valid or viable direction. Hence, it would be necessary to extend our solution capacity to solve $\mathbf{2-Stg\ DRO}$ with mixed integer ambiguity sets.

Also, as noted, existing studies on DRO heavily rely on sophisticated mathematical concepts and derivations with different assumptions or conditions. It is desired to develop simpler, more intuitive, and general reasoning and analyses, which will undoubtedly inspire more scholars and practitioners to study and apply DRO. Indeed, a recent study [29] has made an effort to simplify complex mathematical derivations in the context of Wasserstein ambiguity set based DRO, aiming to provide a better accessibility and applicability. The last and probably the most important issue is our rather limited computational capacity to handle general, large-scale and practical DRO instances, which is significantly inferior to what we currently have for instances of SP and RO.

In this paper, we make an effort to address those critical challenges from the primal perspective. Rather than relying on the duality of convex programs constructed on $\mathcal{P}$ to develop analytical insights and computational methods, we adopt a rather simple approach to directly compute the worst-case expected value (WCEV) and leverage it to solve $\mathbf{2-Stg\ DRO}$. It is worth noting that this strategy offers us three key advantages. First, it helps us establish a clear and intuitive understanding of DRO that facilitates its applications in practice. Second, it allows us to directly investigate and attack more challenging structures that might not be approachable by existing methods. Third, it provides new opportunities to develop stronger and more general solution algorithms. The remainder of this paper will showcase those advantages in our analyses and algorithm development.

3 Computing the WCEV from the Primal Perspective

In this section, we consider a core structure and challenge embedded in solving $\mathbf{2-Stg\ DRO}$, i.e., how to compute the worst-case expected value with respect to $\mathcal{P}$. Since we do not assume any particular value of $\mathbf{x}$, we ignore $\mathbf{x}$ in all derivations and analyses of this section to simplify our exposition.

The basic intuitions behind our study and algorithm design are: (*i*) Generally any probability distribution can be represented or approximated with an arbitrary accuracy by a discrete probability distribution; (*ii*) The WCEV over an ambiguity set can be obtained by solving a typical finite mathematical program or by customizing and implementing some well-known optimization algorithms.

3.1 Expected Value and Worst-Case Expected Value

In this subsection, we present some results regarding the computations of the expected value for any distribution and the WCEV over a set of probability distributions. Different from the current mainstream understandings obtained from a dual perspective, those results are derived by directly taking the primal perspective to appreciate those two expected values. They also provide theoretical support for us to develop corresponding algorithms. Recall that $\mathcal{M}(\Xi,\mathcal{F})$ denotes the collection of probability distributions on $(\Xi,\mathcal{F})$, and consider a real-valued function $Q(\xi)$ defined on Ξ. For the remainder of this paper, we make a couple of very mild assumptions.

Assumption 1. *(i)* Support Ξ is a closed and bounded set; *(ii)* $Q(\xi) > -\infty$ for $\xi \in \Xi$.

We mention that after $\mathbf{x}$ is introduced in Section 4, the second part of this assumption is revised to $Q(\mathbf{x},\xi) > -\infty$ for $\mathbf{x} \in \mathcal{X}$ and $\xi \in \Xi$. Next, we consider the general integration based expected value for an arbitrary continuous distribution.

Theorem 1. Assume that $Q(\Xi)$ is Lebesgue integrable over $\underline{\Xi} \equiv \{\xi \in \Xi : Q(\xi) < +\infty\}$. Let $\Big(\boldsymbol{\xi}(n), P(\boldsymbol{\xi}(n))\Big)$ denote a discrete probability distribution over $\boldsymbol{\xi}(n) \equiv \big(\xi_1(n),\ldots,\xi_n(n)\big)$ with $P(\boldsymbol{\xi}(n)) \equiv \Big(p(\xi_1(n)),\ldots,p(\xi_n(n))\Big)$ being the associated probability. For any probability distribution $P \in \mathcal{M}(\Xi,\mathcal{F})$, there exists a sequence of discrete probability distributions in the form of $\left\{\Big(\boldsymbol{\xi}(n), P(\boldsymbol{\xi}(n))\Big)\right\}_{n=1}^{+\infty}$ such that

$$E_P[Q(\xi)] = \int_{\xi\in\Xi} Q(\xi)P(d\xi) = \lim_{n\to+\infty}\sum_{j=1}^{n} Q\big(\xi_j(n)\big)p(\xi_j(n)). \qquad \square$$

We note that Theorem 1 is rather mathematically intuitive. It indicates that we may be able to replace the integration operation with a weighted sum over a set of discrete scenarios, which should be computationally much more friendly. Actually, if we aim to compute the WCEV over a family of distributions, i.e., the ambiguity set underlying DRO, it has been reported that the worst-case one may reduce to a discrete probability distribution. Next, we consider the following ambiguity set that generalizes all known moment-based and Wasserstein ambiguity sets adopted in the DRO literature.

$$\mathcal{P} = \left\{P \in \mathcal{M}(\Xi,\mathcal{F}) : \int_{\Xi} P(d\xi) = 1,\ \int_{\Xi} \psi_i(\xi)P(d\xi) \le \gamma_i \quad \forall i \in [m]\right\}. \tag{3}$$

Function $\psi_i(\cdot)$ is real valued and bounded on Ξ for $i = 1,\ldots,m$. With this ambiguity set, we define the following optimization problem to compute the WCEV.

$$\textbf{WCEV} : \sup_{P\in\mathcal{P}} E_P[Q(\xi)] \equiv \sup\left\{\int_{\Xi} Q(\xi)P(d\xi) : P \in \mathcal{P}\right\}. \tag{4}$$

When Ξ is a finite discrete set, we have $P \equiv (p_1,\ldots,p_n)$ with $n = |\Xi|$. And the integration in

(3) and (4) reduces to summation. As the analysis for this case is simpler, we, unless noted otherwise, mainly present our derivations and analyses for the case where $|\Xi|$ is not finite. We mention that those results, with little or minor changes, are naturally applicable when $|\Xi|$ is finite.

Remark 1. It is clear that an ambiguity set defined by moment inequalities is a special case of (3). Also, an ambiguity set defined by Wasserstein metric, as explained in Section 5.1, belongs to (3). Regarding a ϕ-divergence-based ambiguity set, we can approximate ϕ, a non-negative and convex function, by a piece-wise linear one with an arbitrary accuracy. As a result, its mathematical representation is also in the form of (3). □

Consider an arbitrary $P \in \mathcal{P}$ and let ε be a sufficiently small positive constant. Following the proof of Theorem 1 in Appendix A.1, it is easy to see that we can always identify a partition of Ξ for ψ_i and derive a discrete distribution that approximates P with a violation of γ_i up to ε. By taking the intersection of those partitions for Q and ψ_i for $i \in [m]$, a finer partition yielding a stronger approximation to P with less violation of all γ_i's will be obtained. Such a construction procedure leads to the next result readily that allows us to leverage a sequence of discrete distributions to compute **WCEV** formulation with an arbitrary accuracy. Before that, we define that a discrete probability distribution $\Big(\boldsymbol{\xi}(n), P(\boldsymbol{\xi}(n))\Big)$ is ε-feasible to $\mathcal{P}$ if $\sum_{j=1}^{n} p(\xi_j(n))\psi_i(\xi_j(n)) \le \gamma_i + \varepsilon$ for $i \in [m]$. Moreover, it is an ε-approximation to $P \in \mathcal{P}$ if $\big|\int_{\Xi} Q(\xi)P(d\xi) - \sum_{j=1}^{n} p(\xi_j(n))Q(\xi_j(n))\big| \le \varepsilon$.

Corollary 2. Suppose that $\sup_{P\in\mathcal{P}} E_P[Q(\xi)]$ is finite. (*i*) For every $P \in \mathcal{P}$, there exists an ε-feasible discrete distribution, denoted by $\Big(\boldsymbol{\xi}(n), P(\boldsymbol{\xi}(n))\Big)$, that is an ε-approximation to P. (*ii*) There exists an ε-feasible discrete distribution, denoted by $\Big(\boldsymbol{\xi}'(n), P'(\boldsymbol{\xi}'(n))\Big)$, that is ε-optimal to the **WCEV** formulation, i.e.,

$$\Big|\sup_{P\in\mathcal{P}} E_P[Q(\xi)] - \sum_{j=1}^{n} p'(\xi'_j(n))Q(\xi'_j(n))\Big| \le \varepsilon. \qquad \square$$

Note that the result in Corollary 2 holds in general as it does not require any restrictive condition. Yet, if an actual solution to the **WCEV** problem is desired, Corollary 2 does not help to ensure the existence of an optimal P or to obtain a discrete distribution that is feasible and (ε-)optimal. Next, we present some sufficient conditions that substantially support us on those issues.

Theorem 3. Suppose $Q(\cdot)$ is an upper semicontinuous function, and $\psi_i(\cdot)$ are lower semicontinuous functions over Ξ for $i \in [m]$. The optimal value of **WCEV** can be attained, i.e., $\sup_{P\in\mathcal{P}} E_P[Q(\xi)] = \max_{P\in\mathcal{P}} E_P[Q(\xi)]$. Moreover, if $\mathcal{P}$ has an interior point P^0 and function $g(t)$, defined as:

$$g(t) = \sup_{P \in \mathcal{M}(\Xi,\mathcal{F})} \left\{ \int_\Xi Q(\xi)P(d\xi) : \int_\Xi P(d\xi) = 1, \int_\Xi \psi_i(\xi)P(d\xi) \leq \gamma_i - t \quad \forall i \in [m] \right\},$$

is continuous at $t = 0$, there exists a sequence of discrete probability distributions such that their associated expected values of $Q(\cdot)$'s converge to $\max_{P \in \mathcal{P}} E_P[Q(\xi)]$. □

Proof. We start with the proof for the first statement. Since Ξ is a compact metric space and $Q(\cdot)$ is upper semicontinuous, $Q(\cdot)$ is bounded and the set of all probability measures $\mathcal{P}(\Xi)$ on $(\Xi, \mathcal{F})$ is weakly compact. Let $\{P^k\}_k \subseteq \mathcal{P}(\Xi)$ be a sequence converging weakly to $P^\infty \in \mathcal{P}(\Xi)$. According to Portmanteau Theorem, given that $\psi_i(\cdot)$ are lower semicontinuous and bounded from below on Ξ for $i \in [m]$, we have

$$\liminf_{k \to \infty} \int_\Xi \psi_i(\xi)P^k(d\xi) \geq \int_\Xi \psi_i(\xi)P^\infty(d\xi),$$

indicating that $P \mapsto \int_\Xi \psi_i(\xi)P(d\xi)$ is weakly lower semicontinuous on $\mathcal{P}(\Xi)$. Given that the pre-image of $(-\infty, \gamma_i]$ under a lower semicontinuous mapping are closed for $i \in [m]$, $\mathcal{P}_{\gamma_i} = \left\{P \in \mathcal{P}(\Xi) : \int_\Xi \psi_i(\xi)P(d\xi) \leq \gamma_i\right\}$ are weakly closed for $i \in [m]$. As $\mathcal{P} = \bigcap_{i \in [m]} \mathcal{P}_{\gamma_i} \subseteq \mathcal{P}(\Xi)$, $\mathcal{P}$ is weakly closed and tight, i.e., $\mathcal{P}$ is weakly compact. Applying Portmanteau Theorem again, we can conclude that $P \mapsto \int_\Xi Q(\xi)P(d\xi)$ is weakly upper semicontinuous on $\mathcal{P}$. Therefore, by Weierstrass' Theorem, the optimal value of **WCEV** problem can be attained.

We next prove the second statement. Our proof depends on the construction of an auxiliary sequence of measures $\{P^k\}_k$:

$$P^k = \operatorname*{argmax}_{P \in \mathcal{M}(\Xi,\mathcal{F})} \left\{ \int_\Xi Q(\xi)P(d\xi) : \int_\Xi P(d\xi) = 1, \int_\Xi \psi_i(\xi)P(d\xi) \leq \gamma_i - \frac{\Delta}{k} \quad \forall i \in [m] \right\} \tag{5}$$

where Δ is a fixed value that ensures the feasible set of (5) is consistent for all k, i.e., (5) is feasible, and thus P^k's are attained for any k. Note that Δ can always be found given the existence of interior point P^0. Clearly, the sequence, $\left\{\int_\Xi Q(\xi)P^k(d\xi)\right\}_k$, is monotonically increasing and bounded above by the optimal value of **WCEV** problem. Because of the continuity of $g(t)$ at $t = 0$ and the monotone convergence theorem, we have

$$\lim_{k \to \infty} \int_\Xi Q(\xi)P^k(d\xi) = \max_{P \in \mathcal{P}} E_P[Q(\xi)].$$

Using the argument presented before Corollary 2, for P^k, we can obtain a discrete distribution based on a partition $\{\Xi_j(n_k)\}_j$ of Ξ that satisfies

$$\text{both} \quad \lim_{n_k \to +\infty} \sum_{j=1}^{n_k} Q\big(\xi_j(n_k)\big)p\big(\xi_j(n_k)\big) = \int_\Xi Q(\xi)P^k(d\xi)$$

$$\text{and} \qquad \lim_{n_k \to +\infty} \sum_{j=1}^{n_k} \psi_i\big(\xi_j(n_k)\big)p\big(\xi_j(n_k)\big) = \int_\Xi \psi_i(\xi)P^k(d\xi).$$

For the discrete distribution associated with P^k, we let

$$N_k = \min\Bigg\{ n_k : \Big| \sum_{j=1}^{n_k} Q\big(\xi_j(n_k)\big)p\big(\xi_j(n_k)\big) - \int_\Xi Q(\xi)P^k(d\xi)\Big| \le \frac{\Delta}{k}, \\ \Big| \sum_{j=1}^{n_k} \psi_i\big(\xi_j(n_k)\big)p\big(\xi_j(n_k)\big) - \int_\Xi \psi_i(\xi)P^k(d\xi)\Big| \le \frac{\Delta}{k} \quad \forall i \in [m] \Bigg\}.$$

Claim: All discrete distributions in $\Big\{\Big(\boldsymbol{\xi}(N_k), P(\boldsymbol{\xi}(N_k))\Big)\Big\}_{k=1}^{+\infty}$ are feasible and

$$\lim_{k\to\infty} \sum_{j=1}^{N_k} Q\big(\xi_j(N_k)\big)p\big(\xi_j(N_k)\big) = \max_{P\in\mathcal{P}} E_P[Q(\xi)].$$

Proof of Claim: Obviously, any discrete distribution in $\Big\{\Big(\boldsymbol{\xi}(N_k), P(\boldsymbol{\xi}(N_k))\Big)\Big\}_{k=1}^{+\infty}$ is feasible as it satisfies all constraints in $\mathcal{P}$. For every $\varepsilon > 0$, there is $k(\varepsilon)$ such that for all $k \ge k(\varepsilon)$, we have

$$\Big|\max_{P\in\mathcal{P}} E_P[Q(\xi)] - \sum_{j=1}^{N_k} Q\big(\xi_j(N_k)\big)p\big(\xi_j(N_k)\big)\Big| \le \Big|\max_{P\in\mathcal{P}} E_P[Q(\xi)] - \int_\Xi Q(\xi)P^k(d\xi)\Big| \\ + \Big|\int_\Xi Q(\xi)P^k(d\xi) - \sum_{j=1}^{N_k} Q\big(\xi_j(N_k)\big)p\big(\xi_j(N_k)\big)\Big| \le \varepsilon.$$

The first inequality follows from the triangle inequality. As for the second inequality, the convergence of $\Big\{\int_\Xi Q(\xi)P^k(d\xi)\Big\}_k$ guarantees the difference between two integrals is not more than $\frac{\varepsilon}{2}$. Furthermore, by the definition of N_k and $\Big\{\Big(\boldsymbol{\xi}(N_k), P(\boldsymbol{\xi}(N_k))\Big)\Big\}_{k=1}^{+\infty}$, the difference between the sum and the integral can be made no more than $\frac{\varepsilon}{2}$ when $k(\varepsilon) \ge \frac{2\Delta}{\varepsilon}$, which results in the second inequality. Hence, the claim holds when $k \longrightarrow +\infty$.

With this claim proved, the second statement of this theorem follows directly. □

By simply making use of Theorems 1 and 3, we have the next result.

Corollary 4. Assuming that all the sufficient conditions in Theorem 3 hold, the next formulation is equivalent to **WCEV** problem in (4).

$$\max_{P\in\mathcal{P}} E_P[Q(\xi)] = \lim_{n\to\infty} \max_{\big(\boldsymbol{\xi}(n), P(\boldsymbol{\xi}(n))\big)} \Big\{ \sum_{j=1}^{n} Q\big(\xi_j(n)\big)p(\xi_j(n)) : P \in \mathcal{P}, \xi_j(n) \in \Xi \;\; \forall j \in [n] \Big\} \tag{6}$$

Proof. By the definition of **WCEV** problem, it is clear that

$$\max_{P\in\mathcal{P}} E_P[Q(\xi)] \geq \lim_{n\to\infty} \max_{\{(\xi_j(n),p(\xi_j(n)))\}_{j=1}^n} \Big\{ \sum_{j=1}^n Q(\xi_j(n))p(\xi_j(n)) : P\in\mathcal{P}, \xi_j(n)\in\Xi \quad \forall j\in[n] \Big\}.$$

According to Theorem 3, there exists a sequence of discrete distributions in $\mathcal{P}$ such that $\max_{P\in\mathcal{P}} E_P[Q(\xi)] = \lim_{n\to\infty} \sum_{j=1}^n Q(\tilde{\xi}_j(n))\tilde{p}(\tilde{\xi}_j(n))$. Given that

$$\sum_{j=1}^n Q(\tilde{\xi}_j(n))\tilde{p}(\tilde{\xi}_j(n)) \leq \max_{\{(\xi_j(n),p(\xi_j(n)))\}_{j=1}^n} \sum_{j=1}^n Q(\xi_j(n))p(\xi_j(n)),$$

we have

$$\max_{P\in\mathcal{P}} E_P[Q(\xi)] \leq \lim_{n\to\infty} \max_{\{(\xi_j(n),p(\xi_j(n)))\}_{j=1}^n} \Big\{ \sum_{j=1}^n Q(\xi_j(n))p(\xi_j(n)) : P\in\mathcal{P}, \xi_j(n)\in\Xi \quad \forall j\in[n] \Big\}.$$

Hence, the desired result simply follows. □

Actually, it is worth highlighting that **WCEV** problem in (4) can be seen as an infinite-column linear program. Under some little bit stronger conditions, a more insightful result has been derived [20]: **WCEV** problem, if subject to equality constraints in (3) and with a finite optimal value, has an optimal solution that has at most $(m+1)$ scenarios with non-zero probabilities. This insight actually is verified in our numerical result presented in Section 6. Extending and making use of this result, we can build a finite mathematical program (FMP) that is concise and helps to solve **WCEV** problem exactly.

Proposition 5. Suppose $Q(\cdot)$ is upper semicontinuous and $\psi_i(\cdot)$ are continuous over Ξ for $i\in[m]$. The WCEV, i.e., the optimal value of **WCEV** problem, is attainable and can be obtained by solving the following FMP

$$\begin{aligned} \mathbf{WCEV-FMP} : \max \Big\{ \sum_{j=1}^{m+1} Q(\xi_j)p_j : \sum_{j=1}^{m+1} p_j = 1,\ \sum_{j=1}^{m+1} \psi_i(\xi_j)p_j \leq \gamma_i \quad \forall i\in[m], \\ \xi_j\in\Xi \quad \forall j\in[m+1],\ p_j\geq 0 \quad \forall j\in[m+1] \Big\}. \end{aligned} \tag{7}$$

Its optimal solution, denoted by (P^*,ξ^*) with $P^*\equiv(p_1^*,\ldots,p_{m+1}^*)$ and $\xi^*\equiv(\xi_1^*,\ldots,\xi_{m+1}^*)$, is then feasible and optimal to **WCEV** problem. □

Proof. Given that $\psi_i(\cdot)$ for $i\in[m]$ are continuous, according to Theorem 3, the optimal value of the original **WCEV** in (4) can be attained. Let $P'(\Xi)$ be an optimal solution and compute $\gamma_i' = \int_\Xi \psi_i(\xi)P'(d\xi)$ for $i\in[m]$. Then, the original **WCEV** problem is equivalent to the following one with $(m+1)$ equality constraints.

$$\max_{P\in\mathcal{M}(\Xi,\mathcal{F})} \Big\{ \int_\Xi Q(\xi)P(d\xi) : \int_\Xi P(d\xi) = 1,\ \int_\Xi \psi_i(\xi)P(d\xi) = \gamma_i' \quad \forall i\in[m] \Big\}. \tag{8}$$

According to Richter-Rogosinski Theorem [20], (8) has an optimal distribution with a support of at most $(m+1)$ scenarios. Clearly, it is also feasible and optimal to the original **WCEV** problem. Consequently, **WCEV** reduces to **WCEV – FMP**, and an optimal solution to **WCEV – FMP** solves **WCEV**. □

Remark 2. (i) We mention that, compared to the integration-based **WCEV** formulation, both (6) and (7) are more accessible. The finite nonlinear program **WCEV – FMP** certainly allows us to take advantage of many existing mathematical programming tools or results. For example, a strong nonlinear programming solver or algorithm can be readily used as an oracle to compute **WCEV – OPT** if the incorporation of $Q(\mathbf{x})$ and ψ_i is computationally friendly. Indeed, even if the oracle is a fast approximation or heuristic one, i.e., the exactness of the derived (P^*, ξ^*) cannot be guaranteed, it provides a basis for applying more sophisticated procedures for refinements. On the other hand, (6) indicates that in general we can gradually augment $\boldsymbol{\xi}(n)$ and associated $P(\boldsymbol{\xi}(n))$ to approach **WCEV** arbitrarily.
(ii) We would like to highlight the critical advantage of the primal representation demonstrated by (6) and (7) in handling much more complex and general ambiguity sets. Note that Corollary 4 does not require $\mathcal{P}$ to be convex in P. It is very different from the existing duality-based methodologies to compute **WCEV** problem. Indeed, if $\mathcal{P}$ is a linear mixed integer set and an upper bound on the number of its constraints is known, it is viable to construct the corresponding **WCEV – FMP** to compute **WCEV**. Certainly, solution procedures for mixed integer nonlinear programs are needed. Such a general situation is addressed in Section 5.2. □

Nevertheless, we mention that **WCEV – FMP** is a challenging non-convex program, given the bilinear terms between P and $Q(\xi)$ or $\psi_i(\xi)$ in (7). Actually, in the context of **2 – Stg DRO**, value function $Q(\xi)$ represents the recourse cost, which is a complex function of ξ and renders **WCEV – FMP** computationally very intractable. Even just with first moment constraints, we observe that it could take an extremely long time for a state-of-the-art professional solver to solve small-scale instances. Hence, regardless of its simplicity and compactness, **WCEV – FMP** is still difficult. On the other hand, (6) inspires us to develop computationally effective procedures to compute **WCEV** problem (including its variants).

