IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Distributionally Robust Optimization using Cost-Aware Ambiguity Sets

Mathijs Schuurmans and Panagiotis Patrinos

***Abstract*— We present a novel framework for distributionally robust optimization (DRO), called cost-aware DRO (CADRO). The key idea of CADRO is to exploit the cost structure in the design of the ambiguity set to reduce conservatism. Particularly, the set specifically constrains the worst-case distribution along the direction in which the expected cost of an approximate solution increases most rapidly. We prove that CADRO provides both a high-confidence upper bound and a consistent estimator of the out-of-sample expected cost, and show empirically that it produces solutions that are substantially less conservative than existing DRO methods, while providing the same guarantees.**

I. INTRODUCTION

We consider the stochastic programming problem

$$\underset{x\in X}{\mathbf{minimize}}\ \mathbb{E}[\ell(x,\xi)] \tag{1}$$

with $X \subseteq \mathbb{R}^n$ a nonempty, closed set of feasible decision variables, $\xi \in \Xi$ a random variable following probability measure $\mathbb{P}$, and $\ell : \mathbb{R}^n \times \Xi \to \mathbb{R}$ a known cost function. This problem is foundational in many fields, including operations research [1], machine learning [2], and control (e.g., stochastic model predictive control) [3].

Provided that the underlying probability measure $\mathbb{P}$ is known exactly, this problem can effectively be solved using traditional stochastic optimization methods [1], [4]. In reality, however, only a data-driven estimate $\hat{\mathbb{P}}$ of $\mathbb{P}$ is typically available, which may be subject to misestimations—known as *ambiguity*. Perhaps the most obvious method for handling this issue is to disregard this ambiguity and instead apply a sample average approximation (SAA) (also known as empirical risk minimization (ERM) in the machine learning literature), where (1) is solved using $\hat{\mathbb{P}}$ as a *plug-in* replacement for $\mathbb{P}$. However, this is known to produce overly optimistic estimates of the optimal cost [4, Prop. 8.1], potentially resulting in unexpectedly high realizations of the cost when deploying the obtained optimizers on new, unseen samples. This downward bias of SAA is closely related to the issue of overfitting, and commonly referred to as the *optimizer's curse* [5], [6].

Several methods have been devised over the years to combat this undesirable behavior. Classical techniques such as regularization and cross-validation are commonly used in machine learning [2], although typically, they are used as heuristics, providing few rigorous guarantees, in particular for small sample sizes. Alternatively, the suboptimality gap of the SAA solution may be statistically estimated by reserving a fraction of the dataset for independent replications [7]. However, these results are typically based on asymptotic arguments, and are therefore not valid in the low-sample regime. Furthermore, although this type of approach may be used to *validate* the SAA solution, it does not attempt to *improve* it, by taking into account possible estimation errors. More recently, distributionally robust optimization (DRO) has garnered considerable attention, as it provides a principled way of obtaining a high-confidence upper bound on the true out-of-sample cost [6], [8], [9]. In particular, its capabilities to provide rigorous performance and safety guarantees has made it an attractive technique for data-driven and learning-based control [10]–[12]. DRO refers to a broad class of methods in which a variant of (1) is solved where $\mathbb{P}$ is replaced with a worst-case distribution within a statistically estimated set of distributions, called an *ambiguity set*.

As the theory essentially requires only that the ambiguity set contains the true distribution with a prescribed level of confidence, a substantial amount of freedom is left in the design of the geometry of these sets. As a result, many different classes of ambiguity sets have been proposed in the literature, e.g., Wasserstein ambiguity sets [9], divergence-based ambiguity sets [6], [12], [13] and moment-based ambiguity sets [8], [14]; See [15], [16] for recent surveys.

Despite the large variety of existing classes of ambiguity sets, a common characteristic is that their design is considered separately from the optimization problem in question. Although this simplifies the analysis in some cases, it may also induce a significant level of conservatism; In reality, we are only interested in excluding distributions from the ambiguity set which actively contribute to increasing the worst-case cost. Requiring that the true distribution deviates little from the data-driven estimate in *all directions* may therefore be unnecessarily restrictive. This intuition motivates the introduction of a new DRO methodology, which is aimed at designing the geometry of the ambiguity sets with the original problem (1) in mind. The main idea is that by only excluding those distributions that maximally affect the worst-case cost, higher levels of confidence can be attained without introducing additional conservatism to the cost estimate.

Contributions: (i) We propose a novel class of ambiguity sets for DRO, taking into account the structure of the underlying optimization problem; (ii) We prove that the DRO cost is both a high-confidence upper bound and a consistent estimate of the optimal cost of the original stochastic program (1); (iii) We demonstrate empirically that the provided ambiguity set outperforms existing alternatives.

Notation: We denote $[n] = \{1, \dots, n\}$, for $n \in \mathbb{N}$. $|S|$ denotes the cardinality of a (finite) set S. $\mathbf{e}_i \in \mathbb{R}^n$ is the ith standard basis vector in $\mathbb{R}^n$. Its dimension n will be clear from context. We write 'a.s.' to signify that a random event occurs *almost surely*, i.e., with probability 1. δ_X is the indicator of a set X: $\delta_X(x) = 0$ if $x \in X$, $+\infty$ otherwise.