3.2 A Decomposition Algorithm for WCEV Problem

In the remainder of this paper, we assume, unless noted otherwise, the sufficient conditions presented in Proposition 5 to ensure the attainability of the WCEV. Rather than treating ambiguity set $\mathcal{P}$ as a whole set, we consider its sample space and the probability distribution in a separate fashion. Recall that $\boldsymbol{\xi}_n \equiv (\xi_1, \dots, \xi_n)$ represents a pool of discrete scenarios, and $P(\boldsymbol{\xi}_n) \equiv \big(p(\xi_1), \dots, p(\xi_n)\big) \in \mathcal{P}(\boldsymbol{\xi}_n) \subseteq \mathcal{P}$ is a discrete probability distribution and the subset of $\mathcal{P}$ defined on $\boldsymbol{\xi}_n$. Also, the optimal value of an infeasible maximization problem is conventionally set to $-\infty$. The next result directly follows from the σ-algebra of Ξ and Corollary 4.

Corollary 6. For **WCEV** problem, i.e., $\max_{P\in\mathcal{P}} E_P[Q(\xi)]$, we have

$$\max_{P\in\mathcal{P}} E_P[Q(\xi)] \geq \max_{\boldsymbol{\xi}_n\subseteq\Xi}\ \max_{P(\boldsymbol{\xi}_n)\in\mathcal{P}} \sum_{j=1}^{n} Q(\xi_j)p(\xi_j) \geq \max_{P(\boldsymbol{\xi}_n^0)\in\mathcal{P}(\boldsymbol{\xi}_n^0)} \sum_{j=1}^{n} Q(\xi_j^0)p(\xi_j^0), \tag{9}$$

where $\boldsymbol{\xi}_n^0 \equiv \{\xi_1^0,\ldots,\xi_n^0\}$ is a set of fixed scenarios. □

The non-convex program in the middle of (9) actually equals the WCEV if $n \geq m+1$. The rightmost optimization problem, although just providing a lower bound, is a very simple linear program (LP) in P.

Remark 3. It is worth highlighting that the lower bound from that LP in (9) is the strongest one we can have for given $\boldsymbol{\xi}_n^0$. That is, if its strength with respect to $\max_{P\in\mathcal{P}} E_P[Q(\xi)]$ is weak, $\boldsymbol{\xi}_n^0$ should be expanded by including additional nontrivial scenarios. Computationally, we can start with a small-sized $\boldsymbol{\xi}_n^0$ and then gradually expand it for stronger lower bounds. □

Actually, if the expansion of $\boldsymbol{\xi}_n^0$ can be managed appropriately, the optimal value of that LP approaches $\max_{P\in\mathcal{P}} E_P[Q(\xi)]$ exactly or with an arbitrary accuracy, which corresponds to the limit operation presented in Corollary 4. We mention that such an expansion process can be realized by customizing the well-known column generation (CG) algorithm, a classical decomposition method proposed to solve large-scale LPs [30, 31, 32, 33]. Next, we first present the explicit form of the LP in (9), referred to as the *pricing master problem* (PMP) in the remainder of this paper.

$$\begin{aligned}\textbf{PMP}:\ \underline{\eta}^*(\boldsymbol{\xi}_n^0) = \max\Big\{ &\sum_{j=1}^{n} Q(\xi_j^0)p(\xi_j^0) : \sum_{j=1}^{n} p(\xi_j^0) = 1,\ \sum_{j=1}^{n} \psi_i(\xi_j^0)p(\xi_j^0) \leq \gamma_i \quad \forall i \in [m],\\ &p(\xi_j^0) \geq 0 \quad \forall j\in[n]\Big\}.\end{aligned} \tag{10}$$

Let α and $\boldsymbol{\beta} \equiv [\beta_1,\ldots,\beta_m]$ be dual variables of its constraints, respectively. Supposing that **PMP** is feasible and its shadow prices are $(\alpha^*,\boldsymbol{\beta}^*)$, the corresponding *pricing subproblem* (PSP) to derive a new scenario with the largest reduced cost is

$$\textbf{PSP}:\ \nu^*(\boldsymbol{\xi}_n^0) = \max_{\xi\in\Xi} Q(\xi) - \alpha^* - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^*. \tag{11}$$

Next, we provide an estimation on the strength of the lower bound derived from $\boldsymbol{\xi}_n^0$.

Proposition 7. Suppose that **PMP** is feasible and both **PMP** and **PSP** are solved to optimality. We have

$$\underline{\eta}^*(\boldsymbol{\xi}_n^0) \leq \max_{P\in\mathcal{P}} E_P[Q(\xi)] \leq \underline{\eta}^*(\boldsymbol{\xi}_n^0) + \nu^*(\boldsymbol{\xi}_n^0).$$ □

Proof. Note that it is sufficient to prove the second inequality. From **PSP** in (11), it can be seen that $v^*(\boldsymbol{\xi}_n^0) + \alpha^* \geq \max_{\xi \in \Xi} Q(\xi) - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^*$, or equivalently

$$v^*(\boldsymbol{\xi}_n^0) + \alpha^* \geq Q(\xi) - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^* \quad \forall \xi \in \Xi.$$

On the other hand, noting that (4) is an infinite-column linear program with the strong duality [20], we have its dual problem

$$\min \left\{ \alpha + \sum_{i=1}^{m} \beta_i \gamma_i : \alpha + \sum_{i=1}^{m} \psi_i(\xi)\beta_i \geq Q(\xi) \ \ \forall \xi \in \Xi, \ \alpha \text{ free}, \ \beta_i \geq 0 \ \ \forall i \in [m] \right\}.$$

Clearly, by setting $\alpha = v^*(\boldsymbol{\xi}_n^0) + \alpha^*$ and $\beta_i = \beta_i^*$ for all i, we obtain a feasible solution to the dual problem. Hence, we have

$$\max_{P \in \mathcal{P}} E_P[Q(\xi)] \leq v^*(\boldsymbol{\xi}_n^0) + \alpha^* + \sum_i \beta^* \gamma_i = v^*(\boldsymbol{\xi}_n^0) + \underline{\eta}^*(\boldsymbol{\xi}_n^0),$$

where the last equality follows from the strong duality of **PMP**. □

According to Proposition 7, for a given $\boldsymbol{\xi}_n^0$ (and hence **PMP** and **PSP** are given), if the optimal value of **PSP** is 0, $\underline{\eta}^*(\boldsymbol{\xi}_n^0)$ equals the WCEV. Otherwise, denoting **PSP**'s optimal solution by ξ^*, we can augment $\boldsymbol{\xi}_n^0$ by including ξ^* and recompute **PMP**. This process is repeated until the reduced price becomes sufficiently small, which is the basic idea of CG algorithm. For simplicity, we refer to the approach directly constructing **WCEV − FMP** and solving it by some stand-alone method(s) as *Oracle-1*, and one using an iterative procedure as *Oracle-2*. In the context of this paper, *Oracle-2* is just the following customized CG algorithm. Note that we do not include subscript n, unless we need to track the number of scenarios in $\boldsymbol{\xi}^0$.

Oracle-2: The CG Algorithm to Compute the WCEV

```
Step 1 Given initial $\boldsymbol{\xi}^0$ and optimality tolerance $\varepsilon$, set the iteration counter $k = 1$.
Step 2 Solve PMP to derive optimal value $\underline{\eta}^{k*}(\boldsymbol{\xi}^0)$ and shadow price $(\alpha^*, \boldsymbol{\beta}^*)$.
Step 3 Solve PSP to derive its optimal solution $\xi^*$ and optimal value $v^*(\boldsymbol{\xi}^0)$.
Step 4 If $v^*(\boldsymbol{\xi}_n^0) \leq \varepsilon$, report $\underline{\eta}^{k*}(\boldsymbol{\xi}^0)$ as the optimal value of (4) and terminate. Otherwise,
       update $\boldsymbol{\xi}^0 = \boldsymbol{\xi}^0 \cup \{\xi^*\}$ and $k = k + 1$, and go to Step 2.
```

We mention that, as an algorithm from the primal perspective to compute **WCEV** problem, it generates a set of discrete scenarios and their respective probabilities. As discussed next, this algorithm brings us several important features that can be further explored.

Remark 4. (i) Similar to **WCEV − FMP**, this CG algorithm mainly works in the primal space and is general in handling different ambiguity sets. Even if $\mathcal{P}$ is a mixed integer set, for which the strong duality-based approaches do not work, it is feasible to compute the WCEV by the extension of *Oracle-2*, i.e., the well-established Brand-and-Price (B&P) algorithm.

Alternatively, we propose a new CG variant in Section 5.2 to handle such ambiguity sets without developing B&P procedures.

(*ii*) At the termination, we also have

$$\underline{\eta}^{k*}(\boldsymbol{\xi}^0) = \max\left\{\sum_{\xi\in\boldsymbol{\xi}^0} Q(\xi)p(\xi) : (p_\xi)_{\xi\in\boldsymbol{\xi}^0} \in \mathcal{P}\right\}. \tag{12}$$

By Proposition 7, the WCEV is bounded by $\underline{\eta}^{k*}(\boldsymbol{\xi}^0) + \varepsilon$ at termination. Instead of utilizing a time-consuming oracle to solve **PSP** exactly, this observation allows us to develop a fast approximation algorithm for complex **PSP** and hence for the WCEV, as long as its approximation bound is available.

(*iii*) Compared to directly computing the nonlinear program **WCEV − FMP** by a professional solver, *Oracle-2* demonstrates a superior capacity that is generally faster by multiple orders of magnitude. Actually, it is rather a vanilla version of CG. More advanced implementation techniques, after customization to fit into the DRO context, should be able to help us accelerate the computation or the convergence.

(*iv*) Although *Oracle-2* is currently faster by several orders of magnitude, we anticipate that this situation may change with the development of specialized techniques to improve *Oracle-1*. Yet, the applicability of *Oracle-2* is less restrictive, as it does not depend on the number of scenarios that are with non-zero probabilities as represented in Proposition 5. □

According to the CG literature, it is not restrictive to assume that **PMP** is feasible, as a modified CG based on Farka's Lemma can generate new scenarios (i.e., columns) to ensure it to be feasible. Alternatively, as shown in Section 4.1, the algorithm framework for **2 − Stg DRO** can be used to address this issue of infeasible **PMP**. In the next subsection, we consider the convergence issue of *Oracle-2* and its computational complexity.

3.3 Convergence and Complexity of *Oracle-2*

Without loss of generality, we assume $\boldsymbol{\xi}_0 = \varnothing$ at the initialization. Since *Oracle-2* derives optimal ξ^* from **PSP** and updates $\boldsymbol{\xi}^0 = \boldsymbol{\xi}^0 \cup \xi^*$ in each iteration, a sequence $\left\{\underline{\eta}^{k*}(\boldsymbol{\xi}^0)\right\}_k$, consisting of optimal values of **PMP**s, can be found. As it can be easily recovered, we omit $\boldsymbol{\xi}_0$ when appropriate to simplify our arguments, unless otherwise stated. The next theorem reveals the relationship between the limit of $\left\{\underline{\eta}^{k*}\right\}_k$ and the WCEV.

Lemma 8. Assuming that $\max_{P\in\mathcal{P}} E_P[Q(\xi)]$ exists and an ε-optimal solution of **WCEV** problem can be derived in finite iterations for $\varepsilon > 0$, $\left\{\underline{\eta}^{k*}\right\}_k$ converges to $\max_{P\in\mathcal{P}} E_P[Q(\xi)]$. □

Proof. Consider the simple case where *Oracle-2* terminates with $v^* = 0$ in the k_t-th iteration and hence it involves $k_t - 1$ different ξ's. According to Proposition 7, we obtain the optimal value of **WCEV** problem, which leads the conclusion. Next, we consider the other case where an ε−optimal solution of **WCEV** problem can be retrieved in finite iterations.

According to Corollary 4, for any given $\varepsilon > 0$, there exists an integer K such that

$$\max_{P \in \mathcal{P}} E_P[Q(\xi)] - \varepsilon < \underline{\eta}^{K*} \le \max_{P \in \mathcal{P}} E_P[Q(\xi)].$$

Otherwise $\max_{P \in \mathcal{P}} E_P[Q(\xi)] - \varepsilon$ would be an upper bound of $\{\underline{\eta}^{k*}\}_k$, which contradicts the derivation of an $\varepsilon-$optimal solution by *Oracle-2*. Because ε can be arbitrarily small, $\{\underline{\eta}^{k*}\}_k$ is increasing with k, and because of the monotone convergence theorem,

$$\lim_{k \to +\infty} \underline{\eta}^{k*} = \max_{P \in \mathcal{P}} E_P[Q(\xi)],$$

which is the expected conclusion. □

The next theorem shows that we actually can always find an optimal or ε-optimal solution of **WCEV** problem in finite iterations under some mild conditions. Before that, consider a continuous function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, for $\mathbf{y} \in \mathcal{Y}$ and for any $\varepsilon > 0$, if there exists a $\delta > 0$, such that

$$|f(\mathbf{x}_1, \mathbf{y}) - f(\mathbf{x}_2, \mathbf{y})| \le \varepsilon \quad \text{if} \quad \|\mathbf{x}_1 - \mathbf{x}_2\| \le \delta \quad \text{for } \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X},$$

we say that $f(\mathbf{x}, \mathbf{y})$ is continuous, uniformly with respect to $\mathbf{x}$ for $\mathbf{y} \in \mathcal{Y}$.

Theorem 9. Assume that the reduced cost function, i.e., $r(\xi; \alpha, \boldsymbol{\beta}) \equiv Q(\xi) - \alpha - \sum_{i=1}^{m} \psi_i(\xi)\beta_i$, is continuous, uniformly with respect to ξ over Ξ for $\boldsymbol{\beta} \ge \mathbf{0}$. Then, *Oracle-2* returns an optimal or ε-optimal solution of **WCEV** problem in finite iterations. □

Proof. According to the assumption, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|r(\xi_1; \alpha, \boldsymbol{\beta}) - r(\xi_2; \alpha, \boldsymbol{\beta})| \le \varepsilon \quad \text{if} \quad \|\xi_1 - \xi_2\| \le \delta \quad \text{for } \xi_1, \xi_2 \in \Xi, \boldsymbol{\beta} \ge \mathbf{0}.$$

Let $B(\xi_1, \delta)$ denote the closed ball with radius δ at $\xi_1 \in \Xi$, i.e., $B(\xi_1, \delta) = \{\xi_2 \in \Xi, \|\xi_1 - \xi_2\| \le \delta\}$. We consider the following claim.

Claim: In any particular iteration with associated $\boldsymbol{\xi}^0$, *Oracle-2* either solves **WCEV** problem to an optimal or ε-optimal solution, or produces a new scenario ξ^* that is not contained in ball $B(\xi, \delta)$ for any $\xi \in \boldsymbol{\xi}^0$.

Proof of Claim: Recall $(\alpha^*, \boldsymbol{\beta}^*)$ denotes the shadow price obtained from computing **PMP**. We prove this claim by contradiction.

Suppose that scenario ξ^* with $r(\xi^*; \alpha^*, \boldsymbol{\beta}^*)$ is identified by **PSP**. Note that it is sufficient to assume that $r(\xi^*; \alpha^*, \boldsymbol{\beta}^*) > \varepsilon$, since otherwise it is an optimal or ε-optimal solution of **WCEV** problem by Proposition 7. Assume further that it is contained in ball $B(\xi', \delta)$ for some $\xi' \in \boldsymbol{\xi}^0$. We have

$$\begin{aligned} v^*(\boldsymbol{\xi}^0) =& r(\xi^*; \alpha^*, \boldsymbol{\beta}^*) \\ &= r(\xi^*; \alpha^*, \boldsymbol{\beta}^*) - r(\xi'; \alpha^*, \boldsymbol{\beta}^*) + r(\xi'; \alpha^*, \boldsymbol{\beta}^*) \end{aligned}$$

$$
\begin{aligned}
&= |r(\xi^*;\alpha^*,\boldsymbol{\beta}^*) - r(\xi';\alpha^*,\boldsymbol{\beta}^*)| + r(\xi';\alpha^*,\boldsymbol{\beta}^*)\\
&\leq \varepsilon + r(\xi';\alpha^*,\boldsymbol{\beta}^*)\\
&\leq \varepsilon.
\end{aligned}
$$

The second equation holds due to an identity transformation, the third follows from the fact that $r(\xi';\alpha^*,\boldsymbol{\beta}^*) \leq 0$ and the first inequality follows the definition of uniformly continuous function. The last inequality, which is valid due to the fact that $r(\xi;\alpha^*,\boldsymbol{\beta}^*) \leq 0$ for $\xi \in \boldsymbol{\xi}^0$, clearly contradicts to our first assumption.

With the aforementioned contradiction, we can conclude that either $\boldsymbol{\xi}^0$ (and the associated probabilities) is an optimal or ε-optimal solution of **WCEV** problem, or ξ^* is not contained in ball $B(\xi,\delta)$ for any $\xi \in \boldsymbol{\xi}^0$. □

It simply follows from the claim that the distance between any two scenarios in Ξ^0 is more than δ, indicating the two balls centered at them with $\frac{\delta}{2}$-radius are disjoint. Let $V(\frac{\delta}{2})$ denote the volume of such a ball. Given that Ξ is compact, it is clear that Ξ is contained in a ball centered at some $\xi^0 \in \Xi$ with radius $\hat{d}$. Consequently, $B(\xi^0, \hat{d} + \frac{\delta}{2})$ is compact, and its volume, denoted by $V(\hat{d} + \frac{\delta}{2})$, is finite. Because $V(\hat{d} + \frac{\delta}{2})/V(\frac{\delta}{2})$ is finite, it follows that an optimal or ε-optimal solution of **WCEV** problem can be found by *Oracle-2* within a finite number of iterations, bounded by $\left\lceil V(\hat{d} + \frac{\delta}{2})/V(\frac{\delta}{2}) \right\rceil$. □

Next, we consider some special case where an optimal solution is guaranteed to be obtained within finite iterations.

Corollary 10. Assume that Ξ is a polytope, $Q(\xi)$ is convex and $\psi_i(\xi)$ are concave over Ξ for $i \in [m]$. Let $\mathsf{XV}(\Xi)$ be the set of extreme points of Ξ. Then, *Oracle-2* terminates with an optimal solution to **WCEV** problem within $|\mathsf{XV}(\Xi)|$ iterations. □

Proof. We note that, under the aforementioned assumptions, an optimal solution to **PSP** in (11) is an extreme point of Ξ, i.e., it belongs to $\mathsf{XV}(\Xi)$. Moreover, it can be seen from the proof of Theorem 9 that a new extreme point of Ξ, which is not contained in $\boldsymbol{\xi}^0$, will be identified and then used to update $\boldsymbol{\xi}^0$ in each iteration before *Oracle-2* terminates.

Given that $\mathsf{XV}(\Xi)$ is a finite set, the expected conclusion follows. □

Remark 5. (*i*) Actually, results of Theorem 9 can be extended to handle some discontinuous $r(\xi;\alpha,\boldsymbol{\beta})$. For example, as long as $r(\xi;\alpha,\boldsymbol{\beta})$ is uniformly continuous within each subset of a finite partition of Ξ for $\boldsymbol{\beta} \geq \mathbf{0}$, where Proposition 5 does not hold, an ε-optimal solution for **WCEV** problem can still be found within finite iterations.

(*ii*) As discussed, $\psi_i(\cdot)$ functions are continuous over Ξ for ambiguity sets defined by moment inequalities and Wasserstein metric, which render Theorem 9 to hold when $Q(\cdot)$ guarantees $r(\xi;\alpha,\boldsymbol{\beta})$ uniformly continuous over Ξ for $\boldsymbol{\beta} \geq \mathbf{0}$. Moreover, in the case where Ξ is a polytope and $\mathcal{P}$ involves the first moment inequalities, the second moment inequalities subject to lower bounds, or the Wasserstein metric restriction defined with L_1 norm, optimal ξ^* of **PSP**, which is to maximize a convex function, can always be obtained at some extreme point of Ξ. According to Corollary 10, an optimal solution for **WCEV** problem can be obtained within finite iterations. □

In the next section, we seek to compute the whole **2 – Stg DRO**, and present a decomposition algorithm that incorporates those primal oracles for **WCEV** problem.

4 From Computing the WCEV to Solving 2 – Stg DRO

In this section, we present a generally applicable and computationally strong solution scheme to solve **2 – Stg DRO**, along with theoretical analyses on its strength, convergence and computational complexity. Compared to all existing methods in the literature, it enhances our understanding on DRO and significantly expands and improves our ability to solve practical instances, particularly those with the challenging infeasible recourse issue.

We highlight that the basic intuitions behind our algorithm design are: (*i*) For a set of given scenarios, the worst-case probability distribution can be readily determined by solving an LP; (*ii*) Together with the first stage decision making, a simple bilevel optimization problem can be constructed to obtain a lower-bound approximation of the original **2 – Stg DRO**; (*iii*) If such lower-bound approximation is not satisfactory, additional scenarios needed to determine the WCEV can be identified by solving **WCEV** problem and then are employed to expand that scenario set through the framework of C&CG.

4.1 Integrating Primal Oracles within C&CG

As noted, different from the current mainstream strategies on solving DRO, our algorithm development takes the primal perspective: it integrates the oracles for **WCEV** problem discussed in the previous section within the C&CG framework, resulting in a simple and intuitive algorithmic structure that helps its adoption among practitioners.

We say that two formulations are *equivalent* to each other if they share the same optimal value, and one's optimal first-stage solution is also optimal to the other one. With this definition and by the results in Corollary 4 and Proposition 5, two equivalent reformulations can be obtained for **2 – Stg DRO**.

Proposition 11. For **2 – Stg DRO** in (1), we have

$$
\begin{aligned}
w^* &= \min\Big\{f_1(\mathbf{x}) + \eta : \mathbf{x} \in \mathcal{X},\ \eta \geq \lim_{n\to+\infty} \max_{\xi_j(n)\in\Xi, j\in[n]} \max\{\sum_{j=1}^{n} Q(\mathbf{x}, \xi_j(n)) p_j : P \in \mathcal{P}\}\Big\} \\
&= \min\Big\{f_1(\mathbf{x}) + \eta : \mathbf{x} \in \mathcal{X},\ \eta \geq \max\Big\{ \sum_{j=1}^{m+1} Q(\mathbf{x}, \xi_j) p_j : P \in \mathcal{P}, \xi_j \in \Xi \ \ \forall j \in [m+1]\Big\}\Big\}.
\end{aligned}
$$
□

Both equivalent reformulations can be treated as bilevel optimization problems with a non-convex lower-level maximization problem. Note that no closed-form is available to characterize the optimal solution set of such lower-level problem to build a single-level reformulation further. Indeed, they can easily yield lower bounds to w^* if we reduce the lower-level problems to simpler ones, e.g., LPs. Let $\mathbf{x}^*$ denote a fixed first-stage solution. The resulting lower bound we obtain might not be strong enough to substantiate $\mathbf{x}^*$'s quality, which may demand us to revise our simplification. Next, we show that, by employing the oracles developed for **WCEV** problem and the C&CG scheme, a complete algorithmic scheme for **2 – Stg DRO** is developed, which ensures an optimal and feasible $\mathbf{x}^*$ can be produced.

4.1.1 Addressing the Challenge of Infeasible Recourse

By convention, we set $Q(\mathbf{x}, \xi) = +\infty$ if $\mathcal{Y}(\mathbf{x}, \xi) = \varnothing$, i.e., the recourse problem in (2) is infeasible. Also, let $\mathcal{E}$ denote the event when $\mathcal{Y}(\mathbf{x}, \xi) = \varnothing$ (or equivalently $Q(\mathbf{x}, \cdot) = +\infty$). Then, by extending the feasibility definition from the deterministic context, we say a first-stage decision $\mathbf{x}$ is *almost surely* feasible to **2 – Stg DRO** if $\sup_{P \in \mathcal{P}} P(\mathcal{E}) = 0$, i.e., the probability of this event $\mathcal{E}$ is 0 for $P \in \mathcal{P}$. To implement this concept in computation, we introduce a supporting problem, whose optimal value corresponds to the status of $\mathcal{Y}(\mathbf{x}, \xi)$. Specifically, we assume that $\mathcal{Y}(\mathbf{x}, \xi) \equiv \left\{\mathbf{y} : \mathbf{g}(\mathbf{x}, \xi, \mathbf{y}) \geq \mathbf{0}\right\}$ and define

$$\tilde{\mathcal{Y}}(\mathbf{x}, \xi) \equiv \left\{(\mathbf{y}, \tilde{\mathbf{y}}) : \mathbf{g}(\mathbf{x}, \xi, \mathbf{y}) + \tilde{\mathbf{y}} \geq \mathbf{0},\ \tilde{\mathbf{y}} \geq \mathbf{0}\right\}, \tag{14}$$

with artificial variable $\tilde{\mathbf{y}}$. Note that $\tilde{\mathcal{Y}}(\mathbf{x}, \xi) \neq \varnothing$ regardless of $\mathbf{x}$ and ξ. With $\|\cdot\|_1$ denoting the L_1 norm, the supporting problem is

$$\tilde{Q}_f(\mathbf{x}, \xi) = \min \left\{\|\tilde{\mathbf{y}}\|_1 : (\mathbf{y}, \tilde{\mathbf{y}}) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi)\right\}. \tag{15}$$

Since (15) is to minimize the L_1 norm of $\tilde{\mathbf{y}}$, the next result follows easily.

Lemma 12. $\tilde{Q}_f(\mathbf{x}, \xi) > 0$ iff $Q(\mathbf{x}, \xi) = +\infty$. Moreover, $\mathbf{x}$ is almost surely feasible to **2 – Stg DRO** iff the optimal value of the following problem equals 0.

$$\mathbf{WCEV}(F): \max_{P \in \mathcal{P}} E_P[\tilde{Q}_f(\mathbf{x}, \xi)] \equiv \max \left\{ \int_{\Xi} \tilde{Q}_f(\mathbf{x}, \xi) P(d\xi) : P \in \mathcal{P}\right\}. \tag{16}$$

As a result, the worst-case expected value of $\tilde{Q}_f(\mathbf{x}, \xi)$ can be used to verify the (almost surely) feasibility of $\mathbf{x}$, and hence is denoted by $\mathbf{WCEV}(F)$. To keep our notation consistent, the one presented in (4), after replacing $Q(\xi)$ by $Q(\mathbf{x}, \xi)$, is interchangeably referred to as $\mathbf{WCEV}(O)$ as it computes the worst-case expected recourse cost. We mention that when a mixed integer ambiguity set, denoted by $\mathcal{P}^I$, is adopted as the ambiguity set, Lemma 12 still holds. Actually, both $\mathbf{WCEV}(O)$ and $\mathbf{WCEV}(F)$, as well as their extensions on $\mathcal{P}^I$, can be computed by the primal oracles developed in the previous section.

Remark 6. When the optimal value of $\mathbf{WCEV}(F)$ equals 0, it has an intuitive interpretation. That is, the probability over the subset of Ξ satisfying $\tilde{Q}_f(\mathbf{x}, \xi) = 0$ equals 1 and the probability over its complement 0 for any distribution within $\mathcal{P}$. When the optimal value of $\mathbf{WCEV}(F)$ is larger than 0, it means that some scenarios become infeasible for the given $\mathbf{x}$ and they have positive probabilities under some distribution in $\mathcal{P}$. For the latter case, the scenario set generated by a primal oracle should be used within C&CG in a way such that the current $\mathbf{x}$ is cut off, i.e., being excluded from future considerations. Note that in the remainder of this paper, we say $\mathbf{x}$ is feasible means that it is almost surely feasible. □

4.1.2 Primary Components of the C&CG Method for 2 – Stg DRO

The C&CG method for **2 – Stg DRO**, which is referred to as C&CG-DRO, involves a master problem and two subproblems. The master problem is a relaxation of **2 – Stg DRO** and yields a lower bound. Two subproblems, which are for feasibility and optimality, respectively, help us strengthen the master problem and derive an upper bound.

(I): **The Master Problem of C&CG-DRO**

The master problem is defined on $\widehat{\boldsymbol{\xi}}^o$ and $\widehat{\boldsymbol{\xi}}^f$, two sets of fixed scenarios in Ξ. To differentiate it from the master problem of *Oracle-2*, we call it the *main master problem* (**MMP**). In the following we first present a form developed using big-M technique, which is rather intuitive and interpretable. Then, with deep insights, we derive a big-M-free one that is compact and more rigorous. Hence, they are referred to as $\mathbf{MMP}_1$ and $\mathbf{MMP}_2$, respectively.

$$\mathbf{MMP}_1 : \underline{w} = \min_{\mathbf{x}\in\mathcal{X}} f_1(\mathbf{x}) + \eta \tag{17a}$$

$$\eta \geq \max\Big\{ \sum_{\xi\in\widehat{\boldsymbol{\xi}}^o} \eta^o_\xi p^o_\xi : (p^o_\xi)_{\xi\in\widehat{\boldsymbol{\xi}}^o} \in \mathcal{P}\Big\} \tag{17b}$$

$$\Big\{\eta^o_\xi = f_2(\mathbf{x},\xi,\mathbf{y}_\xi) + \mathbf{M}\|\tilde{\mathbf{y}}^o_\xi\|_1,\ (\mathbf{y}_\xi,\tilde{\mathbf{y}}^o_\xi) \in \tilde{\mathcal{Y}}(\mathbf{x},\xi)\Big\} \quad \forall\xi\in\widehat{\boldsymbol{\xi}}^o \tag{17c}$$

$$0 \geq \max\Big\{ \sum_{\xi\in\widehat{\boldsymbol{\xi}}^f} \eta^f_\xi p^f_\xi : (p^f_\xi)_{\xi\in\widehat{\boldsymbol{\xi}}^f} \in \mathcal{P}\Big\} \tag{17d}$$

$$\Big\{\eta^f_\xi = \|\tilde{\mathbf{y}}^f_\xi\|_1,\ (\mathbf{y}_\xi,\tilde{\mathbf{y}}^f_\xi) \in \tilde{\mathcal{Y}}(\mathbf{x},\xi)\Big\} \quad \forall\xi\in\widehat{\boldsymbol{\xi}}^f \tag{17e}$$

We mention that in (17c) $\tilde{\mathcal{Y}}(\cdot,\cdot)$ is employed, instead of original $\mathcal{Y}(\cdot,\cdot)$, and the sum of f_2 and big-M penalized $\|\tilde{\mathbf{y}}^o_\xi\|_1$ is assigned to η^o_ξ. It is worth highlighting that these components are fundamental in solving **2 – Stg DRO** when the feasibility issue arises in the recourse problem. Consider a scenario ξ^o in $\widehat{\boldsymbol{\xi}}^o$. On one hand, if p^o_ξ equals 0, i.e., ξ^o virtually does not occur, it is not necessary to have the corresponding $\mathcal{Y}(\cdot,\xi^o)$ being non-empty. Or, alternatively the choice of $\mathbf{x}$ should not be affected by $\mathcal{Y}(\cdot,\xi^o)$. On the other hand, if p^o_ξ is positive, indicating the occurrence of ξ^o cannot be ruled out, $\mathbf{x}$ must render $\mathcal{Y}(\cdot,\xi^o)$ non-empty to ensure a feasible recourse. Or, alternatively, those causing $\mathcal{Y}(\cdot,\xi^o)$ empty cannot be in any optimal solution of $\mathbf{MMP}_1$. Those two points and the logic behind, which are the nature of DRO, can be well achieved by making use of artificial variables $\tilde{\mathbf{y}}^o_\xi$ and big-M coefficient $\mathbf{M}$ in (17c).

Note that when $p^o_\xi = 0$, η in (17b) is not affected by η^o_ξ, and $\tilde{\mathcal{Y}}(\mathbf{x},\xi^o) \neq \emptyset$ for any $\mathbf{x}$.

Hence, we can conclude that ξ^o has no impact on $\mathbf{MMP}_1$. Also, when $p^o_\xi > 0$, the big-M penalty term in η^o_ξ will drive $\mathbf{MMP}_1$ to select an $\mathbf{x}$ such that $\tilde{\mathbf{y}}^o_\xi$ can be 0, i.e., $\mathcal{Y}(\mathbf{x}, \xi^o) \neq \emptyset$. As a result, (17c) will disqualify $\mathbf{x}$ in any optimal solution of $\mathbf{MMP}_1$ if it causes the recourse problem of ξ^o infeasible. Certainly, if the recourse problem is assumed to be feasible for any $\mathbf{x} \in \mathcal{X}$ and $\xi \in \Xi$, it is not needed to introduce variable $\tilde{\mathbf{y}}$ and set $\tilde{\mathcal{Y}}$.

Remark 7. As the two inequalities in (17b) and (17d) refine the feasible set of $\mathbf{x}$ from optimality and feasibility perspectives, we refer to them as *optimality* and *feasibility* cutting planes, and $\widehat{\boldsymbol{\xi}^f}$ and $\widehat{\boldsymbol{\xi}^o}$ feasibility and optimality sets, respectively. Note that different from most of classical cutting plane methods that generate new and independent cutting planes over iterations, $\mathbf{MMP}_1$ strengthens either the *optimality* or the *feasibility* cutting plane based on the augmented $\widehat{\boldsymbol{\xi}^o}$ or $\widehat{\boldsymbol{\xi}^f}$ in every iteration. □

By further investigating the critical interaction between a scenario and the associated probability, we actually can leverage it to eliminate the reliance on big-M. Similar to the unification idea presented in [22] and [34], the two sets of scenarios, for optimality and feasibility, respectively, can be merged to have a unified set. By doing so, we let $\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}^f} \cup \widehat{\boldsymbol{\xi}^o}$, through which we can build the following alternative formulation of $\mathbf{MMP}_1$.

$$\mathbf{MMP}_2 : \underline{w} = \min_{\mathbf{x} \in \mathcal{X}} f_1(\mathbf{x}) + \eta \tag{18a}$$

$$\eta \geq \max \left\{ \sum_{\xi \in \widehat{\boldsymbol{\xi}}} \eta^o_\xi p^o_\xi : (p^o_\xi)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P} \right\} \tag{18b}$$

$$\eta^o_\xi = f_2(\mathbf{x}, \xi, \mathbf{y}_\xi) \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{18c}$$

$$0 \geq \max \left\{ \sum_{\xi \in \widehat{\boldsymbol{\xi}}} \eta^f_\xi p^f_\xi : (p^f_\xi)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P} \right\} \tag{18d}$$

$$\eta^f_\xi = \|\tilde{\mathbf{y}}_\xi\|_1 \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{18e}$$

$$(\mathbf{y}_\xi, \tilde{\mathbf{y}}_\xi) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi) \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{18f}$$

Proposition 13. $\mathbf{MMP}_2$ is a valid relaxation and provides a lower bound to $\mathbf{2-Stg\ DRO}$.

Proof. We first extend $\mathbf{MMP}_1$ by considering $\widehat{\boldsymbol{\xi}}$ that leads to the following auxiliary formulation.

$$\mathbf{MMP}^a_1 : \underline{w} = \min_{\mathbf{x} \in \mathcal{X}} f_1(\mathbf{x}) + \eta \tag{19a}$$

$$\eta \geq \max \left\{ \sum_{\xi \in \widehat{\boldsymbol{\xi}}} \eta^o_\xi p^o_\xi : (p^o_\xi)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P} \right\} \tag{19b}$$

$$\left\{ \eta^o_\xi = f_2(\mathbf{x}, \xi, \mathbf{y}^o_\xi) + \mathbf{M}\|\tilde{\mathbf{y}}^o_\xi\|_1, \ (\mathbf{y}^o_\xi, \tilde{\mathbf{y}}^o_\xi) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi) \right\} \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{19c}$$

$$0 \geq \max \left\{ \sum_{\xi \in \widehat{\boldsymbol{\xi}}} \eta^f_\xi p^f_\xi : (p^f_\xi)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P} \right\} \tag{19d}$$

$$\left\{ \eta^f_\xi = \|\tilde{\mathbf{y}}^f_\xi\|_1, \ (\mathbf{y}^f_\xi, \tilde{\mathbf{y}}^f_\xi) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi) \right\} \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{19e}$$

Since $\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}^f} \cup \widehat{\boldsymbol{\xi}}^o \subseteq \Xi$, $\mathbf{MMP}_1^a$ is a valid relaxation that actually is stronger than $\mathbf{MMP}_1$.

For $\mathbf{MMP}_1^a$, let $\{p_\xi^{o*}\}_{\xi \in \widehat{\boldsymbol{\xi}}}$ denote an optimal solution to the LP in (19b). Considering some $\xi' \in \widehat{\boldsymbol{\xi}}$, we note that (19d) and (19e) will drive $\tilde{\mathbf{y}}_{\xi'}^o$ to be 0 if $p_{\xi'}^{o*} > 0$. If this is not the case, we let $\left(p_{\xi'}^f, \tilde{\mathbf{y}}_{\xi'}^f\right) = \left(p_{\xi'}^{o*}, \tilde{\mathbf{y}}_{\xi'}^o\right)$. Then, we have

$$\max\Big\{\sum_{\xi \in \widehat{\boldsymbol{\xi}}} \eta_\xi^f p_\xi^f : (p_\xi^f)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P}\Big\} \geq \eta_{\xi'}^f p_{\xi'}^f = \|\tilde{\mathbf{y}}_{\xi'}^f\|_1 p_{\xi'}^f > 0,$$

which is contradictory to inequality (19d). Given the arbitrarity of ξ', we can simply remove $\mathbf{M}$ in (19c) without affecting the optimality of $\mathbf{MMP}_1^a$. The updated (19c) is

$$\Big\{\eta_\xi^o = f_2(\mathbf{x}, \xi, \mathbf{y}_\xi^o),\ (\mathbf{y}_\xi^o, \tilde{\mathbf{y}}_\xi^o) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi)\Big\} \quad \forall \xi \in \widehat{\boldsymbol{\xi}}. \tag{20}$$

Next, we prove that two sets of $(\mathbf{y}_\xi, \tilde{\mathbf{y}}_\xi)$ variables are not necessary. Let $\left(\mathcal{P}^a, \widehat{\mathcal{Y}}^a(\mathbf{x})\right)$ denote the feasible set defined by (19d) and (19e), i.e.,

$$\left(\mathcal{P}^a, \widehat{\mathcal{Y}}^a(\mathbf{x})\right) = \Big\{0 \geq \max\big\{\sum_{\xi \in \widehat{\boldsymbol{\xi}}} \eta_\xi^f p_\xi^f : (p_\xi^f)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P}\big\}$$
$$\big\{\eta_\xi^f = \|\tilde{\mathbf{y}}_\xi^f\|_1,\ (\mathbf{y}_\xi^f, \tilde{\mathbf{y}}_\xi^f) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi)\big\} \quad \forall \xi \in \widehat{\boldsymbol{\xi}}\Big\}.$$

Claim: There exists an optimal solution to $\mathbf{MMP}_1^a$ such that its component, $(p_\xi^{o*}, \mathbf{y}_\xi^{o*}, \tilde{\mathbf{y}}_\xi^{o*})_{\xi \in \widehat{\boldsymbol{\xi}}}$, belongs to $\left(\mathcal{P}^a, \widehat{\mathcal{Y}}^a(\mathbf{x})\right)$.

Proof of Claim: Suppose it is not true. So, there exists at least one scenario $\xi' \in \widehat{\boldsymbol{\xi}}$ such that $\eta_{\xi'}^{f*} p_{\xi'}^{f*} > 0$. Again, we let $\left(p_{\xi'}^{o*}, \mathbf{y}_{\xi'}^{o*}, \tilde{\mathbf{y}}_{\xi'}^{o*}\right) = \left(p_{\xi'}^{f*}, \mathbf{y}_{\xi'}^{f*}, \tilde{\mathbf{y}}_{\xi'}^{f*}\right)$. Then, for (19c) we have

$$\eta_{\xi'}^{o*} p_{\xi'}^{o*} = p_{\xi'}^{o*}\Big(f_2(\mathbf{x}, \xi', \mathbf{y}_\xi^{o*}) + \mathbf{M}\tilde{\mathbf{y}}_{\xi'}^{o*}\Big) > p_{\xi'}^{o*} f_2(\mathbf{x}, \xi', \mathbf{y}_\xi^{o*}),$$

which is contradictory to the updated (19c), i.e., (20). □

Hence, $(\mathbf{y}_\xi^o, \tilde{\mathbf{y}}_\xi^o)$ and $(\mathbf{y}_\xi^f, \tilde{\mathbf{y}}_\xi^f)$ can be unified into $(\mathbf{y}_\xi, \tilde{\mathbf{y}}_\xi)$ without affecting the optimality of $\mathbf{MMP}_1^a$. By doing so, $\mathbf{MMP}_1^a$ is equivalent to $\mathbf{MMP}_2$. □

Remark 8. (*i*) Note that "=" in (17c) or (17e) can be changed to "≤" without affecting the optimality of **MMP**, which may lead to some computational improvement when $|\widehat{\boldsymbol{\xi}^o}|$ or $|\widehat{\boldsymbol{\xi}^f}|$ is large. More importantly, both $\mathbf{MMP}_1$ and $\mathbf{MMP}_2$ are bilevel optimization formulations with two lower-level problems for feasibility and optimality, respectively. For $\mathcal{P}$ defined in (3), the lower-level problems in (17b) and (17d) (or in (18b) and (18d)) are LPs and can be directly replaced by their optimality conditions or dual problems as shown in Appendix A.2. Actually, by making use of their specific structures and their dual problems, simpler single-level equivalent reformulations can be obtained. Consider $\mathbf{MMP}_2$ for demonstration, with

$(\alpha^f, \boldsymbol{\beta}^f)$ and $(\alpha^o, \boldsymbol{\beta}^o)$ denoting dual variables of constraints of $\mathcal{P}$ in (18b) and (18d).

$$
\begin{aligned}
\underline{w} = \min \Big\{ f_1(\mathbf{x}) + \eta : \ & \mathbf{x} \in \mathcal{X},\ \eta \geq \alpha^o + \sum_{i=1}^{m} \gamma_i \beta_i^o,\ \ \alpha^o + \sum_{i=1}^{m} \psi_i(\xi)\beta_i^o \geq \eta_\xi^o \ \ \forall \xi \in \widehat{\boldsymbol{\xi}}, \\
& 0 \geq \alpha^f + \sum_{i=1}^{m} \gamma_i \beta_i^f,\ \ \alpha^f + \sum_{i=1}^{m} \psi_i(\xi)\beta_i^f \geq \eta_\xi^f \ \ \forall \xi \in \widehat{\boldsymbol{\xi}}, \\
& (18c),\ (18e),\ (18f),\ \beta_i^o \geq 0,\ \ \forall i \in [m],\ \beta_i^f \geq 0 \ \ \forall i \in [m] \Big\}.
\end{aligned}
\tag{21}
$$

We mention that (21) is a big-M-free single-level formulation. If f_1 and f_2 are linear and $\mathcal{X}$ and $\mathcal{Y}(\cdot,\cdot)$ are linearly representable, it is a mixed integer linear program that is rather tractable by state-of-the-art professional solvers.

(*ii*) For $\mathbf{MMP}_2$, it is not necessary to define the optimality cutting plane with respect to the whole set $\widehat{\boldsymbol{\xi}}$. It is worth noting from our numerical studies that $|\widehat{\boldsymbol{\xi}^f}|$ could be more than an order of magnitude greater than $|\widehat{\boldsymbol{\xi}^o}|$. Since the feasibility issue has not been analyzed in the existing literature, such a huge difference is very new and unexpected. Hence, rather than employing $\widehat{\boldsymbol{\xi}}$ in both optimality and feasibility cutting planes for consistency, it would be computationally more effective to build the first one with respect to $\widehat{\boldsymbol{\xi}}^o$ only. When computing large-scale instances, the benefit of this modification is often obvious.

(*iii*) If the recourse problem is assumed to be feasible all the time, (21) reduces to the duality based master problem appeared in the literature (e.g., [9, 16, 25]). Yet, its derivation from **MMP** is intuitive, involves the knowledge of LP only, and more importantly, allows us to consider sophisticated $\mathcal{P}$ using mature optimization techniques. Unless specified otherwise, (21) is utilized as **MMP** in our algorithm implementation. □

(II): **Subproblems of C&CG-DRO**

The two subproblems are referred to as main subproblems for consistency. As mentioned after Lemma 12, the feasibility one is **WCEV**(F) in (16), and the optimality one **WCEV**(O) in (4) (with $Q(\xi)$ replaced by $Q(\mathbf{x}, \xi)$). Because they can be solved by the oracles developed in Section 3.1 in an almost identical fashion, we next present the customization of those oracles on **WCEV**(O) only. Specifically, for given $\mathbf{x}^*$, the finite mathematical program for **WCEV**(O) can be easily obtained by modifying (7) as in the following.

$$
\begin{aligned}
\mathbf{WCEV}(O) - \mathbf{FMP}: \ \eta^o(\mathbf{x}^*) = \max \Bigg\{ & \sum_{j=1}^{m+1} \eta_j p_j : (p_1, \ldots, p_{m+1}) \in \mathcal{P}, \\
\xi_j \in \Xi \ \ \forall j \in [m+1],\ & \Big\{ \eta_j = \min \big\{ f_2(\mathbf{x}^*, \xi_j, \mathbf{y}_j), \mathbf{y}_j \in \mathcal{Y}(\mathbf{x}^*, \xi_j) \big\} \Big\} \ \ \forall j \in [m+1] \Bigg\}.
\end{aligned}
\tag{22}
$$

Comparing (7) and (22), the basic difference is that variable η_j replaces $Q(\xi_j)$, and η^j is set to the recourse cost for ξ_j through a replica of the recourse problem parameterized by ξ_j for all j. Note that the equality sign associated with η_j can be changed to $\leq$ without affecting the whole formulation's optimality.