II. Problem Statement

We will assume that the random variable ξ is finitely supported, so that without loss of generality, we may write $\Xi = \{1, \dots, d\}$. This allows us to define the probability mass vector $p = (\mathbb{P}[\xi = i])_{i=1}^d$, and enumerate the *cost realizations* $\ell_i = \ell(\cdot, i)$, $i \in [d]$. Furthermore, it will be convenient to introduce the mapping $L : \mathbb{R}^n \to \mathbb{R}^d$ as $L(x) = (\ell_1(x), \dots, \ell_d(x))$. We will pose the following (mostly standard) regularity assumption on the cost function.

Assumption II.1 (Problem regularity)**.** *For all* $i \in [d]$

(i) ℓ_i *is continuous on* X*;*
(ii) $\bar{\ell}_i := \ell_i + \delta_X$ *is level-bounded;*

Since any continuous function is lower semicontinuous (lsc), Assumption II.1 combined with the closedness of X implies *inf-compactness*, which ensures attainment of the minimum [17, Thm. 1.9]. Continuity of ℓ_i is used mainly to establish continuity of the solution mapping $V^\star$—defined below, see (2). However, a similar result can be obtained by replacing condition (i) by *lower semicontinuity* and *uniform level-boundedness* on X. However, for ease of exposition, we will not cover this modification explicitly.

Let $p^\star \in \Delta_d := \{p \in \mathbb{R}^d_+ \mid \sum_{i=1}^d p_i = 1\}$ denote the true-but-unknown probability mass vector, and define $V : \mathbb{R}^n \times \Delta_d \to \mathbb{R} : (x, p) \mapsto \langle p, L(x) \rangle$, to obtain the parametric optimization problem with optimal cost and solution set

$$V^\star(p) = \min_{x \in X} V(x, p) \text{ and } X^\star(p) = \operatorname*{argmin}_{x \in X} V(x, p). \quad (2)$$

The solution of (1) is retrieved by solving (2) with $p = p^\star$.

Assume we have access to a dataset $\hat{\Xi} := \{\xi_1, \dots, \xi_m\} \in \Xi^m$ collected i.i.d. from $p^\star$. In order to avoid the aforementioned downward bias of SAA, our goal is to obtain a data-driven decision $\hat{x}_m$ along with an estimate $\hat{V}_m$ such that

$$\mathbb{P}[V(\hat{x}_m, p^\star) \leq \hat{V}_m] \geq 1 - \beta, \quad (3)$$

where $\beta \in (0, 1)$ is a user-specified confidence level.

We address this problem by means of distributionally robust optimization, where instead of (2), one solves the surrogate problem

$$\hat{V}_m = \min_{x \in X} \max_{p \in \mathcal{A}_m} V(x, p). \quad \text{(DRO)}$$

Here, $\mathcal{A}_m \subseteq \Delta_d$ is a (typically data-dependent, and thus, random) set of probability distributions that is designed to contain the true distribution $p^\star$ with probability $1-\beta$, ensuring that (3) holds. Trivially, (3) is satisfied with $\beta = 0$ by taking $\mathcal{A}_m \equiv \Delta_d$. This recovers a robust optimization method, i.e., $\min_{x \in X} \max_{i \in [d]} \ell_i(x)$. Although it satisfies (3), this robust approach tends to be overly conservative as it neglects all available statistical data. The aim of distributionally robust optimization is to additionally ensure that $\hat{V}_m$ is a consistent estimator, i.e.,

$$\lim_{m \to \infty} \hat{V}_m = V^\star(p^\star), \quad \text{a.s.} \quad (4)$$

We will say that a class of ambiguity sets is *admissible* if the solution $\hat{V}_m$ of the resulting DRO problem (DRO) satisfies (3) and (4). Our objective is to develop a methodology for constructing admissible ambiguity sets that take into account the structure of (DRO) and in doing so, provide tighter estimates of the cost, while maintaining (3) with a given confidence level β.

III. Cost-Aware DRO

In this section, we describe the proposed DRO framework, which we will refer to as *cost-aware DRO* (CADRO). The overall method is summarized in Alg. 1.

A. Motivation

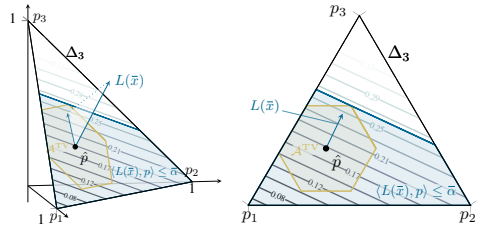

Fig. 1. Conceptual motivation for the structure of the ambiguity set (6). The cost contour lines $\{p \in \Delta_3 \mid \langle L(\bar{x}), p \rangle = \alpha\}$ corresponding to some $\bar{x} \in X$ are shown for increasing values of α (dark to light), together with the sets $\mathcal{A}^{\mathrm{TV}} := \Delta_3 \cap \mathbb{B}_1(\hat{p}, \varrho)$ and $\mathcal{A} := \{p \in \Delta_3 \mid \langle L(\bar{x}), p \rangle \leq \bar{\alpha}\}$. Here, $\varrho > 0$ is determined to satisfy (5) and $\bar{\alpha} = \max_{p \in \mathcal{A}^{\mathrm{TV}}} \langle L(\bar{x}), p \rangle$. Since $\mathcal{A}^{\mathrm{TV}} \subset \mathcal{A}$, $\mathcal{A}$ satisfies (5) with a higher confidence level $1 - \beta$, but nevertheless, we have $\max_{p \in \mathcal{A}} V(\bar{x}, p) = \max_{p \in \mathcal{A}^{\mathrm{TV}}} V(\bar{x}, p)$.