If $\mathbf{WCEV}(O)$ is solved by the CG-based *Oracle-2*, the customized **PMP**, denoted by $\mathbf{PMP}(O)$, is a simple LP with $Q(\xi_j^0)$ replaced by $Q(\mathbf{x}^*, \xi_j^0)$ in (10). The customized **PSP**, denoted by $\mathbf{PSP}(O)$, is in the following format.

$$\mathbf{PSP}(O): \ v^*(\mathbf{x}^*, \boldsymbol{\xi}_n^0) = \max_{\xi \in \Xi} \min_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^*, \xi)} f_2(\mathbf{x}^*, \xi, \mathbf{y}) - \alpha^* - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^*. \tag{23}$$

Both (22) and (23) are bilevel optimization formulations that can be solved by the methods presented in Appendix A.2. To unify our exposition, regardless of using *Oracle-1* or *Oracle-2* to solve $\mathbf{WCEV}(O)$ and $\mathbf{WCEV}(F)$, we let $\eta^o(\mathbf{x}^*)$ denote the optimal value of $\mathbf{WCEV}(O)$, $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*) \equiv \{\xi_1^o, \ldots, \xi^o_{|\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)|}\}$ and $P^o\big(\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)\big)$ the set of resulting scenarios and their probabilities; and $\eta^f(\mathbf{x}^*)$, $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ and $P^f\big(\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)\big)$ to their counterparts for $\mathbf{WCEV}(F)$.

Remark 9. As noted previously, when the recourse problem is an LP and there is no feasibility issue $\mathbf{x} \in \mathcal{X}$, Benders type of algorithms have been developed to solve $\mathbf{2-Stg\ DRO}$ [15, 17, 18, 19, 23]. With the primal oracles for $\mathbf{WCEV}(O)$ and resulting $\hat{\boldsymbol{\xi}}^o$, they naturally can be extended for possible enhancements. Specifically, let the recourse problem be

$$\min \left\{\mathbf{c}_2\mathbf{y}: \ \mathbf{B}\mathbf{y} \geq \mathbf{b}_2 - \mathbf{A}_2\mathbf{x} - \mathbf{H}\xi\right\}. \tag{24}$$

Note that for $\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$ we can derive an optimal solution for the dual problem of (24). Denoting them by $\{\pi_\xi^*\}_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)}$, Benders cutting planes, generated after solving the current $\mathbf{WCEV}(O)$, are presented in the following.

$$\alpha^o + \sum_i \psi_i(\xi)\beta_i^o \geq \eta_\xi^o \quad \forall \xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$$
$$\eta_\xi^o \geq (\mathbf{b}_2 - \mathbf{A}_2\mathbf{x} - \mathbf{H}\xi)^\intercal \pi_\xi^* \quad \forall \xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*) \qquad \square$$

Indeed, although it has not been investigated in the current literature, Benders cutting planes to address the feasibility issue can be produced in the same fashion after solving $\mathbf{WCEV}(F)$. To differentiate from existing implementations and to be consistent, we refer to our new approach as Benders-DRO and the previous ones as basic Benders. Actually, as shown in the numerical study, Benders-DRO performs drastically better than basic Benders.

4.2 Complete C&CG Procedure for $\mathbf{2-Stg\ DRO}$

With the aforementioned problems defined, we are ready to present the overall procedure of C&CG method customized to compute $\mathbf{2-Stg\ DRO}$. Note that LB and UB denote lower and upper bounds, respectively, TOL is the optimality tolerance, and t is the counter for iterations. Also, to facilitate our understanding, we sketch the logic and main steps in Figure 1.

Algorithm 1: C&CG-DRO

Step 1 Set $LB = -\infty$, $UB = +\infty$, $t = 1$, and $\widehat{\boldsymbol{\xi}}^o = \widehat{\boldsymbol{\xi}}^f = \widehat{\boldsymbol{\xi}} = \emptyset$.

Step 2 Solve master problem **MMP**. If it is infeasible, report the infeasibility of **2 − Stg DRO** and terminate. If it is unbounded, select an arbitrary new feasible solution $\mathbf{x}^*$. Otherwise, derive optimal value $\underline{w}$ and solution $\mathbf{x}^*$, and update $LB = \underline{w}$.

Step 3 Solve feasibility subproblem **WCEV**(F), derive its optimal value $\eta^f(\mathbf{x}^*)$, optimal set of scenarios $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ and associated probabilities (by one of oracles from Section 3).

Step 4 Cases based on $\eta^f(\mathbf{x}^*)$

Case A $\eta^f(\mathbf{x}^*) = 0$

(i) Solve optimality subproblem **WCEV**(O), derive its optimal value $\eta^o(\mathbf{x}^*)$, optimal set of scenarios $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$ and their probabilities (by one of oracles from Section 3);

(ii) Update $\widehat{\boldsymbol{\xi}}^o = \widehat{\boldsymbol{\xi}}^o \cup \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$ (and accordingly $\widehat{\boldsymbol{\xi}}$), and augment **MMP** with new variables and constraints accordingly.

Case B $\eta^f(\mathbf{x}^*) > 0$

(i) Update $\widehat{\boldsymbol{\xi}}^f = \widehat{\boldsymbol{\xi}}^f \cup \hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ (and accordingly $\widehat{\boldsymbol{\xi}}$), and augment **MMP** with new variables and constraints accordingly;

(ii) set $\eta^o(\mathbf{x}^*) = +\infty$.

Step 5 Update $UB = \min\{UB, f_1(\mathbf{x}^*) + \eta^o(\mathbf{x}^*)\}$.

Step 6 If $UB - LB \leq TOL$, return $\mathbf{x}^*$ and terminate. Otherwise, set $t = t + 1$ and go to Step 2.

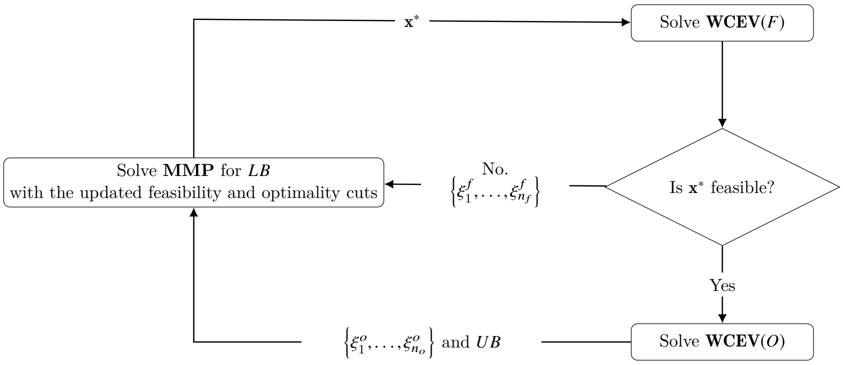

$\mathbf{x}^*$
Solve **WCEV**(F)
Is $\mathbf{x}^*$ feasible?
No.
$\left\{\xi_1^f, \ldots, \xi_{n_f}^f\right\}$
Solve **MMP** for LB
with the updated feasibility and optimality cuts
Yes
Solve **WCEV**(O)
$\left\{\xi_1^o, \ldots, \xi_{n_o}^o\right\}$ and UB

Figure 1: The Schematic Flow of C&CG-DRO

We note that Algorithm 1 is a rather vanilla version of C&CG-DRO. Some simple changes, as described in the following, often can yield significantly computational improvements, especially when subproblems are computed by CG-based *Oracle-2*.

(C1) Rather than letting *Oracle-2* start from scratch, $\widehat{\boldsymbol{\xi}}^f$ and $\widehat{\boldsymbol{\xi}}^o$ obtained up to date can be used as the initial scenario sets for it to solve **WCEV**(F) and **WCEV**(O), respectively. If this is the case, the following operation should be included between Step 2 and Step 3.

– **Step 2.a** For $\xi^f \in \widehat{\boldsymbol{\xi}}^f$, compute $\tilde{Q}(\mathbf{x}^*, \xi^f)$; and for $\xi^o \in \widehat{\boldsymbol{\xi}}^o$, compute $Q(\mathbf{x}^*, \xi^o)$.

With this change, we can simplify **Step 4** as in the following.

– In **Step 4-Case A** and **Step 4-Case B**, update $\widehat{\boldsymbol{\xi}^f} = \hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ and $\widehat{\boldsymbol{\xi}^o} = \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$, respectively.

We note that those changes are included in our default implementation of C&CG-DRO with *Oracle-2*.

(C2) Solving subproblems often leads to many scenarios with zero probability in $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ and $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$, which could be more prominent when *Oracle-2* is applied. To keep **MMP** more tractable, we only set $\widehat{\boldsymbol{\xi}^f}$ and $\widehat{\boldsymbol{\xi}^o}$ to include scenarios of non-zero probabilities from $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ and $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$ in our default implementation of C&CG-DRO. Yet, to keep the following theoretical analyses simple, C&CG-DRO does not screen out scenarios of zero probability.

(C3) If we are concerned with the feasibility of **PMP**(O) (i.e., the pricing master problem of *Oracle-2* for **WCEV**(O)), instead of using Farka's Lemma modified implementation, we can simply use $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ to initialize **PMP**(O). Since **WCEV**(O) is to be solved in **Step 4 – Case A**, employing $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ guarantees that **PMP**(O) is feasible. Yet, as noted earlier, set $\hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)$ might not be small and could slow down the computation of **PMP**(O).

A detailed flow chart describing C&CG-DRO with *Oracle-2* is presented in Figure 4 at the end of this paper. In the next subsection, we analyze C&CG-DRO's strength with respect to existing approaches, and its convergence and iteration complexity.

4.3 Strength, Convergence and Iteration Complexity

As previously mentioned, the classical C&CG implementation, referred to as basic C&CG, is one of the most popular methods for solving **2 – Stg DRO**. As a new and more intricate algorithm, it is significant to analytically demonstrate C&CG-DRO's strength compared to that of basic C&CG. Next, we show that this actually is the case under some mild conditions. Since the infeasibility issue of the recourse problem has not been studied in the literature, we assume that there is no feasibility issue associated with $\mathbf{x} \in \mathcal{X}$. We also note that *Oracle-2* is adopted in C&CG-DRO in the following analysis, because of its available infrastructure.

Proposition 14. (*i*) Assume that both basic C&CG and C&CG-DRO have the same scenario set $\widehat{\boldsymbol{\xi}^o}$ before the start of iteration t^0. We further assume that the solution algorithms for their master and subproblems are deterministic and **PMP**(O) has a unique shadow price with respect to $\widehat{\boldsymbol{\xi}^o}$. Then the lower bound derived from basic C&CG in iteration $t^0 + 1$ is dominated by that of C&CG-DRO.
(*ii*) Assume that the (main) master problems of basic C&CG and C&CG-DRO generate the same first-stage solution in iteration t^0. If the upper bound is updated in basic C&CG in iteration t^0, this upper bound is dominated by that of C&CG-DRO.

Proof. (*i*) As noted in Remark 8, (21), the duality based reformulation of $\mathbf{MMP}_2$, serves as the (main) master problem for both basic C&CG and C&CG-DRO. We let $(\mathbf{x}^*, \alpha^*, \boldsymbol{\beta}^*, \dots)$ denote the optimal solution obtained from computing the (main) master problem for both algorithms, noting that $\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^o$. With **(C1)** implemented and the stated assumptions, it can be seen that $(\alpha^*, \boldsymbol{\beta}^*)$ is the shadow price of $\mathbf{PMP}(O)$ for $\widehat{\boldsymbol{\xi}}^o$ and $\mathbf{x}^*$. Hence, the associated $\mathbf{PSP}(O)$ is

$$\max_{\xi \in \Xi} Q(\mathbf{x}^*, \xi) - \alpha^* - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^*.$$

Also, the subproblem of basic C&CG is

$$\max_{\xi \in \Xi} Q(\mathbf{x}^*, \xi) - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^*.$$

Clearly, they are essentially the same and have the identical optimal solution. Denoting it by ξ', we therefore have $\xi' \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*) \backslash \widehat{\boldsymbol{\xi}}^o$, the set of new scenarios identified after executing *Oracle-2*. As a result, after computing the updated (main) master problems for C&CG and C&CG-DRO, respectively, in iteration $(t^0 + 1)$, the lower bound derived by the former method is weaker than that generated by the latter one.

(ii) Again, in iteration t^0, let $\mathbf{x}^*$ be the optimal first-stage solution output from computing the (main) master problems, ξ' the optimal solution from solving the subproblem of basic C&CG, and its optimal value $\alpha' = Q(\mathbf{x}^*, \xi') - \sum_{i=1}^{m} \psi_i(\xi')\beta_i^*$. Hence, we have

$$\alpha' \geq Q(\mathbf{x}^*, \xi) - \sum_{i=1}^{m} \psi_i(\xi)\beta_i^* \quad \forall \xi \in \Xi.$$

Upon termination of *Oracle-2* with $\epsilon = 0$, the dual problem of $\mathbf{PMP}(O)$ is

$$\underline{\eta}^{o*}\big(\mathbf{x}^*, \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)\big) = \min \Big\{ \alpha + \sum_{i=1}^{m} \beta_i \gamma_i : \alpha + \sum_{i=1}^{m} \psi_i(\xi)\beta_i \geq Q(\mathbf{x}^*, \xi) \ \ \forall \xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*), \ \beta_i \geq 0 \ \ \forall i \in [m] \Big\}.$$

Given that $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*) \subseteq \Xi$, it is clear that $(\alpha', \boldsymbol{\beta}^*)$ is a feasible solution to this dual problem. So, we have $\underline{\eta}^{o*}\big(\mathbf{x}^*, \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)\big) \leq \alpha' + \sum_{i=1}^{m} \beta_i^* \gamma_i$, leading to

$$f_1(\mathbf{x}^*) + \underline{\eta}^{o*}\big(\mathbf{x}^*, \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)\big) \leq f_1(\mathbf{x}^*) + \alpha' + \sum_{i=1}^{m} \beta_i^* \gamma_i.$$

Note that the left-hand-side and right-hand-side of this inequality are benchmarked against the current upper bounds and employed to update them if applicable by C&CG-DRO and basic C&CG, respectively (e.g., Step 5 in C&CG-DRO). Hence, if an update occurs to the upper bound in basic C&CG, the new upper bound is dominated by the upper bound provided by C&CG-DRO. □

Next, we show that C&CG-DRO eventually solves $\mathbf{2-Stg\ DRO}$. We assume that *Oracle-1* is adopted as it does not incur any numerical issue for **WCEV** problem.

Theorem 15. For a fixed $\mathbf{x} \in \mathcal{X}$, let $\Xi^o(\mathbf{x}) = \{\xi \in \Xi : Q(\mathbf{x}, \xi) < +\infty\}$ denote the feasible sample space under $\mathbf{x}$, and the assumption made in Theorem 9 holds for $\mathbf{x} \in \mathcal{X}$, i.e.,

$$Q(\mathbf{x}, \xi) - \sum_i \psi_i(\xi)\beta_i$$

is continuous, uniformly with respect to ξ for $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\beta} \geq \mathbf{0}$.

(i) If there is no feasibility issue for the recourse problem, i.e., all $\mathbf{x} \in \mathcal{X}$ are feasible, C&CG-DRO returns an optimal (or $\varepsilon-$optimal) one in a finite number of iterations.

(ii) When the feasibility issue exists, we further assume Ξ is a polytope, $\tilde{Q}_f(\mathbf{x}, \xi)$ is convex over Ξ, $Q(\mathbf{x}, \xi)$ is uniformly continuous over $\Xi^o(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, and $\psi_i(\xi)$ are concave over Ξ for $i \in [m]$. C&CG-DRO either reports that $\mathbf{2-Stg\ DRO}$ is infeasible or returns an optimal (or $\varepsilon-$optimal) one in a finite number of iterations.

Proof. We first present the proof for the first statement when $\mathbf{2-Stg\ DRO}$ has no feasibility issues. Note that according to the definition of the uniform continuity, we have that for any given $\varepsilon > 0$ there exists a $\delta > 0$ satisfying

$$\left|\left(Q(\mathbf{x}, \xi_1) - \sum_i \psi_i(\xi_1)\beta_i\right) - \left(Q(\mathbf{x}, \xi_2) - \sum_i \psi_i(\xi_2)\beta_i\right)\right| \leq \varepsilon,$$

if $\|\xi_1 - \xi_2\| \leq \delta$ for $\xi_1, \xi_2 \in \Xi$, $\mathbf{x} \in \mathcal{X}$, and $\boldsymbol{\beta} \geq \mathbf{0}$.

Claim 1: Suppose that $\widehat{\boldsymbol{\xi}^o}$ is currently available scenarios, $\mathbf{x}^*$ is an optimal solution to $\mathbf{MMP}$ defined with $\widehat{\boldsymbol{\xi}^o}$, and $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$ has been obtained by solving $\mathbf{WCEV}(O)$ for $\mathbf{x}^*$ in iteration t^0. If the Hausdorff distance between sets $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)$ and $\widehat{\boldsymbol{\xi}^o}$, denoted by $h\big(\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*), \widehat{\boldsymbol{\xi}^o}\big)$, is less than or equal to δ, we have $UB \leq LB + \varepsilon$, i.e., $\mathbf{x}^*$ is an $\varepsilon-$optimal solution.

Proof of Claim 1: Note that, for $\mathbf{x}^*$, its optimal value of $\mathbf{WCEV}(O)$ reduces to

$$\begin{aligned}
\eta^o(\mathbf{x}^*) &= \max\Big\{\sum_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)} p_\xi Q(\mathbf{x}^*, \xi) : (p_\xi)_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)} \in \mathcal{P}\Big\} \\
&= \min\Big\{\alpha + \sum_{i=1}^m \gamma_i\beta_i : \ \alpha + \sum_i \psi_i(\xi)\beta_i \geq Q(\mathbf{x}^*, \xi) \ \ \forall \xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)\Big\}.
\end{aligned}$$

For the dual form (21) of $\mathbf{MMP}$, let $\big(\mathbf{x}^*, \alpha^{o*}, \boldsymbol{\beta}^{o*}, (\eta^{o*}_\xi)_{\xi \in \widehat{\boldsymbol{\xi}^o}}, \dots\big)$ be components associated with $\mathbf{x}^*$. Without loss of generality, we assume that $\eta^{o*}_\xi = Q(\mathbf{x}^*, \xi)$. We have

$$\begin{aligned}
LB = \underline{w}^{t^0} &= f_1(\mathbf{x}^*) + \eta^* \\
\eta^* =& \alpha^{o*} + \sum_{i=1}^m \gamma_i\beta_i^{o*} \\
\alpha^{o*} + \sum_i \psi_i(\xi)\beta_i^{o*} &\geq Q(\mathbf{x}^*, \xi) \ \ \forall \xi \in \widehat{\boldsymbol{\xi}^o}.
\end{aligned}$$

By the definition of Hausdorff distance, we have $\max_{\xi_1 \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)} \Big\{\min_{\xi_2 \in \widehat{\boldsymbol{\xi}^o}} \{\|\xi_1 - \xi_2\|\}\Big\} \leq \delta$.

Following the uniform continuity assumption on $Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i$, we have

$$
\begin{aligned}
&\max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)} \{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} - \max_{\xi \in \widehat{\boldsymbol{\xi}}^o}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} \\
\leq &\max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)} \Big\{\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} - \max_{\xi' \in \widehat{\boldsymbol{\xi}}^o : \|\xi'-\xi\| \leq \delta} \{Q(\mathbf{x},\xi') - \sum_i \psi_i(\xi')\beta_i^{o*}\}\Big\} \\
\leq &\max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)} \varepsilon = \varepsilon.
\end{aligned}
$$

Then, it can be inferred that

$$
\begin{aligned}
\alpha^{o*} &\geq \max_{\xi \in \widehat{\boldsymbol{\xi}}^o}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} \\
&= \max_{\xi \in \widehat{\boldsymbol{\xi}}^o}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} - \max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} + \\
&\max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} \\
&\geq \max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} - \Big|\max_{\xi \in \widehat{\boldsymbol{\xi}}^o}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} - \\
&\max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\}\Big| \\
&\geq \max_{\xi \in \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)}\{Q(\mathbf{x},\xi) - \sum_i \psi_i(\xi)\beta_i^{o*}\} - \varepsilon.
\end{aligned}
$$

Therefore, $(\alpha^{o*} + \varepsilon,\ \boldsymbol{\beta}^{o*})$ is feasible to the dual problem of **WCEV**(O), i.e., $UB \leq f_1(\mathbf{x}^*) + \eta^o(\mathbf{x}^*) \leq f_1(\mathbf{x}^*) + \alpha^{o*} + \varepsilon + \sum_{i=1}^m \gamma^i \beta_i^{o*} \leq \underline{w}^{t_0} + \varepsilon = LB + \varepsilon$. □

Given the definition of $\eta^o(\mathbf{x}^*)$, we can infer that $\mathbf{x}^*$ is an $\varepsilon-$optimal solution to $\mathbf{2-Stg\ DRO}$. Moreover, when $\hat{\boldsymbol{\xi}}^o(\mathbf{x}^*) \subseteq \widehat{\boldsymbol{\xi}}^o$, it is clear that we have $UB \leq LB$, rendering $\mathbf{x}^*$ an optimal solution to $\mathbf{2-Stg\ DRO}$.

Similar to the proof of Theorem 9 on compact sample space Ξ, we introduce balls with volume equal to $V(\frac{\delta}{2})$. According to Claim 1, we can conclude that after a finite number of iterations, C&CG-DRO either returns an optimal solution or an $\varepsilon-$optimal solution when no feasibility issue occurs.

Next, we consider the second statement and assume that *Oracle-1* returns an extreme point solution of Ξ if such type of optimal solutions is available.

Claim 2: If $\mathbf{2-Stg\ DRO}$ is infeasible, C&CG-DRO reports its infeasibility within a finite number of iterations.

Proof of Claim 2: Since Ξ is a polytope, $\tilde{Q}_f(\mathbf{x},\xi)$ is convex and $\psi_i(\xi)$ are concave over Ξ for $i \in [m]$, it follows from Corollary 10 that, if **MMP** is feasible, at least one new extreme point from $\mathsf{XV}(\Xi)$ (which does not belong to existing $\widehat{\boldsymbol{\xi}}^f$) will be derived after solving **WCEV**(F). Given that $\mathsf{XV}(\Xi)$ is a finite set, it follows that **MMP** will become infeasible within a finite number of iterations. Hence, it certifies that $\mathbf{2-Stg\ DRO}$ is infeasible. □

Note that the feasibility check presented in **WCEV**(F) will only be performed a finite number of times. By combining Claims 1 and 2, C&CG-DRO either reports that

$\mathbf{2-Stg\ DRO}$ is infeasible or returns an optimal (or $\varepsilon-$optimal) one in a finite number of iterations. □

Next, we present a few results regarding the number of iterations of the algorithm. The following one can be easily obtained by Theorem 9.

Corollary 16. For a given ε, assume δ in the proof Theorem 9 is available. Then, the number of iterations of C&CG-DRO before termination is bounded by $\left\lceil V(\hat{d}+\frac{\delta}{2})/V(\frac{\delta}{2})+|\mathsf{XV}(\Xi)|\right\rceil$.

When $\mathcal{X}$ or Ξ is a finite set, a natural upper bound on the iteration complexity can be derived. Note that it is easy to show that repeated $\mathbf{x}$ or ξ leads to $LB = UB$.