We start by providing some intuitive motivation. Consider the problem (DRO). In order to provide a guarantee of the form (3), it obviously suffices to design $\mathcal{A}_m$ such that

$$\mathbb{P}[p^\star \in \mathcal{A}_m] \geq 1 - \beta. \quad (5)$$

However, this condition alone still leaves a considerable amount of freedom to the designer. A common approach is to select $\mathcal{A}_m$ to be a ball (expressed in some statistical metric/divergence) around an empirical estimate $\hat{p}$ of the distribution. Depending on the choice of metric/divergence (e.g., total variation [18], Kullback-Leibler [6], Wasserstein [9], ...), several possible variants may be obtained. Using concentration inequalities, one can then select the appropriate radius of this ball, such that (5) is satisfied. A drawback of this approach, however, is that the construction of $\mathcal{A}_m$ is decoupled from the original problem (1). Indeed, given that $\mathcal{A}_m$ takes the form of a ball, (5) essentially requires the deviation of $\hat{p}$ from $p^\star$ to be small *along every direction*. If one could instead enlarge the ambiguity set without increasing the worst-case cost, then (5) could be guaranteed for smaller values of β without introducing additional conservatism. This idea is illustrated in Fig. 1.

Conversely, for a fixed confidence level β, one could thus construct a smaller upper bound $\hat{V}_m$, by restricting the choice of p only in a judiciously selected direction. Particularly, we may set $\mathcal{A}_m = \{p \in \Delta_d \mid \langle L(\bar{x}), p \rangle \leq \alpha_m\}$ for some candidate solution $\bar{x} \in X$, where α_m is the smallest (potentially data-dependent) quantity satisfying (5). This directly yields an upper bound on the estimate $\hat{V}_m$. Namely,

for $x^\star \in X^\star(p^\star)$, we have with probability $1-\beta$,

$$V(x^\star,p^\star) \overset{(a)}{\leq} V(\hat{x}_m,p^\star) \leq \max_{p\in\mathcal{A}_m} V(\hat{x}_m,p) = \hat{V}_m$$
$$= \min_{x\in X}\max_{p\in\mathcal{A}_m} V(x,p) \overset{(b)}{\leq} \max_{p\in\mathcal{A}_m} V(\bar{x},p) = \alpha_m.$$

Here, inequalities (a) and (b) become equalities when $\hat{x}_m = x^\star = \bar{x}$. Thus, a reasonable aim would be to select $\bar{x}$ to be a good approximation of $x^\star$. We will return to the matter of selecting $\bar{x}$ in §III-C. First, however, we will assume $\bar{x}$ to be given and focus on establishing the coverage condition (5).

B. Ambiguity set parameterization and coverage

Motivated by the previous discussion, we propose a family of ambiguity sets parameterized as follows. Let $v \in \mathbb{R}^d$ be a fixed vector (we will discuss the choice of v in §III-C). Given a sample $\hat{\Xi} = \{\xi_1,\dots,\xi_m\}$ of size $|\hat{\Xi}| = m$ drawn i.i.d. from $p^\star$, we consider ambiguity sets of the form

$$\mathcal{A}_{\hat{\Xi}}(v) := \{p \in \Delta_d \mid \langle p, v\rangle \leq \alpha_{\hat{\Xi}}(v)\}, \tag{6}$$

where $\alpha : \Xi^m \times \mathbb{R}^d \ni (\hat{\Xi}, v) \mapsto \alpha_{\hat{\Xi}}(v) \in \mathbb{R}$ is a data-driven estimator for $\langle p^\star, v\rangle$, selected to satisfy the following assumption, which implies that (5) holds for $\mathcal{A}_m = \mathcal{A}_{\hat{\Xi}}(v)$.

Assumption III.1. $\mathbb{P}[\langle p^\star, v\rangle \leq \alpha_{\hat{\Xi}}(v)] \geq 1-\beta,\ \forall v \in \mathbb{R}^d$.

Note that the task of selecting α to satisfy Assumption III.1 is equivalent to finding a high-confidence upper bound on the mean of the scalar random variable $\langle v, \mathbf{e}_\xi\rangle$, $\xi \sim p^\star$. It is straightforward to derive such bounds by bounding the deviation of a random variable from its empirical mean using classical concentration inequalities like Hoeffding's inequality [19, Prop. III.2]. Although attractive for its simplicity, this type of bounds has the drawback that it applies a constant offset (depending only on the sample *size*, not the data) to the empirical mean, which may be conservative, especially for small samples. Considerably sharper bounds can be obtained through a more direct approach. In particular, we will focus our attention on the following result due to Anderson [20], which is a special case of the framework presented in [21].

Proposition III.2 (Ordered mean bound [21])**.** *Let $\eta_k := \langle v, \mathbf{e}_{\xi_k}\rangle$, $k \in [m]$, so that $\mathbb{E}[\eta_k] = \langle v, p^\star\rangle$. Let $\eta_{(1)} \leq \eta_{(2)} \leq \cdots \leq \eta_{(m)} \leq \bar{\eta}$ denote the sorted sequence, with ties broken arbitrarily, where $\bar{\eta} := \max_{i\in[d]} v_i$. Then, there exists a $\gamma \in (0,1)$ such that Assumption III.1 holds for*

$$\alpha_{\hat{\Xi}}(v) = \left(\tfrac{\kappa}{m} - \gamma\right)\eta_{(\kappa)} + \textstyle\sum_{i=\kappa+1}^{m} \frac{\eta_{(i)}}{m} + \gamma\bar{\eta},\ \kappa = \lceil m\gamma\rceil. \tag{7}$$

For finite m, the smallest value of γ ensuring that Proposition III.2 holds, can be computed efficiently by solving a scalar root-finding problem [21, Rem. IV 3]. Furthermore, it can be shown that the result holds for [22, Thm. 11.6.2]

$$\gamma = \sqrt{\tfrac{\log(1/\beta)}{2m}},\ \text{for sufficiently large } m. \tag{8}$$

This asymptotic expression will be useful when establishing theoretical guarantees in Section IV.