Proposition 17. (i) If $\mathcal{X}$ is a finite set, the number of iterations is bounded by the cardinality of $\mathcal{X}$, i.e., $|\mathcal{X}|$; (ii) If sample space Ξ is finite, the number of iterations is bounded by the cardinality of Ξ.

Next, we discuss the case where both the recourse problem and the ambiguity set are linear, noting that it can be exactly solved within a finite number of iterations.

Proposition 18. Assume that the recourse problem is an LP, Ξ is a polytope, and $\psi_i(\xi)$ are linear over Ξ for $i \in [m]$. C&CG-DRO either reports that $\mathbf{2-Stg\ DRO}$ is infeasible, or converges to its exact solution within $|\mathsf{XV}(\Xi)|$ iterations.

Proof. According to the proof of Corollary 10 and Theorem 15, since the set of extreme points of Ξ, i.e., $|\mathsf{XV}(\Xi)|$, is finite, C&CG-DRO either reports that $\mathbf{2-Stg\ DRO}$ is infeasible, or converges to an exact solution that belongs to $\mathsf{XV}(\Xi)$. Hence, the whole procedure completes in $|\mathsf{XV}(\Xi)|$ iterations. □

5 Further Investigations on Solving $\mathbf{2-Stg\ DRO}$

In this section, a couple of new variants of the primary C&CG-DRO algorithm are presented to support us with a stronger solution capacity. We first develop a variant to compute $\mathbf{2-Stg\ DRO}$ with Wasserstein metric-based ambiguity set. Then, we describe another variant to handle MIP ambiguity set, which has not been investigated in the literature yet and can be used to capture more sophisticated uncertainties.

On one hand, the basic intuition behind the variant for $\mathbf{2-Stg\ DRO}$ with Wasserstein metric-based ambiguity set is: given that we are considering a distribution close to an empirical one, the critical scenarios underlying that distribution should also be close to those of the empirical one. Hence, we should identify some non-trivial scenarios around each empirical sample (in an independent fashion) and then aggregate them to provide a basis for selection. We repeat this operation in the framework of C&CG-DRO. On the other hand, the basic intuition to handle the MIP ambiguity set is: we generate scenarios and build a bilevel MIP (instead of a bilevel LP for convex ambiguity set) to obtain a lower bound approximation of the original $\mathbf{2-Stg\ DRO}$. As bilevel MIP is very hard to solve, we fix discrete variables in the lower-level problem to convert it into an LP and thus to derive a weaker lower bound. Such a lower bound can be improved through C&CG-DRO iteratively.

5.1 A Variant to Handle Wasserstein Metric-Based Ambiguity Set

Consider two distributions P and P^e. The Wasserstein metric between them is defined as

$$\mathcal{W}(P,P^e) = \Big\{ \inf_{K\in S(P,P^e)} \int_{\Xi}\int_{\Xi^e} \|\xi-\zeta\|_{\mathbf{p}} K(d\zeta,d\xi) \Big\}, \tag{25}$$

where $S(P,P^e)$ denotes the collection of joint probability distributions of ξ and ζ with marginal distributions being P and P^e, respectively, and $\|\cdot\|_{\mathbf{p}}$ denotes $L_{\mathbf{p}}$ norm. Accordingly, the Wasserstein metric-based ambiguity set, which has received a lot of attention in the literature of data-driven decision making, is

$$\mathcal{P}^W = \Big\{ P\in\mathcal{M}(\Xi,\mathcal{F}) : \mathcal{W}(P,P^e)\le \mathrm{r} \Big\},$$

where r is the radius bound imposed on the Wasserstein metric. When P^e is defined on a set of empirical samples, i.e., $\Xi^e \equiv \{\xi_1^e,\dots,\xi_N^e\}$ with $\{p_i^e\}_{i=1}^N$ being its probability mass function, (25) can be simplified to the following one by taking advantage of conditional probabilities [8, 9],

$$\mathcal{W}(P,P^e) = \Big\{ \min_{P_i\in\mathcal{M}(\Xi,\mathcal{F}), i\in[N]} \sum_{i=1}^N p_i^e \int_{\Xi} \|\xi-\xi_i^e\| P_i(d\xi) : \int_{\Xi} P_i(d\xi) = 1 \quad \forall i\in[N] \Big\}, \tag{26}$$

noting that P_i denotes the conditional probability distribution under the condition of ξ_i^e. Clearly, we have $P=\sum_{i=1}^N p_i^e P_i$.

Remark 10. We mention that, with the help of the indicator function, Wasserstein metric-based ambiguity set can be described in the form of (3) as in the following.

$$\begin{aligned}\mathcal{P}^W = \Big\{ P\in\mathcal{M}(\Xi,\mathcal{F}) : &\int_{\Xi} P(d\xi) = 1, \int_{\Xi} \mathbb{1}_{\{i=k\}} P(d\xi) = p_k^e \quad \forall k\in[N], \\ &\int_{\Xi} \sum_{i=1}^N\sum_{k=1}^N p_i^e \mathbb{1}_{\{i=k\}} \|\xi-\xi_k^e\|_{\mathbf{p}} P_i(d\xi) \le r \Big\},\end{aligned}$$

where the indicator function $\mathbb{1}_{\{i=k\}}$ equals 1 if $i=k$ and 0 otherwise. □

With $\mathcal{P}^W$ defined on Ξ^e, it has been noted that the true data-generating probability distribution contained in $\mathcal{P}^W$ is of a high confidence, and, under some mild assumption, a finite-sample guarantee can be established for a solution to $\mathcal{P}^W$-based DRO [8]. We next present a new variant of $\mathbf{C\&CG-DRO}$ to compute such $\mathcal{P}^W$-based $\mathbf{2-Stg\ DRO}$, which is referred to as $\mathbf{C\&CG-DRO}(\mathcal{P}^W)$ for simplicity. Also, in the context of Wasserstein metric-based ambiguity set, unless explicitly stated otherwise, we omit the subscript $\mathbf{p}$ for clarity in our exposition.

5.1.1 Computing WCEV Problems for Given $\mathbf{x}^*$

To describe new variants of solution oracles for $\mathcal{P}^W$-based **WCEV**'s, we use **WCEV**(F) for illustration. Note that all results developed for **WCEV**(F) are applicable to its counterpart **WCEV**(O) by simply replacing $\tilde{Q}_f(\xi_j, \mathbf{x}^*)$ with $Q(\xi_j, \mathbf{x}^*)$. Specifically, consider the following **WCEV**(F) formulated with respect to $\mathcal{P}^W$.

$$\begin{aligned} \mathbf{WCEV}(F) : \eta^f(\mathbf{x}^*) = \sup_{P_i \in \mathcal{M}(\Xi,\mathcal{F}), i\in[N]} & \sum_{i=1}^N p_i^e \int_\Xi \tilde{Q}_f(\mathbf{x}^*, \xi) P_i(d\xi) \\ \text{s.t. } & \int_\Xi P_i(d\xi) = 1 \quad \forall i \in [N] \\ & \sum_{i=1}^N p_i^e \int_\Xi \|\xi - \xi_i^e\| P_i(d\xi) \le r \end{aligned} \tag{27}$$

In the following, we leverage a structural insight and the Pigeonhole principle, a classical result in combinatorics, to develop a novel and compact finite mathematical program that solves **WCEV**(F) problem.

Proposition 19. An optimal solution to the following finite mathematical program solves **WCEV**(F) problem in (27) exactly.

$$\begin{aligned} \mathbf{WCEV}(F)\text{–}\mathbf{FMP} : \max \Big\{ & \sum_{i=1}^N p_i^e [\tilde{Q}_f(\mathbf{x}^*, \xi_{i1}) p_{i1} + \tilde{Q}_f(\mathbf{x}^*, \xi_{i2}) p_{i2}] : p_{i1} + p_{i2} = 1 \quad \forall i \in [N], \\ & \sum_{i=1}^N p_i^e (\|\xi_{i1} - \xi_i^e\| p_{i1} + \|\xi_{i2} - \xi_i^e\| p_{i2}) \le r, \ \xi_{i1}, \xi_{i2} \in \Xi \quad \forall i \in [N] \Big\} \end{aligned}$$

Proof. According to Theorm 3, the optimal value of (27) can be obtained. Following similar steps to the proof of Proposition 5 and applying the Richter-Rogosinski Theorem [20], (27) has an optimal distribution with at most $(N+1)$ scenarios of non-zero probabilities. Hence, given the definition of P_i, (27) is equivalent to

$$\begin{aligned} \mathbf{WCEV}(F)\text{–}\mathbf{FMP}^0 : \max \Big\{ & \sum_{i=1}^N \sum_{j=1}^{N+1} p_i^e \tilde{Q}_f(\mathbf{x}^*, \xi_{ij}) p_{ij} : \sum_{j=1}^{N+1} p_{ij} = 1 \quad \forall i \in [N], \\ & \sum_{i=1}^N \sum_{j=1}^{N+1} p_i^e \|\xi_{ij} - \xi_i^e\| p_{ij} \le r, \xi_{ij} \in \Xi \quad \forall i \in [N], \ \forall j \in [N+1] \Big\}. \end{aligned}$$

Assume an optimal solution to **WCEV**(F)–**FMP**^{0}, denoted by $(\xi_{ij}^*, p_{ij}^*)_{i\in[N], j\in[N+1]}$, is available. By fixing $\xi_{ij} = \xi_{ij}^*$, **WCEV**(F)–**FMP**^{0} reduces to an LP with $(N+1)$ constraints. Hence, according to the theory of LP, there exists an optimal solution with at most $(N+1)$ probability variables p_{ij} being non-zeros.

Given that $\sum_{j=1}^{N+1} p_{ij} = 1$ must be satisfied, it follows that we have at least one $p_{ij} > 0$ for every i. Then, by the Pigeonhole principle, we further have that no more than two $p_{ij} > 0$

for every i. As a result, for i, we can reduce $(N+1)$ variables to just two variables, i.e., p_{i1} and p_{i2}, through which **WCEV**(F)–**FMP**^{0} becomes **WCEV**(F)–**FMP**. □

It is worth pointing out that (27) (as well as its counterpart for **WCEV**(O)) actually has a decomposable structure. Note in the objective function of (27) that

$$\begin{aligned}\sum_{i=1}^{N} p_i^e \int_{\Xi} \tilde{Q}_f(\mathbf{x}^*, \xi) P_i(d\xi) &= p_1^e \int_{\Xi} \tilde{Q}_f(\mathbf{x}^*, \xi) P_1(d\xi) + \cdots + p_N^e \int_{\Xi} \tilde{Q}_f(\mathbf{x}^*, \xi) P_N(d\xi) \\ &= p_1^e \int_{\Xi} \tilde{Q}_f(\mathbf{x}^*, \xi^1) P_1(d\xi^1) + \cdots + p_N^e \int_{\Xi} \tilde{Q}_f(\mathbf{x}^*, \xi^N) P_N(d\xi^N),\end{aligned}$$

which allows us to modify *Oracle-2* to tackle P_i parallelly. Let $\hat{\boldsymbol{\xi}}_i^f = \{\xi_{i,1}^f, \ldots, \xi_{i,n_i}^f\}$ denote a set of scenarios introduced for P_i, $i = 1, \ldots, N$, and $\widehat{\boldsymbol{\xi}}^f = (\hat{\boldsymbol{\xi}}_1^f, \ldots, \hat{\boldsymbol{\xi}}_N^f)$. We construct the following **PMP**(F) defined on top of $\widehat{\boldsymbol{\xi}}^f$. Note that dual variables are indicated after the colon for each constraint.

$$\mathbf{PMP}(F) : \underline{\eta}^{f*}(\mathbf{x}^*, \widehat{\boldsymbol{\xi}}^f) = \max \sum_{i=1}^{N} p_i^e \sum_{j=1}^{n_i} \tilde{Q}_f(\mathbf{x}^*, \xi_{i,j}^f) p_i(\xi_{i,j}^f) \tag{28a}$$

$$\text{s.t.} \sum_{j=1}^{n_i} p_i(\xi_{i,j}^f) = 1 \quad \forall i \in [N] \qquad : \alpha_i^f \tag{28b}$$

$$\sum_{i=1}^{N} p_i^e \sum_{j=1}^{n_i} \|\xi_{ij}^f - \xi_i^e\| p_i(\xi_{i,j}^f) \le r \qquad : \beta^f \tag{28c}$$

$$p_i(\xi_{i,j}^f) \ge 0 \quad \forall j \in [n_i],\ \forall i \in [N] \tag{28d}$$

Then, with **PMP**(F)'s shadow price $(\alpha_1^{f*}, \cdots, \alpha_N^{f*}, \beta^{f*})$ and the stricture of **PMP**(F), we define **PSP**$^i(F)$ for $i \in [N]$ as follows.

$$\mathbf{PSP}^i(F) : v^{i*}(\mathbf{x}^*, \widehat{\boldsymbol{\xi}}^f) = \max_{\xi \in \Xi}\ p_i^e \big(\tilde{Q}_f(\mathbf{x}^*, \xi) - \beta^{f*} \|\xi_i^e - \xi\|\big) - \alpha_i^{f*}. \tag{29}$$

Remark 11. (i) Note that, except $\tilde{Q}_f(\mathbf{x}^*, \xi)$, the objective function of **PSP**$^i(F)$ is concave for $L_{\mathbf{p}}$-norm if $\mathbf{p} \ge 1$ or linearizable (by introducing binary variables) if $\mathbf{p} = 0$. So, it does not impose additional computational challenge, compared to standard **PSP**(F) constructed in Section 3.2.

(ii) Regarding customization of *Oracle-2* to handle $\mathcal{P}^w$, for shadow price $(\alpha_1^{f*}, \cdots, \alpha_N^{f*}, \beta^{f*})$, all **PMP**$^i(F)$'s will be solved in an independent or parallel fashion for $i \in [N]$. Once a scenario with positive reduced cost, denoted by ξ_i^{f*}, is identified by **PSP**$^i(F)$, $\hat{\boldsymbol{\xi}}_i$ will be augmented as $\hat{\boldsymbol{\xi}}_i^f = \hat{\boldsymbol{\xi}}_i^f \cup \{\xi_i^{f*}\}$. Then, **PMP**$(F)$, with the updated $\hat{\boldsymbol{\xi}}_i^f$ for $i \in [N]$, will be used to generate a new shadow price. Hence, the CG procedure terminates if none of **PMP**$^i(F)$ produces a scenario of positive reduced cost, which is the primary difference to *Oracle-2* presented in Section 3.2. Upon termination, we note that $\hat{\boldsymbol{\xi}}_i^f \cap \hat{\boldsymbol{\xi}}_j^f$ might not be empty for $i \neq j$. It suggests that scenarios in the overlap are critical to both ξ_i^e and ξ_j^e. □

5.1.2 Customizing C&CG-DRO to Handle $\mathcal{P}^W$

Similar to $\widehat{\boldsymbol{\xi}}^f$, we let $\widehat{\boldsymbol{\xi}}^o = (\hat{\boldsymbol{\xi}}^o_1, \dots, \hat{\boldsymbol{\xi}}^o_N)$, i.e., optimality sets for $i \in [N]$, and $\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^f \cup \widehat{\boldsymbol{\xi}}^o$. Note that they have been updated up to now after solving **WCEV**(F) and **WCEV**(O) in previous iterations, respectively. Then, the main master problem of $\mathbf{2-Stg\ DRO}$ defined on $\mathcal{P}^W$ can be reformulated as:

$$\mathbf{MMP}^W : \underline{w} = \min_{\mathbf{x} \in \mathcal{X}} f_1(\mathbf{x}) + \eta \tag{30a}$$

$$\eta \geq \max \left\{ \sum_{i=1}^N p_i^e \sum_{\xi \in \hat{\boldsymbol{\xi}}_i} \eta^o_\xi p^o_\xi : (p^o_\xi)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P}^W \right\} \tag{30b}$$

$$\eta^o_\xi = f_2(\mathbf{x}, \xi, \mathbf{y}_\xi) \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{30c}$$

$$0 \geq \max \left\{ \sum_{i=1}^N p_i^e \sum_{\xi \in \hat{\boldsymbol{\xi}}_i} \eta^f_\xi p^f_\xi : (p^f_\xi)_{\xi \in \widehat{\boldsymbol{\xi}}} \in \mathcal{P}^W \right\} \tag{30d}$$

$$\eta^f_\xi = \|\tilde{\mathbf{y}}_\xi\|_1 \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{30e}$$

$$(\mathbf{y}_\xi, \tilde{\mathbf{y}}_\xi) \in \tilde{\mathcal{Y}}(\mathbf{x}, \xi) \quad \forall \xi \in \widehat{\boldsymbol{\xi}} \tag{30f}$$

Similar to $\mathbf{MMP}_2$ in (18), $\mathbf{MMP}^W$ is a bilevel optimization formulation and can be easily converted into a single-level. We next present the single-level one derived by using the strong duality-based reformulation technique.

$$\begin{aligned} \mathbf{MMP}^W : \underline{w} = \min \Big\{ & f_1(\mathbf{x}) + \eta : \mathbf{x} \in \mathcal{X},\ \eta \geq \sum_{i=1}^N \alpha_i^o + r\beta^o, \\ & \left\{ \alpha_i^o + p_i^e \|\xi - \xi_i^e\| \beta^o \geq p_i^e \eta^o_\xi \quad \forall \xi \in \hat{\boldsymbol{\xi}}_i \right\} \quad \forall i \in [N] \\ & 0 \geq \sum_{i=1}^N \alpha_i^f + r\beta^f, \\ & \left\{ \alpha_i^f + p_i^e \|\xi - \xi_i^e\| \beta^f \geq p_i^e \eta^f_\xi \quad \forall \xi \in \hat{\boldsymbol{\xi}}_i \right\} \quad \forall i \in [N] \\ & (30c),\ (30e),\ (30f),\ \beta^o \geq 0,\ \beta^f \geq 0 \Big\}. \end{aligned} \tag{31}$$

With the aforementioned $\mathbf{MMP}^W$ and those defined to solve **WCEV** problems, the customization of C&CG-DRO to solve $\mathbf{2-Stg\ DRO}$ with $\mathcal{P}^W$ can be easily obtained. As the necessary changes are rather straightforward, we do not provide detailed descriptions.

Remark 12. (i) Actually, as Wasserstein metric-based ambiguity set is defined on top of Ξ^e, it is rather straightforward to initialize $\hat{\boldsymbol{\xi}}^o_i$ by ξ^e_i. Moreover, given that $p^e_i > 0$, the associated recourse problem for ξ^e_i must be feasible for any choice of $\mathbf{x}$. In our numerical study, when the radius r is small, this initialization strategy is computationally very effective, while it is less effective when r is large. Also, when N is large, using the whole Ξ^e for initialization might not be computationally effective. If this is the case, we can consider employing a more adversarial subset of Ξ^e for initialization.

(ii) We can also modify the procedure of generating Benders cutting planes according to the one described in Remark 9 to handle $\mathcal{P}^W$. We consider the recourse problem in (24) and Benders cuts for optimality for demonstration. Those cuts, generated after solving the current $\mathbf{WCEV}(O)$ with $\mathbf{x} = \mathbf{x}^*$, are in the following forms.

$$\Big\{\alpha_i^o + p_i^e \|\xi - \xi_i^e\|\beta^o \geq p_e^i \eta_\xi^o \;\; \forall \xi \in \hat{\boldsymbol{\xi}}_i^o(\mathbf{x}^*)\Big\} \;\; \forall i \in [N]$$
$$\Big\{\eta_\xi^o \geq (\mathbf{b}_2 - \mathbf{A}_2\mathbf{x} - \mathbf{H}\xi)^\intercal \pi_\xi^* \;\; \forall \xi \in \hat{\boldsymbol{\xi}}_i^o(\mathbf{x}^*)\Big\} \;\; \forall i \in [N]$$

□

5.2 An Extension to Handle Mixed Integer Ambiguity Set

As mentioned, so far all ambiguity sets in the DRO literature are assumed to be convex in P, although the underlying sample spaces can be either continuous or discrete. Such an assumption is crucial to duality based reformulations, and actually is the enabling structure for all known solution methods. Nevertheless, it seriously restricts our modeling capacity to describe and analyze real world problems. Even some simple situations, as shown next, cannot be represented by any convex set, while they can be captured by mixed integer sets.

Example 1. In addition to constraints in (3), we consider a situation where the first moment only belongs to one of two intervals, i.e., $[l_1, u_1]$ and $[l_2, u_2]$. It can be seen that it is not convex in P, as a convex combination of two legitimate distributions may not yield a legitimate one. Yet, it can be represented by introducing a new binary variable, z, and incorporating the following constraints

$$l_1 z + l_2(1 - z) \leq \int_\Xi \xi P(d\xi) \leq u_1 z + u_2(1 - z), \tag{32}$$

resulting in a mixed integer representation for the updated ambiguity set $\mathcal{P}$. □

It is easy to see that for both $z = 0$ and 1, the corresponding **WCEV** problems have optimal discrete probability distributions. Hence the original one does, too. Nevertheless, such mixed integer ambiguity sets are infeasible to the popular duality based approaches. We highlight that they actually can be addressed by solution methods developed from the primal perspective, through which our modeling and solution capacity on DRO can be greatly improved. Specifically, we mainly consider the following mixed 0-1 ambiguity set to present our algorithm development for the associated $\mathbf{2-Stg\ DRO}$, where $\mathcal{Z} \subseteq \{0,1\}^{n_{\xi^I}}$ denotes the feasible set for binary vector $\mathbf{z}$. Note that it can be further extended to handle mixed integer Wasserstein metric-based ambiguity sets.

$$\mathcal{P}^I = \Big\{(P, \mathbf{z}) \in \mathcal{M}(\Xi, \mathcal{F}) \times \mathcal{Z} : \int_\Xi P(d\xi) = 1,\, E_P[\psi_i(\xi, \mathbf{z})] \leq \gamma_i(\mathbf{z}) \;\; \forall i \in [m]\Big\}. \tag{33}$$

5.2.1 Computing WCEV Problems for A Given $\mathbf{x}^*$

It can be seen that for any $\mathbf{z} \in \mathcal{Z}$, there are still $m+1$ constraints defining P. Hence, the continuous finite mathematical program **WCEV − FMP** presented in Section 3.1 can be easily extended to the following mixed integer one to compute **WCEV**(F) problem. We note again that all results developed for **WCEV**(F) problem are applicable to its counterpart **WCEV**(O).