C. Selection of v

The proposed ambiguity set (6) depends on a vector v. As discussed in §III-A, we would ideally take $v = L(x^\star)$ with $x^\star \in X^\star(p^\star)$. However, since this ideal is obviously out of reach, we instead look for suitable approximations. In particular, we propose to use the available dataset $\hat{\Xi}$ in part to select v to approximate $L(x^\star)$, and in part to calibrate the mean bound α.

To this end, we will partition the available dataset $\hat{\Xi}$ into a *training* set and a *calibration* set. Let $\tau : \mathbb{N} \to \mathbb{N}$ be a user-specified function determining the size of the *training* set, which satisfies

$$\tau(m) \leq cm \ \text{ for some } c \in (0,1);\ \text{and} \tag{9a}$$
$$\tau(m) \to \infty \ \text{ as } m \to \infty. \tag{9b}$$

Correspondingly, let $\{\hat{\Xi}_{\mathrm{T}}, \hat{\Xi}_{\mathrm{C}}\}$ be a partition of $\hat{\Xi}$, i.e., $\hat{\Xi}_{\mathrm{T}} \cap \hat{\Xi}_{\mathrm{C}} = \emptyset$ and $\hat{\Xi}_{\mathrm{T}} \cup \hat{\Xi}_{\mathrm{C}} = \hat{\Xi}$. Given that $|\hat{\Xi}| = m$, we ensure that $|\hat{\Xi}_{\mathrm{T}}| = \tau(m)$ and thus $|\hat{\Xi}_{\mathrm{C}}| = m' := m - \tau(m)$. Note that by construction, $m' \geq (1-c)m$, with $c \in (0,1)$, and thus, both $|\hat{\Xi}_{\mathrm{T}}| \to \infty$ and $|\hat{\Xi}_{\mathrm{C}}| \to \infty$ as $m \to \infty$. Due to the statistical independence of the elements in $\hat{\Xi}$, it is inconsequential how exactly the individual data points are divided into $\hat{\Xi}_{\mathrm{T}}$ and $\hat{\Xi}_{\mathrm{C}}$. Therefore, without loss of generality, we may take $\hat{\Xi}_{\mathrm{T}} = \{\xi_1,\dots,\xi_{\tau(m)}\}$ and $\hat{\Xi}_{\mathrm{C}} = \{\xi_{\tau(m)+1},\dots,\xi_m\}$.

With an independent dataset $\hat{\Xi}_{\mathrm{T}}$ at our disposal, we may use it to design a mapping $v_{\tau(m)} : \Xi^{\tau(m)} \to \mathbb{R}^d$, whose output will be a data-driven estimate of $L(x^\star)$. For ease of notation, we will omit the explicit dependence on the data, i.e., we write $v_{\tau(m)}$ instead of $v_{\tau(m)}(\hat{\Xi}_{\mathrm{T}})$. We propose the following construction. Let $\hat{p}_{\tau(m)} = \frac{1}{\tau(m)}\sum_{k=1}^{\tau(m)} \mathbf{e}_{\xi_k}$ denote the empirical distribution of $\hat{\Xi}_{\mathrm{T}}$ and set

$$\begin{aligned} &v_{\tau(m)} = L(\bar{x}_{\tau(m)}),\ \text{with} \\ &\bar{x}_{\tau(m)} \in \operatorname*{\mathbf{argmin}}_{x\in X} V(x, \hat{p}_{\tau(m)}). \end{aligned} \tag{10}$$

Remark III.3. We underline that although (10) is a natural choice, several alternatives for the *training* vector could in principle be considered. To guide this choice, Lemma IV.2 provides sufficient conditions on the combination of α and $v_{\tau(m)}$ to ensure consistency of the method.

Given $v_{\tau(m)}$ as in (10), we will from hereon use the following shorthand notation whenever convenient:

$$\mathcal{A}_m := \mathcal{A}_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)}), \quad \alpha_m := \alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)}), \tag{11}$$

with $\mathcal{A}_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)})$ as in (6). We correspondingly obtain the cost estimate $\hat{V}_m$ according to (DRO).

D. Selection of τ

Given the conditions in (9), there is still some flexibility in the choice of $\tau(m)$, which defines a trade-off between the quality of $v_{\tau(m)}$ as an approximator of $L(x^\star)$ and the size of the ambiguity set $\mathcal{A}_m$.

An obvious choice is to reserve a fixed fraction of the available data for the *training* set, i.e., set $\tau(m)/m$ equal to some constant. However, for low sample counts m, the mean bound α_m will typically be large and thus $\mathcal{A}_m$ will not be

substantially smaller than the unit simplex Δ_d, regardless of $v_{\tau(m)}$. As a result, the obtained solution will also be rather insensitive to $v_{\tau(m)}$. In this regime, it is therefore preferable to reduce the conservatism of α_m quickly by using small values of $\tau(m)/m$ (i.e., large values of $m' = m - \tau(m)$).