Corollary 20. Assume that the sufficient conditions presented in Proposition 5 are satisfied for every $\mathbf{z} \in \mathcal{Z}$. Then, **WCEV** problem is equivalent to

$$\mathbf{WCEV}^{\boldsymbol{I}}(F) - \mathbf{FMP} : \max \Big\{ \sum_{j=1}^{m+1} p_j \tilde{Q}_f(\xi_j, \mathbf{x}^*) : \sum_{j=1}^{m+1} p_j = 1, \ \sum_{j=1}^{m+1} p_j \psi_i(\xi_j, \mathbf{z}) \le \gamma_i(\mathbf{z}) \ \ \forall i \in [m],$$
$$\xi_j \in \Xi \ \ \forall j \in [m+1], \ p_j \ge 0 \ \ \forall j \in [m+1], \ \mathbf{z} \in \mathcal{Z} \Big\}.$$

Example 1 (Continued)**.** As constraints in (32) are appended to those in (3), the number of constraints is $m+3$. Then, the corresponding **WCEV**$^{\boldsymbol{I}}$ **− FMP** is obtained by augmenting the original **WCEV − FMP** with binary variable z, continuous variables p_{m+2} and p_{m+3}, and ξ_{m+2} and ξ_{m+3}, and with constraint

$$l_1 z + l_2(1-z) \le \sum_{j=1}^{m+3} p_j \xi_j \le u_1 z + u_2(1-z). \qquad \square$$

Compared to original **WCEV**(F) **− FMP**, this mixed integer **WCEV**$^{\boldsymbol{I}}$(F) **− FMP** is even more computationally challenging. With the strong performance of *Oracle-2*, it would be desired to extend it to compute **WCEV** problems with respect to $\mathcal{P}^{\boldsymbol{I}}$. As noted earlier, the challenge of a mixed integer master problem (e.g., **PMP** with a mixed integer ambiguity set) can be addressed by customizing the classical B&P procedures on top of CG [35]. Rather than designing and implementing traditional branch-and-bound subroutines, we extend *Oracle-2* in a novel fashion that leverages strong features of professional solvers to minimize the extra development burden. In this subsection, we assume that $\mathcal{P}^{\boldsymbol{I}}$ is not empty for any $\mathbf{z} \in \mathcal{Z}$.

First, we present the new **PMP** defined for $\mathcal{P}^{\boldsymbol{I}}$, denoted by **PMP**$^{\boldsymbol{I}}$.

$$\mathbf{PMP}^{\boldsymbol{I}}(F) : \underline{\eta}^*(\boldsymbol{\xi}_n^0) = \max \Big\{ \sum_{j=1}^{n} \tilde{Q}_f(\xi_j^0, \mathbf{x}) p(\xi_j^0), \ \sum_{j=1}^{n} p(\xi_j^0) = 1, \ p(\xi_j^0) \ge 0 \ \ \forall j \in [n],$$
$$\sum_{j=1}^{n} \psi_i(\xi_j^0, \mathbf{z}) p(\xi_j^0) \le \gamma_i(\mathbf{z}) \ \ \forall i \in [m], \ \mathbf{z} \in \mathcal{Z} \Big\}. \tag{34}$$

Note that **PMP**$^{\boldsymbol{I}}$ is a mixed integer program, not a linear program. In the following, we define the pricing subproblem **PSP**$^{\boldsymbol{I}}$. It has a bilevel optimization structure, which is substantially different from classical pricing subproblems.

$$\mathbf{PSP}^{\boldsymbol{I}}(F): v^*(\boldsymbol{\xi}_n^0) = \max \Bigg\{ \tilde{Q}_f(\xi,\mathbf{x}) - \hat{\alpha} - \sum_{i=1}^{m} \psi_i(\xi,\mathbf{z})\hat{\beta}_i : \xi \in \Xi,\ \mathbf{z} \in \mathcal{Z} \tag{35}$$

$$\begin{aligned} \big(\hat{P},(\hat{\alpha},\hat{\boldsymbol{\beta}})\big) \in \arg\max \Big\{ \sum_{j\in 1}^{n} \tilde{Q}_f(\xi_i^0,\mathbf{x})p(\xi_i^0) : \sum_{j=1}^{n} p(\xi_j^0) = 1 \\ \sum_{j=1}^{n} \psi_i(\xi_j^0,\mathbf{z})p(\xi_j^0) \le \gamma_i(\mathbf{z}) \ \ \forall i \in [m],\ p(\xi_j^0) \ge 0 \ \ \forall j \in [n] \Big\} \Bigg\}. \end{aligned} \tag{36}$$

Note that the lower-level problem in (36) is actually the continuous portion of $\mathbf{PMP}^{\boldsymbol{I}}(F)$. With a slight abuse of notation, we let $\big(\hat{P},(\hat{\alpha},\hat{\boldsymbol{\beta}})\big)$ denote a pair of optimal primal and dual solutions to (36) for $\mathbf{z}$ selected by the upper-level problem. So, the overall bilevel optimization problem seeks to choose $(\xi,\mathbf{z})$ that maximizes the reduced cost based on $(\hat{\alpha},\hat{\boldsymbol{\beta}})$ feedback from the lower-level problem. We can further furnish this bilevel optimization problem with the following inequality. It stipulates that, with new ξ, the upper bound on the WCEV should be larger than or equal to the optimal value of $\mathbf{PMP}^{\boldsymbol{I}}(F)$, which helps us to reduce the generation of unnecessary columns in the execution of *Oracle-2*.

$$\sum_{j\in 1}^{n} \tilde{Q}_f(\xi_j^0,\mathbf{x})\hat{p}(\xi_j^0) + \tilde{Q}_f(\xi,\mathbf{x}) - \hat{\alpha} - \sum_{i=1}^{m} \psi_i(\xi,\mathbf{z})\hat{\beta}_i \ge \underline{\eta}^*(\boldsymbol{\xi}_n^0).$$

Indeed, bilevel optimization formulation $\mathbf{PSP}^{\boldsymbol{I}}$ can be treated as an integration of the master and subproblems of the conventional CG algorithm, where all $\mathbf{z}$'s in $\mathcal{Z}$ are evaluated in (35) when maximizing the reduced cost. Hence, the next result follows directly, which extends that presented in Proposition 7.

Proposition 21. Suppose that $\mathbf{PMP}^{\boldsymbol{I}}(F)$ is feasible and both $\mathbf{PMP}^{\boldsymbol{I}}(F)$ and $\mathbf{PSP}^{\boldsymbol{I}}(F)$ are solved to optimality. We have

$$\underline{\eta}^*(\boldsymbol{\xi}_n^0) \le \max_{P\in\mathcal{P}^{\boldsymbol{I}}} E_P[\tilde{Q}_f(\xi)] \le \underline{\eta}^*(\boldsymbol{\xi}_n^0) + v^*(\boldsymbol{\xi}_n^0). \qquad \square$$

Regarding the modification of *Oracle-2* to handle $\mathcal{P}^{\boldsymbol{I}}$, it can be done by simply using $\mathbf{PMP}^{\boldsymbol{I}}(F)$ and $\mathbf{PSP}^{\boldsymbol{I}}(F)$ to replace their counterparts, and removing the shadow price output in **Step 2**. The resulting variant is referred to as *Oracle-2*$^{\boldsymbol{I}}$.

Remark 13. (*i*) When solving bilevel optimization problem $\mathbf{PSP}^{\boldsymbol{I}}$ by a contemporary MIP solver through its single-level reformulation (see Appendix A.2), it is common to have both optimal primal and dual solutions, i.e., $\big(\hat{P},(\hat{\alpha},\hat{\boldsymbol{\beta}})\big)$ are available. Hence, *Oracle-2*$^{\boldsymbol{I}}$ is rather straightforward to implement, compared to the traditional B&P method.

(*ii*) In our computational study, a hybrid implementation including both *Oracle-2* and variant *Oracle-2*$^{\boldsymbol{I}}$ is often computationally more efficient. Specifically, let $\mathcal{P}(\mathbf{z})$ denote the ambiguity set for given $\mathbf{z}$. For a fixed $\mathbf{z}$ and the associated $\mathcal{P}(\mathbf{z})$, we first run *Oracle-2* to generate a set of ξ's until $v^*(\boldsymbol{\xi}_n^0)$ becomes 0. Then, *Oracle-2*$^{\boldsymbol{I}}$ is called to derive a new $\mathbf{z}$ with positive

$v^*(\xi_n^0)$, which allows us to switch back to run *Oracle-2*. We repeat those steps until no $\mathbf{z}$ with $v^*(\xi_n^0) > 0$, which terminates the whole procedure. Note that this fashion of implementation reduces the number of calls to solve bilevel optimization problems. □

5.2.2 Customizing C&CG-DRO to Handle $\mathcal{P}^I$

With an augmented $\widehat{\boldsymbol{\xi}}$, we can build a main master problem with respect to $\mathcal{P}^I$, which is a bilevel optimization formulation with MIP lower-level problem(s). Its optimal solution can be obtained by employing a C&CG variant specialized for such type of bilevel MIP optimization [36]. Nevertheless, rather than searching deeply for strongest $\mathbf{x}$ by solving a bilevel MIP program, we can build and solve the following main master problem that is simpler and can be solved more efficiently.

Note that, for a given $\mathbf{x}^*$, computing **WCEV**(F) (and **WCEV**(O), respectively) problem actually returns $\big(\mathbf{z}^{f*}(\mathbf{x}^*), \hat{\boldsymbol{\xi}}^f(\mathbf{x}^*)\big)$ (and $\big(\mathbf{z}^{o*}(\mathbf{x}^*), \hat{\boldsymbol{\xi}}^o(\mathbf{x}^*)\big)$, respectively). Hence, similar to $\widehat{\boldsymbol{\xi}}^f$ and $\widehat{\boldsymbol{\xi}}^o$, we assume that, before a new C&CG iteration, two sets of $\mathbf{z}$'s have been obtained accumulatively from computing **WCEV**(F) and **WCEV**(O) problems in previous iterations, denoted by $\widehat{\mathcal{Z}}^f$ and $\widehat{\mathcal{Z}}^o$, respectively. Also, let $\widehat{\mathcal{Z}} = \widehat{\mathcal{Z}}^f \cup \widehat{\mathcal{Z}}^o$, and $\mathcal{P}^I(\mathbf{z}')$ denote $\mathcal{P}^I$ with $\mathbf{z}$ fixed to $\mathbf{z}'$. Then, we construct and consider the following main master problem. As it is an extension of **MMP** in (18), we refer to it as **MMP**^{I}.

$$\mathbf{MMP}_2^I : \underline{w}' = \min_{\mathbf{x}\in\mathcal{X}} f_1(\mathbf{x}) + \eta \tag{37a}$$

$$\eta \geq \max\Big\{\sum_{\xi\in\widehat{\boldsymbol{\xi}}} \eta_\xi^o p_{\xi,\mathbf{z}}^o : (p_{\xi_1,\mathbf{z}}^o, \ldots, p_{\xi_{|\widehat{\xi}|},\mathbf{z}}^o) \in \mathcal{P}(\mathbf{z})\Big\} \quad \forall \mathbf{z} \in \widehat{\mathcal{Z}} \tag{37b}$$

$$0 \geq \max\Big\{\sum_{\xi\in\widehat{\boldsymbol{\xi}}} \eta_\xi^f p_{\xi,\mathbf{z}}^f : (p_{\xi_1,\mathbf{z}}^f, \ldots, p_{\xi_{|\widehat{\xi}|},\mathbf{z}}^f) \in \mathcal{P}(\mathbf{z})\Big\} \quad \forall \mathbf{z} \in \widehat{\mathcal{Z}} \tag{37c}$$

$$(18c),\ (18e), (18f)$$

Since the lower-level problems in (37b) and (37c) are LPs, **MMP**^{I} can be converted into a single-level formulation. Especially by applying the duality-based technique, a big-M-free one, similar to (21), can be obtained to facilitate easy computation. We also note that (37b) and (37c) are defined over $\widehat{\boldsymbol{\xi}}$ and $\widehat{\mathcal{Z}}$ to simplify **MMP**^{I}'s structure. Actually, given the scale of $|\widehat{\boldsymbol{\xi}}| \times |\widehat{\mathcal{Z}}|$, it could be computationally more friendly to define (37b) on $\widehat{\boldsymbol{\xi}}^o$ and $\widehat{\mathcal{Z}}^o$ only, which might not be significantly weaker than the current one.

With all master and subproblems revised according to $\mathcal{P}^I$, we next list modifications on particular steps of **C&CG − DRO**.

Step 1: Additional initialization includes $\widehat{\mathcal{Z}}^f = \widehat{\mathcal{Z}}^o = \varnothing$.

Step 2: **MMP** is replaced by (the single-level reformulation of) **MMP**^{I}.

Step 3: **WCEV**(F) is replaced by **WCEV**$^I(F)$, and the variant of the adopted oracle also outputs optimal $\mathbf{z}^f(\mathbf{x}^*)$.

Step 4-Case A: $\mathbf{WCEV}(O)$ and $\mathbf{MMP}$ are replaced by $\mathbf{WCEV}^{\boldsymbol{I}}(O)$ and $\mathbf{MMP}^{\boldsymbol{I}}$, the adopted oracle also outputs optimal $\mathbf{z}^o(\mathbf{x}^*)$, and additional update operation $\widehat{Z}^o = \widehat{Z}^o \cup \{\mathbf{z}^o(\mathbf{x}^*)\}$ is included.

Step 4-Case B: $\mathbf{WCEV}(F)$ is replaced by $\mathbf{WCEV}^{\boldsymbol{I}}(F)$, and additional update operation $\widehat{Z}^f = \widehat{Z}^f \cup \{\mathbf{z}^f(\mathbf{x}^*)\}$ is included.

With the aforementioned changes, the updated algorithm is referred to as **C&CG − DRO**($\mathcal{P}^{\boldsymbol{I}}$). We mention that it can be further extended to handle Wasserstein metric-based mixed integer ambiguity sets. As this extension is rather a straightforward integration with the results presented in Section 5.1, we omit its description to avoid redundancy.

Remark 14. (i) Analyses on the convergence and iteration complexity for **C&CG − DRO**($\mathcal{P}^{\boldsymbol{W}}$) follows directly from Sections 3.3 and 4.3 as $\mathcal{P}^{\boldsymbol{W}}$ belongs to the general ambiguity set presented in (3). Such analyses for **C&CG − DRO**($\mathcal{P}^{\boldsymbol{I}}$) can be developed in a way similar to those presented Sections 3.3 and 4.3, noting that $\mathcal{P}^{\boldsymbol{I}}$ reduces to the form of $\mathcal{P}$ for any fixed $\mathbf{z}$ and set $\mathcal{Z}$ is finite.
(ii)We mention that the presented C&CG-DRO, together with oracles for **WCEV** problems, actually provides a strong and flexible platform to compute a broad class of **2 − Stg DRO** problems. On one hand, the ambiguity set does not need to adhere to any standard form or be defined by any specific mathematical structure. For example, it can be captured by mixing moment inequalities and Wasserstein metric-based consideration. Also, it is comparable to solution methodologies developed for RO and SP. Note that, as long as master and sub- problems can be solved, there are no strict restrictions imposed on the underlying decision-making problem. More importantly, the overall development requires only a basic understanding of probability, LP, and MIP. This simplicity makes it easy to understand, modify, and debug. □

6 Numerical Studies

In this section, we present and discuss numerical results obtained from computational experiments. Our focus is on testing, evaluating and analyzing **C&CG − DRO** and its variants in computing **2 − Stg DRO** instances. The facility location model with different structures or considerations is employed as the testbed, which often arises from various applications in logistics, supply chain and healthcare systems. Data regarding clients and facilities' locations, distances and basic demands are adopted from [37]. All solution methods are implemented by Python 3.6, with professional MIP solver Gurobi 9.5.2 on a Windows PC with E5-1620 CPU and 32G RAM. Unless noted otherwise, the time limit is set to 7200s, and the relative optimality tolerance of any algorithm/solver is set to .5%.

6.1 Distributionally Robust Facility Location Models

Consider a facility location problem that builds $p \geq 1$ facilities. Let I represent the set of client sites, and $J \subseteq I$ the set of potential facility sites. The parameter c_{ij} captures the service cost of a unit demand from client i served by facility j, with $c_{ij} = 0$ if $i = j$. Moreover, d_i is the demand of client i, and f_j is the fixed construction cost of a facility at j with F_j being its capacity once the facility is established. The decision maker seeks a solution of the minimum cost consisting of the fixed construction cost and the weighted sum of the service cost for the normal situation (i.e., nominal demand $\bar{\mathbf{d}}$ with no disruptions) and the expected worst-case service cost within an ambiguity set. Under the DRO scheme, we investigate the impact of three major factors. They are the sample space, which could be either continuous or discrete, the ambiguity set, which could be either moment- or Wasserstein metric-based, and the recourse problem, which may or may not have the feasibility issue. Hence, there are 8 different combinations that will be used to generate instances and to perform our computational study. In the following, we present mathematical formulations, with ρ being the weight coefficient.

The first basic formulation considers continuous random demands $\mathbf{d} = (d_1, \ldots, d_{|I|})$.

$$\mathbf{FL-DRO(d)}: \min_{(\mathbf{x},\mathbf{y})\in\mathcal{X}} \sum_{j\in J} f_j y_j + \rho \sum_{i\in I}\sum_{j\in J} c_{ij}x_{ij} + (1-\rho)\max_{P\in\mathcal{P}_{\mathrm{d}}} E_P[Q^{\mathbf{d}}(\mathbf{y},\mathbf{d})] \tag{38}$$

with $\mathcal{X} = \Big\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{|I|\times|J|} \times \{0,1\}^{|J|} : \sum_{j\in J} x_{ij} \geq \bar{d}_i \ \ \forall i \in I, \ \sum_{j\in J} y_j = p, \ \sum_{i\in I} x_{ij} \leq F_j y_j \ \ \forall j \in J\Big\}$, and

$$Q^{\mathbf{d}}(\mathbf{y},\mathbf{d}) = \min\Big\{\sum_{i\in I}\sum_{j\in J} c_{ij}w_{ij} : \sum_{j\in J} w_{ij} \geq d_i \ \ \forall i \in I, \ \sum_{i\in I} w_{ij} \leq F_j y_j \ \ \forall j \in J, \\ w_{ij} \geq 0 \ \ \forall i \in I, \ \forall j \in J\Big\}.$$

Note that variables $(\mathbf{x},\mathbf{y})$ in $\mathcal{X}$ are continuous and binary, respectively, representing demand allocations and yes/no construction decisions. Constraints in $\mathcal{X}$ require that all nominal demands are satisfied, the demand allocation can only be made if the facility is constructed, and the total allocation to that facility is subject to its capacity. For $Q^{\mathbf{d}}(\mathbf{y},\mathbf{d})$, $\mathbf{w}$ represents the demand allocation after the randomness of demand is materialized.

We highlight that when F_j is sufficiently large, the whole formulation reduces to the uncapacitated model and $Q^{\mathbf{d}}(\mathbf{y},\mathbf{d})$ is always feasible regardless of the choice of $\mathbf{y}$. Otherwise, $\mathbf{y}$ needs to be selected to ensure the feasibility of $Q^{\mathbf{d}}(\mathbf{y},\mathbf{d})$, which requires to eliminate infeasible $\mathbf{y}$'s in our computation. We consider both cases in our numerical study to understand and evaluate the challenge and the impact of Q's feasibility issue in $\mathbf{2-Stg\ DRO}$. As noted earlier, existing algorithms are not able to handle this challenge.

Regarding the underlying ambiguity set, the sample space of random demand is $\mathcal{D} = \Big\{\mathbf{d} \in$

$\mathbb{R}^{|I|} : d_i^- \le d_i \le d_i^+, \forall i \in [I], \Big\}$. We mainly consider the following moment- and Wasserstein metric-based ones (using L_1 norm), denoted by $\mathcal{P}_{\mathbf{d}}^m$ and $\mathcal{P}_{\mathbf{d}}^w$, respectively.

$$\mathcal{P}_{\mathbf{d}}^m = \Big\{P \in \mathcal{M}(\mathcal{D}, \mathcal{F}) : E_P[\mathbf{d}] \le \tilde{\boldsymbol{d}}\Big\}; \quad \mathcal{P}_{\mathbf{d}}^W = \Big\{P \in \mathcal{M}(\mathcal{D}, \mathcal{F}) : \mathcal{W}(P, P^e) \le \mathrm{r}_{\mathbf{d}}\Big\} \tag{40}$$

As C&CG-DRO (and its variants) is generally applicable, more sophisticated ambiguity sets are considered in Section 6.5.

Empirical distribution P^e in $\mathcal{P}_{\mathbf{d}}^W$ consists of a set of random samples drawn from $\mathcal{D}$, each with equal probability. Besides continuous demand, we also study discrete disruptions that cause a facility to be unavailable. The demands are then served by the survived facilities. The basic formulation for binary random disruptions, $\mathbf{u} = (u_1, \ldots, u_{|J|})$, is

$$\mathbf{FL-DRO(u)} : \min_{(\mathbf{x},\mathbf{y}) \in \mathcal{X}} \sum_{j \in J} f_j y_j + \rho \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} + (1-\rho) \max_{P \in \mathcal{P}_{\mathbf{u}}} E_P[Q^{\mathbf{u}}(\mathbf{y}, \mathbf{u})]. \tag{41}$$

For this random factor, we consider up to $k(\le p-1)$ disruptions in set J, and hence the sample space is $\mathcal{U} = \{\mathbf{u} \in \{0,1\}^{|J|} : \sum_{j \in J} u_j \le k \, \forall j \in J\}$. Such type of sample space, although is finite, is generally exponential with respect to $|J|$, rendering the enumeration is practically infeasible. The corresponding recourse problem is

$$Q^{\mathbf{u}}(\mathbf{y}, \mathbf{u}) = \min \Big\{ \sum_{i \in I} \sum_{j \in J} c_{ij} w_{ij} : \sum_{j \in J} w_{ij} \ge d_i \quad \forall i \in I, \ \sum_{i \in I} w_{ij} \le F_j y_j \quad \forall j \in J, \tag{42a}$$

$$\sum_{i \in I} w_{ij} \le F_j(1-u_j) \quad \forall j \in J, \ w_{ij} \ge 0 \quad \forall i \in I, \ \forall j \in J \Big\}. \tag{42b}$$

Similarly, following ambiguity sets are investigated.

$$\mathcal{P}_{\mathbf{u}}^m = \Big\{P \in \mathcal{M}(\mathcal{U}, \mathcal{F}) : E_P[\mathbf{u}] \le \tilde{\boldsymbol{k}}\Big\}; \quad \mathcal{P}_{\mathbf{u}}^W = \Big\{P \in \mathcal{M}(\mathcal{U}, \mathcal{F}) : \mathcal{W}(P, P^e) \le \mathrm{r}_{\mathbf{u}}\Big\}. \tag{43}$$

Also, empirical distribution P^e includes random samples of equal probability drawn from discrete sample space.

6.2 Computational Results of Oracles for WCEV Problems

We first perform a set of experiments to evaluate two primal oracles for **WCEV** problems, including **WCEV**(F) and **WCEV**(O). Instances belonging to "Small" group are with $|I| = |J| = 5$, $p = 3$. For the applicable combinations, the size of empirical set is 5, $\mathrm{r}_{\mathbf{d}} = 2$, $\mathrm{r}_{\mathbf{u}} = 0.5$, and $k = 2$ if $\mathcal{U}$ is adopted. For instances belonging to "Large" group, all parameters are same except that $|I| = |J| = 8$ and the size of the empirical set is 10.