Conversely, for large sample sizes, α_m is typically a good approximation of $\langle p^\star, v_{\tau(m)} \rangle$ and the solution to (DRO) will be more strongly biased to align with $v_{\tau(m)}$. Thus, the marginal benefit of improving the quality of $v_{\tau(m)}$ takes priority over reducing α_m, and large fractions $\tau(m)/m$ become preferable. Based on this reasoning, we propose the heuristic

$$\tau(m) = \lfloor \mu\nu \tfrac{m(m+1)}{\mu m + \nu} \rfloor, \quad \mu, \nu \in (0, 1). \tag{12}$$

Note that μ and ν are the limits of $\tau(m)/m$ as $m \to 0$ and $m \to \infty$, respectively. Eq. (12) then interpolates between these extremes, depending on the total amount of data available. We have found $\mu = 0.01, \nu = 0.8$ to be suitable choices for several test problems.

E. Tractable reformulation

The proposed ambiguity set takes the form of a polytope, and thus, standard reformulations based on conic ambiguity sets apply directly [23]. Nevertheless, as we will now show, a tractable reformulation of (DRO) specialized to the ambiguity set (6) may be obtained, which requires fewer auxiliary variables and constraints.

Proposition III.4 (Tractable reformulation of (DRO))**.** *Fix parameters $\hat{p} \in \Delta$, $v \in \mathbb{R}^d$, and $\alpha \in \mathbb{R}$ and let $\mathcal{A} = \{p \in \Delta_d \mid \langle p, v \rangle \leq \alpha\}$ be an ambiguity set of the form* (6)*. Denoting $V_{\mathcal{A}} := \mathbf{min}_{x \in X} \mathbf{max}_{p \in \mathcal{A}} V(x, p)$, we have*

$$V_{\mathcal{A}} = \min_{x \in X, \lambda \geq 0} \lambda\alpha + \max_{i \in [d]} \{\ell_i(x) - \lambda v_i\}. \tag{13}$$

Proof. Let $g(z) := \mathbf{max}_{p \in \Delta_d} \{\langle p, z \rangle \mid \langle p, v \rangle \leq \alpha\}$, where $z \in \mathbb{R}^d$ and α are constants with respect to p. By strong duality of linear programming [24],

$$\begin{aligned} g(z) &= \min_{\lambda \geq 0} \max_{p \in \Delta_d} \langle p, z \rangle - \lambda(\langle p, v \rangle - \alpha) \\ &= \min_{\lambda \geq 0} \lambda\alpha + \max_{p \in \Delta_d} \langle p, z - \lambda v \rangle \end{aligned}$$

Noting that $\mathbf{max}_{p \in \Delta_d} y = \mathbf{max}_{i \in [d]} y_i$, $\forall y \in \mathbb{R}^d$ and that $V_{\mathcal{A}} = \mathbf{min}_{x \in X} g(L(x))$, we obtain (13). □

If the functions $\{\ell_i\}_{i \in [d]}$ are convex, then (13) is a convex optimization problem, which can be solved efficiently using off-the-shelf solvers. In particular, if they are convex, piecewise affine functions, then it reduces to a linear program (LP). For instance, introducing a scalar epigraph variable, one may further rewrite (13) as

$$\min_{x \in X, \lambda \geq 0, z \in \mathbb{R}} \{\lambda\alpha + z \mid L(x) - \lambda v \leq z\mathbf{1}\}, \tag{14}$$

which avoids the non-smoothness of the pointwise maximum in (13) at the cost of a scalar auxiliary variable. Even for general (possibly nonconvex) choices of ℓ_i, (13) is a standard nonlinear program, which can be handled by existing solvers.

We conclude the section by summarizing the described steps in Alg. 1.

Algorithm 1 CADRO

Require: i.i.d. dataset $\hat{\Xi} = \{\xi_1, \ldots, \xi_m\}$; $\tau(m)$ (cf. (9)); Confidence parameter $\beta \in (0, 1)$.
Ensure: $(\hat{V}_m, \hat{x}_m)$ satisfy (3)–(4) ▷ Cf. §IV
$\hat{\Xi}_{\mathrm{T}} \leftarrow \{\xi_1, \ldots, \xi_{\tau(m)}\}$, $\hat{\Xi}_{\mathrm{C}} \leftarrow \{\xi_{\tau(m)+1}, \ldots, \xi_m\}$
$v_{\tau(m)} \leftarrow$ evaluate (10)
$(\hat{V}_m, \hat{x}_m) \leftarrow$ solve (DRO) with $\mathcal{A}_m = \mathcal{A}_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)})$ ▷ Use (13)

IV. THEORETICAL PROPERTIES

We will now show that the proposed scheme possesses the required theoretical properties, namely to provide (i) an upper bound to the out-of-sample cost, with high probability (cf. (3)); and (ii) a consistent estimate of the true optimal cost (cf. (4)). The first guarantee follows almost directly by construction, and its proof is therefore omitted here. See [19] for more details.

Theorem IV.1 (Out-of-sample guarantee)**.** *Fix $m > 0$, and let $\hat{V}_m$, $\hat{x}_m$ be generated by Alg. 1. Then,*

$$\mathbb{P}[V(\hat{x}_m, p^\star) \leq \hat{V}_m] \geq 1 - \beta. \tag{15}$$

We now turn our attention to the matter of consistency. That is, we will show that under suitable conditions on the mean bound α and the *training* vector v in (6), $\hat{V}_m$ converges almost surely to the true optimal value, as the sample size m grows to infinity. We will then conclude the section by demonstrating that for the choices proposed in §III-B and III-C, the aforementioned conditions hold.

Lemma IV.2 (consistency conditions)**.** *Let $\hat{\Xi}_{\mathrm{T}}$, $\hat{\Xi}_{\mathrm{C}}$ be two independent samples from $p^\star$, with sizes $|\hat{\Xi}_{\mathrm{T}}| = \tau(m)$ and $|\hat{\Xi}_{\mathrm{C}}| = m' := m - \tau(m)$. Let $\hat{p}_{m'} := \frac{1}{m'} \sum_{\xi \in \hat{\Xi}_{\mathrm{C}}} \mathbf{e}_\xi$ denote the empirical distribution of the calibration set $\hat{\Xi}_{\mathrm{C}}$. If $v_{\tau(m)} = L(\bar{x}_{\tau(m)})$, with $\bar{x}_{\tau(m)}$, $\alpha_m = \alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)})$ chosen to ensure*
(i) $\langle \hat{p}_{m'}, v_{\tau(m)} \rangle \leq \alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)})$, a.s.;
(ii) $\limsup_{m \to \infty} \alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)}) \leq V^\star(p^\star)$, a.s.
Then $\hat{V}_m \to V^\star(p^\star)$, a.s., where $\hat{V}_m$ is given by (DRO)*.*

Proof. Let $\overline{V}_m(x) := \mathbf{max}_{p \in \mathcal{A}_m} \langle p, L(x) \rangle$. It is clear from condition (i) and (6) that $\hat{p}_{m'} \in \mathcal{A}_m$. Let us furthermore define $\varepsilon_m(x) = L(x) - L(\bar{x}_{\tau(m)})$. Then, by [19, Lem. A.2], we have for all $x \in X$, $\langle \hat{p}_{m'}, L(x) \rangle \leq \overline{V}_m(x) \leq \alpha_m + \|\varepsilon_m(x)\|_\infty$. Minimizing with respect to x yields that for all m,

$$\hat{V}^{\mathrm{SAA}}_{m'} \leq \hat{V}_m \leq \alpha_m, \tag{16}$$

where $\hat{V}^{\mathrm{SAA}}_{m'} := V^\star(\hat{p}_{m'})$ (cf. (2)). By the law of large numbers, $\hat{p}_{m'} \to p^\star$, a.s. Furthermore, under Assumption II.1, [19, Lem. A.4] states that the optimal value mapping $V^\star(p)$ is continuous, which implies that also $\hat{V}^{\mathrm{SAA}}_{m'} \to V^\star(p^\star)$, a.s. The claim then follows directly from condition (ii). □

Informally, Lemma IV.2 requires that the mean bound is bounded from below by the empirical mean, and from above by a consistent estimator of the optimal cost. The latter excludes choices such as the robust minimizer $\bar{x}_{\tau(m)} \in \mathbf{argmin}\, \mathbf{max}_{i \in [d]} \ell_i(x)$ in the construction of $v_{\tau(m)}$. However, besides (10), one could consider alternatives, such as

a separate DRO scheme to select $v_{\tau(m)}$. A more extensive study of such alternatives, however, is left for future work. We now conclude the section by showing that (10) and the mean bound given by Proposition III.2 satisfy the requirements of Lemma IV.2.

Theorem IV.3 (Consistency – Ordered mean bound)**.** *Let $\hat{V}_m$ be generated by Alg. 1, for $m > 0$. If $\alpha_m = \alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)})$ is selected according to Proposition III.2, with $v_{\tau(m)}$ as in* (10)*, then, $\hat{V}_m \to V^\star(p^\star)$, a.s.*

Proof. It suffices to show that conditions (i) and (ii) of Lemma IV.2 are satisfied by $\alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)})$.

Condition (i): Consider $\alpha_{\hat{\Xi}_{\mathrm{C}}}(v)$ as in (7) for an arbitrary $v \in \mathbb{R}^d$, and let $(\eta_{(i)})_{i\in[m']}$ denote $(\langle v, \mathbf{e}_\xi\rangle)_{\xi\in\hat{\Xi}_{\mathrm{C}}}$, sorted in increasing order, then, we may write

$$\langle \hat{p}_{m'}, v\rangle = \tfrac{1}{m'}\textstyle\sum_{i=1}^{m'} \eta_{(i)}, \tag{17}$$

and thus,

$$\begin{aligned}\alpha_{\hat{\Xi}_{\mathrm{C}}}(v) - \langle \hat{p}_{m'}, v\rangle &= \left(\tfrac{\kappa}{m'} - \gamma\right)\eta_{(\kappa)} - \textstyle\sum_{i=1}^{\kappa} \tfrac{\eta_{(i)}}{m'} + \gamma\bar{\eta},\\ &\overset{\text{(a)}}{\geq} \left(\tfrac{\kappa}{m'} - \gamma\right)\eta_{(\kappa)} - \tfrac{\kappa}{m'}\eta_{(k)} + \gamma\bar{\eta},\\ &= \gamma(\bar{\eta} - \eta_{(k)}) \overset{(\gamma\geq 0)}{\geq} 0, \quad \forall v \in \mathbb{R}^d,\end{aligned}$$

where (a) follows from the fact that $\eta_{(i)}$ are sorted.