All results are presented in Table 1, with the time limit set to 10 minutes. where label "T" indicates the algorithm terminates before generating an optimal solution. Under this situation, the optimality gap is reported if available, or marked as "-" otherwise. On one hand, it can be easily seen that *Oracle-1* is practically infeasible to apply. Only 3 out of 16 instances can be solved to optimality, with non-trivial amount of computational times. On the other hand, the computational time of *Oracle-2* for all instances is negligible. Overall, we can estimate that *Oracle-2* is drastically faster, potentially by many orders of magnitude. Hence, we just adopt *Oracle-2* in our remaining experiments.

Table 1: Computational Results of Two Oracles for **WCEV** Problems

Prob. Size	Ambiguity Sets	**WCEV**	Continuous Sample Space $\mathcal{D}$				Discrete Sample Space $\mathcal{U}$			
			Oracle-1		*Oracle-2*		*Oracle-1*		*Oracle-2*	
			Time(s)	Gap(%)	Time(s)	Gap(%)	Time(s)	Gap(%)	Time(s)	Gap(%)
Small	Moment	**WCEV**(F)	T	249	0.08	0	T	140	0.02	0
		WCEV(O)	T	11	0.07	0	113.9	0	0.04	0
	Wasserstein	**WCEV**(F)	T	-	0.13	0	23.5	0	0.01	0
		WCEV(O)	T	89	0.9	0	6.03	0	0.03	0
Large	Moment	**WCEV**(F)	T	353	0.00	0	T	1604	0.03	0
		WCEV(O)	T	56.5	0.44	0	T	385	0.14	0
	Wasserstein	**WCEV**(F)	T	-	0.23	0	T	14.5	0.05	0
		WCEV(O)	T	578	3.99	0	T	-	0.07	0

6.3 Computational Results of the Uncapacitated Case

Recall that the existing approaches of implementing C&CG and Benders decomposition to solve $\mathbf{2-Stg\ DRO}$ are referred to as basic C&CG and basic Benders decomposition methods, respectively. In this subsection, we set F_j large than all possible total demands for all j, i.e., the case without capacity restrictions. So, there is no feasibility challenge associated with the recourse problem, allowing us to benchmark our work with respect to the current literature. Next, we first investigate instances of $\mathbf{FL-DRO(d)}$ and then those of $\mathbf{FL-DRO(u)}$. Regarding moment-based ambiguity sets, we introduce parameter $\mathfrak{r}$ and a random vector $\mathfrak{v}$ to generate random $\tilde{\boldsymbol{d}}$ or $\tilde{\boldsymbol{k}}$: $\tilde{\boldsymbol{d}} = \bar{\mathbf{d}}(1 + \mathfrak{r} \times \mathfrak{v})$ with $\mathfrak{v} \in [-0.5 * \mathbf{1}, 0.5 * \mathbf{1}]$ and $\tilde{\boldsymbol{k}} = \mathfrak{r} \times \mathfrak{v}$ with $\mathfrak{v} \in [\mathbf{0}, \mathbf{1}]$. Also, we have $\rho = 0.5$ in all computational studies.

6.3.1 Results for Instances of Continuous Sample Space

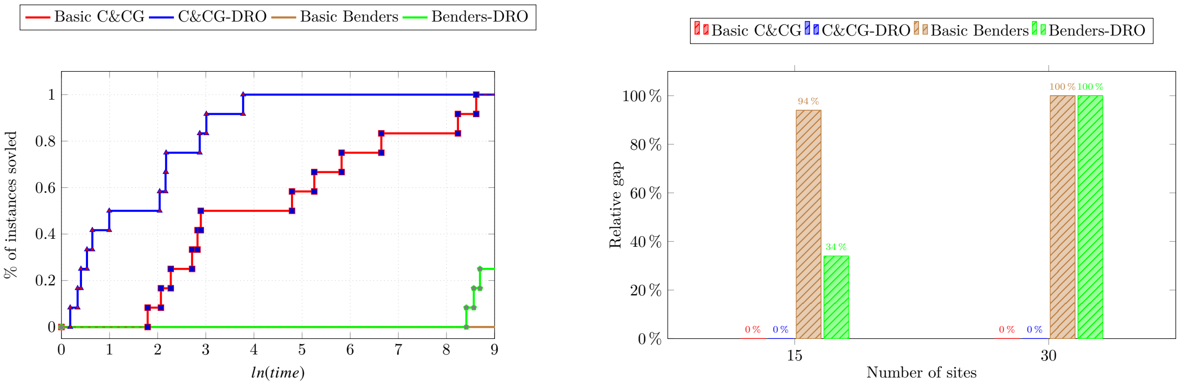
For instances of $\mathbf{FL-DRO(d)}$ with moment-based ambiguity set $\mathcal{P}^m_{\mathbf{d}}$, numerical results of basic C&CG, C&CG-DRO, basic Benders, and Benders-DRO algorithms are presented in Table 2. Columns "LB", "UB", "Iter." and "$|\widehat{\boldsymbol{\xi}}|$" report the lower and upper bounds, the number of main iterations and the number of scenarios generated when an algorithm terminates, respectively. Similar to Table 1, we mark the entry with "T" for "Time(s)" or "-" for "Gap(%)" if the relevant data is not available. Obviously, Benders type of algorithms are practically infeasible, noting that they are extremely uncompetitive compared to their C&CG counterparts. As for basic C&CG and C&CG-DRO, we also observe that the latter exhibits a clear and consistent advantage. For relatively easy instances that basic C&CG can solve within a few minutes, C&CG-DRO is faster by an order of magnitude. For more challenging instances, C&CG-DRO can be up to several hundred times faster. One explanation for such drastically different performances is that basic C&CG needs to perform many more iterations, given that it only generates a single scenario for every iteration. Another explanation is that C&CG-DRO identifies critical scenarios in an effective fashion. Note that it produces much fewer scenarios for main master problems, compared to basic C&CG, that are still able to provide a strong support to define worst-case distributions. Similar observations can be made between Basic Benders and Benders-DRO, i.e., Benders decomposition with *Oracle-2* to generate Benders cuts, where the latter one is also much faster. Also, we observe that Benders type of algorithms always produce a huge number of Benders cuts, which, however, are rather ineffective in the derivation of optimal solutions.

The overall performance profiles of all these four algorithms are plotted in Figure 2, where the dominance of C&CG-DRO is straightforward. To deeply understand algorithms' dynamic behaviors, Figure 3 presents the convergence trajectories of basic C&CG and C&CG-DRO for the case where $|I| = 30$, $\mathrm{r} = 0.65$ and $p = 6$. We do not include those of Benders type of algorithms due to their very weak performances. It can be seen that C&CG-DRO quickly converges to an optimal solution, especially with the optimal solution identified within a couple of iterations. This observation also confirms the importance of solving **WCEV** problem. It not only provides an exact evaluation for a first-stage solution, but also generates a set of scenarios that are critical to select a high quality first-stage solution.

The aforementioned pattern among those algorithms holds for computational results of instances with Wasserstein metric-based ambiguity set $\mathcal{P}_{\mathbf{d}}^{W}$, which are presented in Table 3. Typically, C&CG-DRO is faster than basic C&CG by 1 to 2 orders of magnitude, while Benders type of algorithms remain very uncompetitive. It is also interesting to point out, regardless the solution algorithm we employed, one difference between results for $\mathbf{FL-DRO(d)}$ with $\mathcal{P}_{\mathbf{d}}^{m}$ and with $\mathcal{P}_{\mathbf{d}}^{W}$: the number of (main) iterations for the former one is much larger than that of the latter one. We believe that the reason behind is that the empirical set underlying $\mathcal{P}_{\mathbf{d}}^{W}$ has a determinant impact on its structure. When more samples are available, it is easier to identify the worst-case distribution, which is observed in Table 3. Moreover, it is worth highlighting that for some instances, $|\widehat{\boldsymbol{\xi}}|$ is equal to the product between the number of iterations and $N+1$, e.g., instances (in the form of $I-\mathrm{r}_{\mathbf{d}}-\mathrm{N}-\mathrm{p}$) $15-10-50-6$ and $30-10-100-10$. Recall that the theoretical analysis presented in Propositions 5 and the structure of $\mathcal{P}_{\mathbf{d}}^{W}$ indicate that the number of scenarios with non-zero probability is not more than $N+1$ (for every iteration). Clearly, it is verified by the results of those instances, which actually are obtained by *Oracle-2* that does not depend on Proposition 5.

Table 2: Computational Results of Uncapacitated Models with $\mathcal{P}_{\mathbf{d}}^{m}$

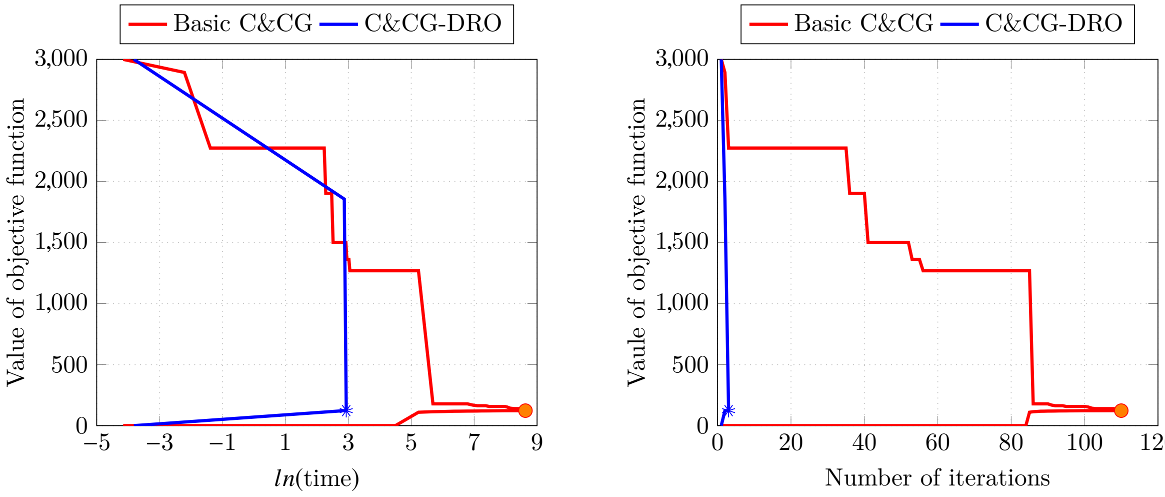
$\lvert I\rvert$	r	p	Basic C&CG						C&CG-DRO					
			LB	UB	Gap(%)	Iter.	$\lvert\widehat{\xi}\rvert$	Time(s)	LB	UB	Gap(%)	Iter.	$\lvert\widehat{\xi}\rvert$	Time(s)
15	0.65	6	51.3	51.3	0	52	52	18.1	51.3	51.3	0	3	13	1.9
		10	16.5	16.5	0	71	71	9.7	16.5	16.5	0	3	12	1.4
	0.80	6	49.2	49.2	0	83	83	16.9	49.1	49.2	0	3	19	1.7
		10	15.8	15.8	0	50	50	6.0	15.8	15.8	0	3	14	1.5
	0.95	6	47.2	47.2	0	58	58	15.1	47.2	47.2	0	4	22	2.7
		10	15.1	15.1	0	66	66	7.9	15.1	15.1	0	3	12	1.2
	Average				**0**	**63.3**	**63.3**	**12.3**			**0**	**3.2**	**15.3**	**1.7**
30	0.65	6	122.8	122.8	0	109	109	5529.5	122.7	122.8	0	3	29	20.3
		10	70.8	70.8	0	40	40	120.4	70.8	70.8	0	3	29	8.7
	0.80	6	118.3	118.7	0	46	46	770.0	118.2	118.3	0	3	41	17.7
		10	68.2	68.2	0	47	47	191.2	68.2	68.2	0	3	25	8.8
	0.95	6	113.9	114.1	0	76	76	3782.2	113.9	113.9	0	3	43	43.6
		10	65.1	65.1	0	58	58	337.3	65.1	65.1	0	3	27	7.7
	Average				**0**	**62.6**	**62.6**	**1788.43**			**0**	**3.0**	**32.3**	**17.8**
$\lvert I\rvert$	r	p	Basic Benders						Benders-DRO					
			LB	UB	Gap(%)	Iter.	Cuts	Time(s)	LB	UB	Gap(%)	Iter.	Cuts	Time(s)
15	0.65	6	13.5	97.0	86	2889	2889	T	51.1	51.3	0	250	24975	5977.9
		10	0.0	38.9	100	13646	13646	T	5.5	17.3	68	389	27631	T
	0.80	6	11.1	94.3	88	3231	3231	T	49.1	49.2	0	231	21545	4503.7
		10	0.0	42.7	100	13755	13755	T	4.8	15.8	70	390	29488	T
	0.95	6	9.2	93.7	90	3424	3424	T	47.1	47.2	0	235	23769	5257.1
		10	0.0	44.5	100	13760	13760	T	5.6	15.1	63	329	28968	T
	Average				**94**	**8450.8**	**8450.8**				**34**	**304.0**	**26062.7**	
30	0.65	6	0.0	248.8	100	8063	8063	T	1.0	141.0	99	95	28135	T
		10	0.0	171.7	100	9423	9423	T	0.0	79.3	100	422	90730	T
	0.80	6	0.0	270.8	100	8108	8108	T	0.0	127.2	100	98	28444	T
		10	0.0	146.5	100	9249	9249	T	0.0	82.4	100	421	91524	T
	0.95	6	0.0	276.7	100	8173	8173	T	1.1	120.8	99	110	31051	T
		10	0.0	163.1	100	8816	8816	T	0.0	79.2	100	400	95203	T
	Average				**100**	**8638.7**	**8638.7**				**100**	**257.7**	**60847.8**	



(a) % of Instances Solved over $ln(time)$

(b) Average Gaps upon Termination

Figure 2: Overall performance profiles of Four Algorithms



(a) Convergence over Time

(b) Convergence over Iterations

Figure 3: Convergence Plots of Computing $\mathbf{FL-DRO(u)}$ with $\mathcal{P}_{\mathbf{d}}^{m}$

Table 3: Computational Results of Uncapacitated Models with $\mathcal{P}_{\mathbf{d}}^{W}$

I	$\mathrm{r}_{\mathbf{d}}$	N	p	Basic C&CG						C&CG-DRO					
				LB	UB	Gap(%)	Iter.	$\lvert\widehat{\xi}\rvert$	Time(s)	LB	UB	Gap(%)	Iter.	$\lvert\widehat{\xi}\rvert$	Time(s)
15	2	50	6	62.3	62.6	0	5	250	114.1	62.4	62.4	0	2	80	26.0
			10	20.1	20.2	0	3	150	12.9	20.1	20.1	0	1	14	5.1
		100	6	62.1	62.3	0	6	600	726.7	62.3	62.3	0	2	139	52.5
			10	20.0	20.1	0	3	300	40.3	20.0	20.0	0	1	27	10.4
		200	6	62.1	62.3	0	5	1000	3077.7	62.1	62.1	0	2	185	98.7
			10	19.9	20.0	0	3	600	134.5	19.9	19.9	0	1	31	16.7
		Average			**41.3**	**0**	**4.2**	**483.3**	**684.4**		**41.2**	**0**	**1.5**	**79.3**	**34.9**
	10	50	6	63.7	63.9	0	5	250	214.0	63.8	64.0	0	3	153	70.2
			10	21.2	21.2	0	3	150	12.5	21.2	21.2	0	1	51	5.4
		100	6	63.6	63.8	0	5	500	770.8	63.7	63.8	0	3	303	179.4
			10	21.1	21.2	0	3	300	43.9	21.1	21.1	0	1	101	12.0
		200	6	63.4	63.7	0	4	800	1719.3	63.3	63.6	0	2	402	190.5
			10	21.0	21.0	0	3	600	129.5	21.0	21.0	0	1	148	20.9
		Average			**42.5**	**0**	**3.8**	**433.3**	**481.6**		**42.5**	**0**	**1.8**	**193.0**	**79.7**
30	2	50	6	143.8	144.4	0	4	200	1105.4	144.2	144.2	0	1	44	130.0
			10	84.1	84.4	0	4	200	400.4	84.4	84.4	0	1	51	52.4
		100	6	143.6	144.1	0	4	400	4803.8	144.0	144.0	0	1	85	596.8
			10	83.9	84.2	0	4	400	1561.6	84.2	84.2	0	1	101	222.3
		200	6	143.0	144.8	1	3	600	T	143.5	143.5	0	1	77	1435.9
			10	83.7	84.0	0	4	800	6587.7	83.9	83.9	0	1	109	487.9
		Average			**114.3**	**0**	**3.8**	**433.3**			**114.0**	**0**	**1.0**	**77.8**	**487.6**
	10	50	6	146.6	147.2	0	3	150	1235.2	147.1	147.1	0	1	51	154.6
			10	86.3	86.3	0	3	150	500.5	86.3	86.3	0	1	51	49.2
		100	6	146.3	147.0	0	3	300	5120.4	146.8	146.8	0	1	101	723.6
			10	86.1	86.2	0	3	300	1664.0	86.2	86.2	0	1	101	216.8
		200	6	144.8	146.7	1	2	400	T	146.5	146.5	0	1	201	2303.9
			10	85.8	86.3	1	2	400	T	86.0	86.0	0	1	201	727.3
		Average			**116.6**	**0**	**2.7**	**283.3**			**116.5**	**0**	**1.0**	**117.7**	**695.9**

6.3.2 Results for Instances of Discrete Sample Space

Given the extremely poor performance of Benders type of algorithms, we do not consider them in the remainder of this section, unless otherwise noted. For **FL − DRO(u)**, the number of sites under consideration is smaller than that for **FL − DRO(d)** due to its computational challenge. Except for parameter k, other parameters are kept the same as those in Tables 2 and 3. Computational results for **FL − DRO(u)** with $\mathcal{P}^m_{\mathbf{u}}$ and $\mathcal{P}^W_{\mathbf{u}}$ are presented in Tables 4 and 5, respectively.

From Tables 4 and 5, it can be seen that both algorithms are more sensitive to parameters determining the size of instances, including $|I|$ and N. A few instances have not been solved to the exactness by either algorithm, indicating that **FL − DRO(u)** is more complex than **FL − DRO(d)**. The number of iterations is generally much greater than that of its counterpart **FL − DRO(d)**. The same observation holds for the number of scenarios involved in. One explanation is that the non-convex structure of the discrete sample space renders the associated ambiguity sets more intricate. As a result, a large number of iterations needs to be performed to generate many scenarios that are critical to capture the worst-case distributions. Especially for Wasserstein metric-based $\mathcal{P}^W_{\mathbf{u}}$, we note that the number of scenarios generated could be larger than 2,000, which is drastically more than that for moment-based $\mathcal{P}^m_{\mathbf{u}}$. As the number of scenarios largely determines the scale of **MMP**, it deserves a deep investigation to develop efficient algorithms for **MMP** when the number of empirical samples is large. Another observation from handling $\mathcal{P}^W$ is that the numbers of iterations could become smaller with respect to larger empirical sets, which verifies the effectiveness to employ those samples for initialization. Between basic C&CG and C&CG-DRO, they have roughly the same performance for simple instances. Yet, for more challenging instances, C&CG-DRO demonstrates a much stronger capacity. It either can solve instances that cannot be computed by basic C&CG, or produces solutions with significantly smaller gaps.

Table 4: Computational Results of Uncapacitated Models with $\mathcal{P}^m_{\mathbf{u}}$

$\lvert I\rvert$	r	p	k	Basic C&CG						C&CG-DRO					
				LB	UB	Gap(%)	Iter.	$\lvert\hat{\xi}\rvert$	Time(s)	LB	UB	Gap(%)	Iter.	$\lvert\hat{\xi}\rvert$	Time(s)
15	0.65	6	3	230.1	230.1	0	120	120	1105.3	229.1	230.1	0	67	99	676.1
		10	5	182.4	182.4	0	164	164	790.3	181.6	182.4	0	64	179	530.5
	0.80	6	3	236.3	236.3	0	105	105	640.9	235.5	236.3	0	30	59	124.5
		10	5	195.1	195.1	0	87	87	173.0	194.2	195.1	0	26	81	109.1
	0.95	6	3	242.6	242.6	0	93	93	417.9	242.2	242.6	0	12	25	15.8
		10	5	207.8	207.8	0	45	45	51.7	207.8	207.8	0	13	40	35.8
	Average				**215.7**	**0**	**102.3**	**102.3**	**529.9**		**215.7**	**0**	**35.3**	**80.5**	**248.6**
20	0.65	6	3	215.9	299.2	28	107	107	T	266.4	280.3	5	74	118	T
		10	5	197.2	237.6	17	138	138	T	208.1	227.6	9	59	234	T
	0.80	6	3	220.5	299.2	26	104	104	T	280.3	288.6	3	76	124	T
		10	5	212.2	246.6	14	167	167	T	226.3	233.2	3	59	219	T
	0.95	6	3	226.2	299.2	24	116	116	T	293.3	296.5	1	92	123	T
		10	5	237.6	254.3	7	165	165	T	239.5	240.2	0	45	151	2318.8
	Average				**272.7**	**19**	**132.8**	**132.8**			**261.1**	**3**	**67.5**	**161.5**	

Table 5: Computational Results of Uncapacitated Models with $\mathcal{P}_{\mathbf{u}}^{W}$

$\lvert I \rvert$	$\mathrm{r}_{\mathbf{u}}$	N	p	k	Basic C&CG						C&CG-DRO					
					LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)	LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)
15	0.5	50	6	3	201.8	201.9	0	4	150	29.4	201.9	201.9	0	3	91	25.8
			10	5	181.9	181.9	0	4	138	113.2	181.5	181.9	0	4	104	162.3
		100	6	3	194.3	194.3	0	4	292	72.9	194.2	194.5	0	3	251	49.7
			10	5	181.1	181.9	0	3	263	176.9	181.0	181.1	0	3	162	271.4
		200	6	3	189.8	189.8	0	4	568	204.1	189.4	189.8	0	4	551	173.1
			10	5	168.7	169.5	0	3	390	365.1	168.9	169.6	0	2	246	418.3
		Average				**186.5**	**0**	**3.7**	**300.2**	**160.3**		**186.5**	**0**	**3.2**	**234.2**	**183.4**
	2	50	6	3	234.8	234.8	0	11	536	317.6	234.1	234.8	0	11	488	236.4
			10	5	200.2	200.7	0	6	207	189.9	200.1	200.7	0	7	280	555.9
		100	6	3	229.0	229.0	0	9	681	442.0	228.6	229.3	0	7	667	332.2
			10	5	198.2	198.6	0	7	522	559.7	198.2	198.6	0	6	541	865.9
		200	6	3	226.4	227.4	0	10	1721	2934.4	226.3	226.4	0	9	1884	3291.1
			10	5	192.8	193.1	0	7	980	1214.7	192.6	193.4	0	6	851	1985.6
		Average				**213.9**	**0**	**8.3**	**774.5**	**943.4**		**213.9**	**0**	**7.7**	**785.2**	**1211.2**
20	0.5	50	6	3	173.2	386.1	55	16	800	T	181.2	192.8	6	36	1767	T
			10	5	183.8	184.7	0	6	249	334.6	184.0	184.7	0	4	120	303.2
		100	6	3	173.5	475.8	64	8	800	T	178.2	205.3	13	22	2118	T
			10	5	171.2	171.3	0	4	275	407.5	170.5	171.3	0	2	137	233.7
		200	6	3	169.8	452.5	62	5	1000	T	183.2	318.0	42	8	1471	T
			10	5	146.4	147.0	0	3	537	967.9	146.5	146.5	0	2	325	480.6
		Average				**302.9**	**30**	**7**	**610.2**			**203.1**	**10**	**12.3**	**989.7**	
	2	50	6	3	187.8	320.8	41	14	700	T	220.6	259.8	15	22	1112	T
			10	5	213.5	215.9	1	21	858	T	212.8	213.8	0	13	513	1604.5
		100	6	3	178.8	521.6	66	8	800	T	189.4	312.6	39	12	1124	T
			10	5	210.1	210.8	0	9	663	4590.6	210.4	210.8	0	11	624	3920.4
		200	6	3	171.5	521.6	67	4	800	T	183.2	318.0	42	8	1471	T
			10	5	196.9	197.4	0	5	650	2740.6	196.9	197.4	0	5	923	3028.5
		Average				**331.3**	**29**	**10.2**	**745.2**			**252.1**	**16**	**11.8**	**961.2**	

6.4 Computational Results of the Capacitated Case

In this subsection, we present computational results for instances of various models with capacity considerations. Capacity parameters F_j are set to the total demands from a random number of neighboring sites for $j \in J$. Parameters determining the size of instances, e.g., $|I|$, p and k, are set to smaller values compared to those for the uncapacitated case, due to the increased computational challenge. All results are presented in Tables 6 and 7, noting that column “$|\widehat{\boldsymbol{\xi}}|^o$” reports the number of scenarios in the optimality set, and column “$|\widehat{\boldsymbol{\xi}}|$” shows the number of scenarios in the complete set from solving both **WCEV**(O) and **WCEV**(F). Hence, the number of scenarios in the feasibility set is the difference between them.