Condition (ii): There exists a constant $\bar{v} \geq \|v_{\tau(m)}\|_\infty$, $\forall m > 0$, a.s. [19, Lem. A.4]. Therefore, using (7) and (17),

$$\begin{aligned}\alpha_m - \langle \hat{p}_{m'}, v_{\tau(m)}\rangle &\leq \left(\tfrac{\kappa}{m'} - \gamma\right)\bar{v} + \tfrac{\kappa}{m'}\bar{v} + \gamma\bar{v}\\ &= 2\bar{v}\left(\tfrac{\kappa}{m'}\right) \overset{\text{(b)}}{\leq} 2\bar{v}(\gamma + \tfrac{1}{m'}),\end{aligned} \tag{18}$$

for all $m' > 0$, where (b) follows from $\kappa = \lceil m'\gamma\rceil \leq m'\gamma + 1$. By construction (see (9) and below), we have that both $\tau(m) \to \infty$ and $m' \to \infty$. Thus, using (8), $\gamma + \frac{1}{m'} = \sqrt{\frac{\log(1/\beta)}{2m'}} + \frac{1}{m'} \to 0$. Combined with (18), this yields that

$$\limsup_{m\to\infty} \alpha_{\hat{\Xi}_{\mathrm{C}}}(v_{\tau(m)}) - \langle v_{\tau(m)}, \hat{p}_{\hat{\Xi}_{\mathrm{C}}}\rangle \leq 0. \tag{19}$$

Finally, by the law of large numbers, $\hat{p}_{m'} \to p^\star$ and $\hat{p}_{\tau(m)} \to p^\star$, a.s. Thus (under Assumption II.1), [19, Cor. A.6] ensures that $\lim_{m\to\infty}\langle \hat{p}_{m'}, L(\bar{x}_{\tau(m)})\rangle = V^\star(p^\star)$, which, combined with (19) yields the required result. □

V. Illustrative Example

As an illustrative example, we consider the following *facility location problem*, adapted from [25, Sec. 8.7.3]. Consider a bicycle sharing service setting out to determine locations $x^{(i)} \in X_i \subseteq \mathbb{R}^2$, $i \in [n_x]$, at which to build stalls where bikes can be taken out or returned. We will assume that X_i are given (polyhedral) sets, representing areas within the city suitable for constructing a new bike stall. Let $z^{(k)} \in \mathbb{R}^2$, $k \in [d]$, be given points of interest (public buildings, tourist attractions, parks, etc.). Suppose that a person located in the vicinity of some point $z^{(k)}$ decides to rent a bike. Depending on the availability at the locations $x^{(i)}$, this person may be required to traverse a distance $\ell_k(x) = \mathbf{max}_{i\in[n_x]}\|x^{(i)} - z^{(k)}\|_2$, where $x = (x^{(i)})_{i\in[n_x]}$. With this choice of cost, (13) can be cast as a second order cone program. Thus, if the demand is distributed over $(z^{(k)})_{k\in[d]}$ according to the probability mass vector $p^\star \in \Delta_d$, then the average cost to be minimized over $X = X_1 \times \cdots \times X_d$ is given by $V(x, p^\star)$ as in (2). We will solve a randomly generated instance of the problem, illustrated in Fig. 2.

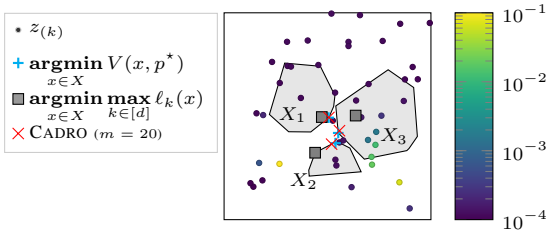

Fig. 2. Illustration of the facility location problem. The colors of the points $z^{(k)}$ represent their probability $p^\star_k$.

As $p^\star$ is unknown, one has to collect data, e.g., by means of counting passersby at the locations $z^{(k)}$. As this may be a costly operation, it is important that the acquired data is used efficiently. Furthermore, in order to ensure that the potentially large up-front investment is justified, we are required to provide a certificate stating that, with high confidence, the quality of the solution will be no worse than what is predicted. Thus, given our collected sample of size m, our aim is to compute estimates $\hat{V}_m$, satisfying (3). We compare the following data-driven methods.

CADRO Solves (DRO) according to Alg. 1, setting $\tau(m)$ as in (12), with $\mu = 0.01, \nu = 0.8$.

$\mathcal{D}$-DRO Solves (DRO), with an ambiguity set of the form $\mathcal{A}_m = \{p \in \Delta_d \mid \mathcal{D}(\hat{p}_m, p) \leq r_m^{\mathcal{D}}\}$, with $\mathcal{D} \in \{\mathrm{TV}, \mathrm{KL}, \mathrm{W}\}$ the total variation, Kullback-Leibler, and Wasserstein distance/divergence[1] (cf. [12, Tb. I]). $r_m^{\mathrm{TV}}, r_m^{\mathrm{KL}}$ are selected according to [26, Thm 2.1][2], [6, Thm. 5], respectively, and $r_m^{\mathrm{W}} = \mathbf{max}_{i,j\in[d]} K_{ij} r_m^{\mathrm{TV}}$ [28], ensuring that (5) is satisfied.

SAA Using the same data partition $\{\hat{\Xi}_{\mathrm{T}}, \hat{\Xi}_{\mathrm{C}}\}$ as CADRO, we use $\hat{\Xi}_{\mathrm{T}}$ to compute $x_m = \bar{x}_{\tau(m)}$ as in (10), and we use $\hat{\Xi}_{\mathrm{C}}$ to obtain a high-confidence upper bound $\hat{V}_m = \alpha_{\hat{\Xi}_{\mathrm{C}}}(L(\bar{x}_{\tau(m)}))$, utilizing Proposition III.2.

Note that $\mathcal{D}$-DRO does not require an independent data sample in order to satisfy (3).