From those tables, it is clear that those instances are more difficult to solve. On average, the number of iterations, the computational time, and the optimality gap are all significantly more than those of uncapacitated instances. It is worth noting that the number of scenarios in the feasibility sets is often non-trivial. Especially for instances with a discrete sample space, it can outnumber the optimality one by an order of magnitude. This indicates that the importance of detecting and addressing the infeasibility issue in practice, although it has been overlooked in the existing literature. Compared to the uncapacitated case, we also note that it is more clear that the number of iterations is often smaller for larger N, which shows that empirical samples are very useful in addressing the complex structure of the capacitated case.

Table 6: Computational Results of Capacitated Models on $\mathcal{P}^m$

$\lvert I\rvert$	r	p	$\mathcal{P}^m_{\mathbf{d}}$							p	k	$\mathcal{P}^m_{\mathbf{u}}$						
			LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}^o\rvert$	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)			LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}^o\rvert$	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)
15	0.65	6	52.8	52.8	0	5	27	38	23.0	6	3	322.3	322.3	0	88	12	106	754.8
		10	16.5	16.5	0	4	11	24	5.0	10	5	183.6	184.3	0	53	107	107	236.2
	0.80	6	51.4	51.4	0	5	24	39	24.8	6	3	331.4	332.6	0	89	9	105	642.1
		10	15.8	15.8	0	4	16	30	5.4	10	5	200.7	201.7	0	45	88	88	158.5
	0.95	6	50.8	50.8	0	4	24	31	27.5	6	3	340.3	340.7	0	100	5	125	676.1
		10	15.1	15.1	0	4	12	25	6.6	10	5	213.3	214.0	0	17	37	37	29.4
	Average				**0**	**4.3**	**19.0**	**31.2**	**15.4**	**Average**				**0**	**65.3**	**43.0**	**94.7**	**416.2**
20	0.65	6	81.8	81.9	0	5	47	64	66.1	6	3	280.5	324.9	14	98	66	125	T
		10	37.3	37.3	0	4	22	38	24.1	10	5	211.1	228.7	8	70	172	172	T
	0.80	6	80.7	81.0	0	5	46	64	76.2	6	3	295.7	334.0	11	105	61	133	T
		10	35.9	35.9	0	4	25	47	28.5	10	5	230.9	245.8	6	73	178	178	T
	0.95	6	79.7	79.9	0	5	41	53	87.9	6	3	306.8	339.9	10	115	63	146	T
		10	34.3	34.4	0	4	26	40	34.6	10	5	247.0	255.8	3	85	199	199	T
	Average				**0**	**4.5**	**34.5**	**51.0**	**52.9**	**Average**				**9**	**91.0**	**123.2**	**158.8**	

Table 7: Computational Results of Capacitated Models on $\mathcal{P}^W$

$\lvert I\rvert$	$\mathrm{r}_{\mathbf{d}}$	N	p	$\mathcal{P}^W_{\mathbf{d}}$							$\mathrm{r}_{\mathbf{u}}$	p	k	$\mathcal{P}^W_{\mathbf{u}}$						
				LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}^o\rvert$	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)				LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}^o\rvert$	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)
15	2	50	6	93.2	93.2	0	2	2	4	114.4	0.5	6	3	286.3	286.3	0	18	13	233	148.2
			10	20.1	20.1	0	1	14	14	194.5		10	5	184.0	184.4	0	1	13	13	26.8
		100	6	92.8	92.8	0	2	5	9	272.5		6	3	255.3	255.3	0	6	33	133	117.6
			10	20.0	20.0	0	1	27	27	371.6		10	5	183.2	183.4	0	3	50	50	138.5
		200	6	92.3	92.3	0	2	8	15	445.3		6	3	241.5	241.5	0	3	75	125	206.1
			10	19.9	19.9	0	1	31	31	585.3		10	5	171.3	171.4	0	2	72	72	343.5
		Average				**0**	**1.5**	**14.5**	**16.7**	**330.6**		**Average**				**0**	**5.5**	**42.7**	**104.3**	**163.4**
	10	50	6	98.2	98.2	0	2	10	18	118.7	2	6	3	331.8	331.8	0	45	71	1341	1640.4
			10	21.2	21.2	0	1	51	51	191.1		10	5	204.9	205.7	0	6	208	208	209.0
		100	6	97.9	97.9	0	2	20	35	262.6		6	3	316.5	316.7	0	22	125	1334	1083.4
			10	21.1	21.1	0	1	101	101	362.2		10	5	202.1	203.0	0	5	308	308	358.7
		200	6	97.4	97.4	0	2	38	68	447.7		6	3	309.6	310.2	0	13	233	1404	1346.4
			10	21.0	21.0	0	1	148	148	589.3		10	5	196.1	197.0	0	6	595	595	1001.4
		Average				**0**	**1.5**	**61.3**	**70.2**	**328.6**		**Average**				**0**	**16.2**	**256.7**	**865.0**	**939.9**
20	2	50	6	114.3	114.3	0	1	51	51	331.5	0.5	6	3	197.0	197.5	0	30	173	213	5465.2
			10	44.1	44.1	0	1	43	43	594.6		10	5	184.5	185.2	0	3	25	25	99.5
		100	6	114.0	114.0	0	1	95	95	717.2		6	3	190.0	199.5	5	18	243	269	T
			10	44.0	44.0	0	1	60	60	1037.5		10	5	170.9	171.7	0	2	29	29	158.4
		200	6	113.7	113.7	0	1	98	98	1485.1		6	3	180.8	210.2	14	9	464	464	T
			10	43.8	43.8	0	1	41	41	1556.9		10	5	147.3	147.6	0	1	50	50	198.8
		Average				**0**	**1.0**	**64.7**	**64.7**	**953.8**		**Average**				**3**	**10.5**	**164.0**	**175.0**	
	10	50	6	118.4	118.4	0	1	51	51	316.0	2	6	3	207.3	277.2	25	17	397	397	T
			10	45.5	45.5	0	1	51	51	443.3		10	5	213.4	214.2	0	10	188	188	723.7
		100	6	118.1	118.1	0	1	101	101	695.8		6	3	191.2	290.1	34	9	432	432	T
			10	45.4	45.4	0	1	101	101	781.1		10	5	212.0	212.8	0	7	321	321	1925.9
		200	6	117.8	117.8	0	1	150	150	2047.9		6	3	186.9	357.5	48	6	697	697	T
			10	45.4	45.4	0	1	191	191	1901.9		10	5	198.7	199.4	0	5	568	568	3802.5
		Average				**0**	**1.0**	**107.5**	**107.5**	**1031.0**		**Average**				**18**	**9.0**	**433.8**	**433.8**	

6.5 Computational Results of Complex Ambiguity Sets

In this subsection, we consider uncapacitated **FL − DRO** defined on more sophisticated ambiguity sets. Note that our purpose here is to demonstrate C&CG-DRO's (with *Oracle-2*) capacity to handle complex problems, rather than to provide a comprehensive evaluation. One type of ambiguity sets is Wasserstein metric-based set defined by L_2 norm, denoted by $\mathcal{P}_{\mathbf{d}}^{W_2}$. Note that it has a second-order conic (SOC) structure and requires SOC programming solver within *Oracle-2* to compute **PSP** problems. Another type is mixed integer ambiguity sets extending $\mathcal{P}_{\mathbf{u}}^{m}$ as in the following, which is denoted by $\mathcal{P}_{\mathbf{u}}^{mI}$.

$$\mathcal{P}_{\mathbf{u}}^{mI} = \Big\{P \in \mathcal{M}(\mathcal{U},\mathcal{F}) : \underline{\boldsymbol{k}} - \theta\mathbf{z} \le E_P[\mathbf{u}] \le \tilde{\boldsymbol{k}} - \theta\mathbf{z},\ \sum_i z_i \ge z^0,\ z_i \in \{0,1\}\ \ \forall i \in I\Big\}.$$

In our numerical study, $\underline{\boldsymbol{k}} = 0.4 * \mathbf{1}$, $\tilde{\boldsymbol{k}} = 0.8 * \mathbf{1}$, θ is set to 0.4 and z^0 to 2. All results are presented in Table 8. Compared to the results in Tables 3 and 4, there is no significant difference in the computational performance, including time and the number of iterations. The C&CG-DRO method remains effective for handling L_2 norm-defined Wasserstein metric-based ambiguity sets, while it is still sensitive to the instance scale when the sample space is discrete. However, we note that this observation is based on this small-scale study and is not conclusive.

Table 8: Computational Results of Uncapacitated Models with $\mathcal{P}_{\mathbf{d}}^{W_2}$ and $\mathcal{P}_{\mathbf{u}}^{mI}$

$\lvert I\rvert$	$\mathrm{r}_{\mathbf{d}}^2$	N	p	SOC Ambiguity Set ($\mathcal{P}_{\mathbf{d}}^{W_2}$)						r	p	k	MIP Ambiguity Set ($\mathcal{P}_{\mathbf{u}}^{mI}$)					
				LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)				LB	UB	Gap(%)	Iter.	$\lvert\widehat{\boldsymbol{\xi}}\rvert$	Time(s)
15	2	50	6	62.4	62.7	0	1	51	31.6	0.65	6	3	235.6	236.4	0	34	69	216.4
		100	6	62.3	62.5	0	1	101	58.5		10	5	193.3	194.2	0	30	76	103.0
		200	6	62.1	62.4	0	1	201	135.2	0.80	6	3	235.4	236.4	0	34	64	195.3
	30	50	6	64.4	64.7	0	2	98	72.3		10	5	194.7	195.1	0	27	58	83.4
		100	6	64.3	64.6	0	2	198	150.9	0.95	6	3	235.4	236.4	0	33	63	191.1
		200	6	64.2	64.4	0	2	397	333.3		10	5	195.0	195.2	0	27	60	68.1
	Average					**0**	**1.5**	**174.3**	**130.3**	**Average**					**0**	**30.8**	**65.0**	**68.1**
20	2	50	6	93.6	93.6	0	1	51	72.0	0.65	6	3	276.6	289.3	4	70	130	T
		100	6	93.4	93.4	0	1	101	217.3		10	5	212.5	225.2	6	69	165	T
		200	6	93.2	93.2	0	1	201	776.9	0.80	6	3	279.7	290.5	4	74	122	T
	30	50	6	96.2	96.6	0	1	51	69.2		10	5	225.4	233.2	3	65	155	T
		100	6	96.1	96.4	0	1	101	222.4	0.95	6	3	280.5	293.2	4	65	133	T
		200	6	96.0	96.3	0	1	201	687.3		10	5	224.9	233.2	4	66	170	T
	Average					**0**	**1.0**	**117.7**	**340.9**	**Average**					**4**	**68.2**	**145.8**	

7 Conclusions

In this paper, rather than following the dual perspective that is popular in the literature, we present a new study on two-stage DRO by taking the primal perspective. This perspective allows us to gain a deeper and more intuitive understanding on DRO, to develop a general and fast decomposition algorithm (and its variants) by leveraging existing powerful solution methods, and to address a couple of unsolved issues underlying two-stage DRO. Theoretical analyses regarding the strength, convergence, and iteration complexity of the developed algorithm are also presented. A systematic numerical study on the distributionally robust facility location problem has been conducted, taking into account multiple critical factors.. Results clearly demonstrate that our new solution algorithm (and its variants) is generally applicable and achieves remarkable superiority over existing methods, often solving problems up to several orders of magnitude faster.

Regarding future research directions, it would be interesting to extend our primal perspective to consider other types of risk measures beyond the expected value under the DRO framework. Also, enhancing the developed algorithm (and its variants) and promoting it in solving large-scale and complex data-driven problems will be carried out to support various real-world systems.

References

[1] Herbert Scarf. A min max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 1958.

[2] Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

[3] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[4] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[5] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[6] Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[7] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[8] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[9] Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018.

[10] Aharon Ben-Tal, Alexander Goryashko, Elana Guslitzer, and Arkadi Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical programming*, 99(2):351–376, 2004.

[11] Dimitris Bertsimas, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

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

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

[14] Ali Bagheri, Chaoyue Zhao, Feng Qiu, and Jianhui Wang. Resilient transmission hardening planning in a high renewable penetration era. *IEEE Transactions on Power Systems*, 34(2):873–882, 2018.

[15] Manish Bansal, Kuo-Ling Huang, and Sanjay Mehrotra. Decomposition algorithms for two-stage distributionally robust mixed binary programs. *SIAM Journal on Optimization*, 28(3):2360–2383, 2018.

[16] Ahmed Saif and Erick Delage. Data-driven distributionally robust capacitated facility location problem. *European Journal of Operational Research*, 291(3):995–1007, 2021.

[17] 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.

[18] Harsha Gangammanavar and Manish Bansal. Stochastic decomposition method for two-stage distributionally robust linear optimization. *SIAM Journal on Optimization*, 32(3):1901–1930, 2022.

[19] Daniel Duque, Sanjay Mehrotra, and David P Morton. Distributionally robust two-stage stochastic programming. *SIAM Journal on Optimization*, 32(3):1499–1522, 2022.

[20] Alexander Shapiro. *On Duality Theory of Conic Linear Problems*, pages 135–165. Springer US, Boston, MA, 2001.

[21] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[22] Bo Zeng and Long Zhao. Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters*, 41(5):457–461, 2013.

[23] Fengqiao Luo and Sanjay Mehrotra. A decomposition method for distributionally-robust two-stage stochastic mixed-integer conic programs. *Mathematical Programming*, 196(1):673–717, 2022.

[24] Mohamed El Tonbari, George Nemhauser, and Alejandro Toriello. Distributionally robust disaster relief planning under the wasserstein set. *Computers & Operations Research*, page 106689, 2024.

[25] Daniel Zhuoyu Long, Jin Qi, and Aiqi Zhang. Supermodularity in two-stage distributionally robust optimization. *Management Science*, 70(3):1394–1409, 2024.

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

[27] Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618, 2019.

[28] Angelos Georghiou, Angelos Tsoukalas, and Wolfram Wiesemann. On the optimality of affine decision rules in robust and distributionally robust optimization. *Available at Optimization Online*, 2021.

[29] Luhao Zhang, Jincheng Yang, and Rui Gao. A simple and general duality proof for wasserstein distributionally robust optimization. *arXiv preprint arXiv:2205.00362*, 2022.

[30] L.V. Kantorovic and V.A. Zalgaller. *Rational cutting of industrial materials.*

[31] Lester Randolph Ford Jr and Delbert R Fulkerson. A suggested computation for maximal multi-commodity network flows. *Management Science*, 5(1):97–101, 1958.

[32] George B Dantzig and Philip Wolfe. Decomposition principle for linear programs. *Operations research*, 8(1):101–111, 1960.

[33] Paul C Gilmore and Ralph E Gomory. A linear programming approach to the cutting stock problem. *Operations research*, 9(6):849–859, 1961.

[34] Bo Zeng and Wei Wang. Two-stage robust optimization with decision dependent uncertainty. *arXiv preprint arXiv:2203.16484*, 2022.

[35] Cynthia Barnhart, Ellis L Johnson, George L Nemhauser, Martin WP Savelsbergh, and Pamela H Vance. Branch-and-price: Column generation for solving huge integer programs. *Operations research*, 46(3):316–329, 1998.

[36] Bo Zeng and Yu An. Solving bilevel mixed integer program by reformulations and decomposition. *Optimization online*, pages 1–34, 2014.

[37] Lawrence V Snyder and Mark S Daskin. Reliability models for facility location: the expected failure cost case. *Transportation science*, 39(3):400–416, 2005.

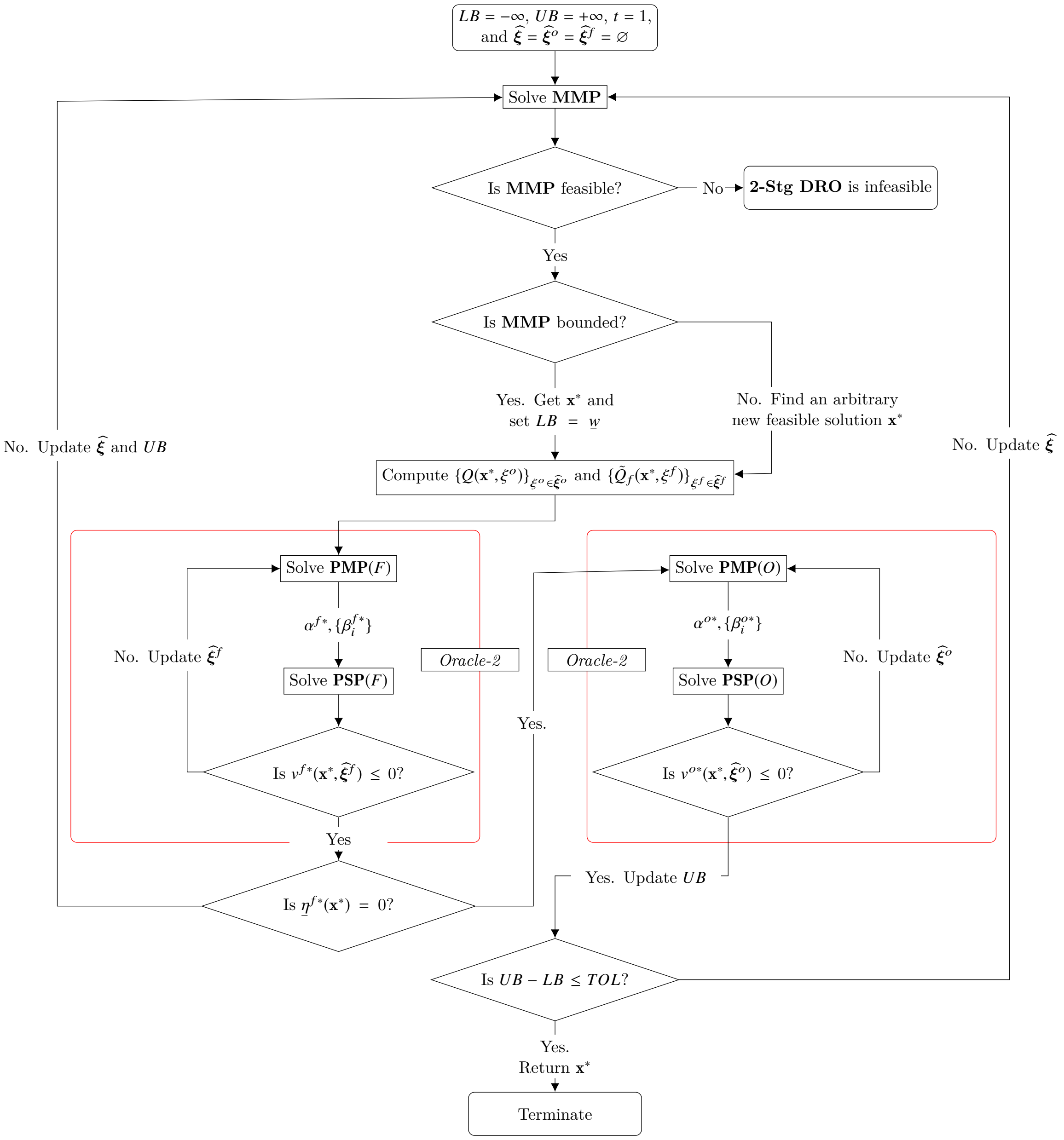

$LB = -\infty$, $UB = +\infty$, $t = 1$, and $\widehat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^o = \widehat{\boldsymbol{\xi}}^f = \varnothing$
Solve **MMP**
Is **MMP** feasible?
No
2-Stg DRO is infeasible
Yes
Is **MMP** bounded?
Yes. Get $\mathbf{x}^*$ and set $LB = \underline{w}$
No. Find an arbitrary new feasible solution $\mathbf{x}^*$
No. Update $\widehat{\boldsymbol{\xi}}$ and UB
No. Update $\widehat{\boldsymbol{\xi}}$
Compute $\{Q(\mathbf{x}^*, \xi^o)\}_{\xi^o \in \widehat{\boldsymbol{\xi}}^o}$ and $\{\tilde{Q}_f(\mathbf{x}^*, \xi^f)\}_{\xi^f \in \widehat{\boldsymbol{\xi}}^f}$
Solve **PMP**(F)
$\alpha^{f*}, \{\beta_i^{f*}\}$
No. Update $\widehat{\boldsymbol{\xi}}^f$
Oracle-2
Solve **PSP**(F)
Is $v^{f*}(\mathbf{x}^*, \widehat{\boldsymbol{\xi}}^f) \leq 0$?
Yes
Is $\underline{\eta}^{f*}(\mathbf{x}^*) = 0$?
Yes.
Oracle-2
Solve **PMP**(O)
$\alpha^{o*}, \{\beta_i^{o*}\}$
No. Update $\widehat{\boldsymbol{\xi}}^o$
Solve **PSP**(O)
Is $v^{o*}(\mathbf{x}^*, \widehat{\boldsymbol{\xi}}^o) \leq 0$?
Yes. Update UB
Is $UB - LB \leq TOL$?
Yes. Return $\mathbf{x}^*$
Terminate

Figure 4: Flowchart for C&CG-DRO algorithm with *Oracle-2*