Remark V.1. Other methods could be used to validate SAA (e.g., cross-validation [2], replications [7]), but these methods only guarantee the required confidence level asymptotically. In order to obtain a fair comparison, we instead use the same mean bound, namely (7) for both CADRO and SAA, so both methods provide the same theoretical guarantees. Moreover, we note that a different data partition could be used for SAA. However, preliminary experiments have indicated that significantly increasing or decreasing $\tau(m)$ resulted in deteriorated bounds on the cost.

We set $n_x = 3$, $d = 50$, $\beta = 0.01$, and apply each method for 100 independently drawn datasets of size m. In Fig. 3, we

[1] We use $K_{ij} = \|z^{(i)} - z^{(j)}\|_2$, $i, j \in [d]$ as the transportation cost.

[2] This is a slightly improved version of the classical Bretagnolle-Huber-Carol inequality [27, Prop. A.6.6].

plot the estimated costs $\hat{V}_m$ and the achieved out-of-sample cost $V(\hat{x}_m, p^\star)$, for increasing values of m. We observe that CADRO provides a sharper cost estimate $\hat{V}_m$ than the other approaches. In particular, the classical DRO formulations require relatively large amounts of data before obtaining a non-vacuous upper bound on the cost. The right-hand panel in Fig. 3 shows that additionally, CADRO returns solutions which exhibit superior out-of-sample performance than the compared approaches, illustrating that it does not rely on conservative solutions to obtain better upper bounds.

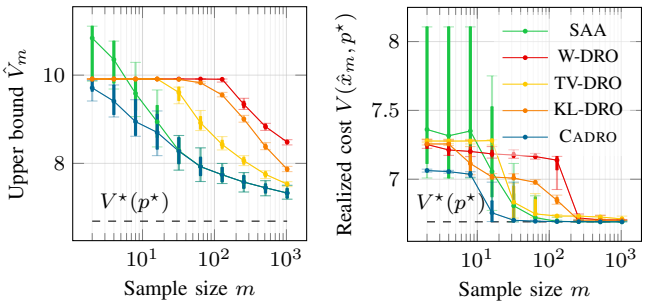

Fig. 3. Results of the facility location problem of Section V. (left): The cost estimates $\hat{V}_m$ satisfying (3) and (4); (right): True out of sample cost $V(\hat{x}_m, p^\star)$. The points indicate the sample mean, the solid errorbars indicate the empirical 0.95 quantiles and the semi-transparent errorbars indicate the largest and smallest values over 100 independent runs.

VI. CONCLUSION AND FUTURE WORK

We proposed a DRO formulation, in which the ambiguity set is designed to only restrict errors in the distribution that are predicted to have significant effects on the worst-case expected cost. We proved out-of-sample performance bounds and consistency of the resulting DRO scheme, and demonstrated empirically that this approach may be used to robustify against poor distribution estimates at small sample sizes, while remaining considerably less conservative than existing DRO formulations. In future work, we aim to extend the work to continuous distributions.

REFERENCES

[1] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*. MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 3 ed., July 2021.

[2] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, New York, NY: Springer, 2nd ed ed., 2009.

[3] A. Mesbah, "Stochastic Model Predictive Control: An Overview and Perspectives for Future Research," *IEEE Control Systems Magazine*, vol. 36, pp. 30–44, Dec. 2016.

[4] J. O. Royset and R. J.-B. Wets, *An Optimization Primer*. Springer Series in Operations Research and Financial Engineering, Cham, Switzerland: Springer, 2021.

[5] J. E. Smith and R. L. Winkler, "The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis," *Management Science*, vol. 52, pp. 311–322, Mar. 2006.

[6] B. P. G. Van Parys, P. M. Esfahani, and D. Kuhn, "From Data to Decisions: Distributionally Robust Optimization Is Optimal," *Management Science*, vol. 67, pp. 3387–3402, June 2021.

[7] G. Bayraksan and D. P. Morton, "Assessing solution quality in stochastic programs," *Mathematical Programming*, vol. 108, pp. 495–514, Sept. 2006.

[8] E. Delage and Y. Ye, "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems," *Operations Research*, vol. 58, pp. 595–612, June 2010.

[9] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, pp. 115–166, Sept. 2018.

[10] A. Hakobyan and I. Yang, "Distributionally Robust Risk Map for Learning-Based Motion Planning and Control: A Semidefinite Programming Approach," *IEEE Transactions on Robotics*, pp. 1–20, 2022.

[11] M. Schuurmans, A. Katriniok, C. Meissen, H. E. Tseng, and P. Patrinos, "Safe, learning-based MPC for highway driving under lane-change uncertainty: A distributionally robust approach," *Artificial Intelligence*, vol. 320, p. 103920, July 2023.

[12] M. Schuurmans and P. Patrinos, "A General Framework for Learning-Based Distributionally Robust MPC of Markov Jump Systems," *IEEE Transactions on Automatic Control*, pp. 1–16, 2023.

[13] G. Bayraksan and D. K. Love, "Data-Driven Stochastic Programming Using Phi-Divergences," in *The Operations Research Revolution* (D. Aleman, A. Thiele, J. C. Smith, and H. J. Greenberg, eds.), pp. 1–19, INFORMS, Sept. 2015.

[14] P. Coppens, M. Schuurmans, and P. Patrinos, "Data-driven distributionally robust LQR with multiplicative noise," in *Learning for Dynamics and Control*, pp. 521–530, PMLR, July 2020.

[15] H. Rahimian and S. Mehrotra, "Frameworks and Results in Distributionally Robust Optimization," *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022.

[16] F. Lin, X. Fang, and Z. Gao, "Distributionally Robust Optimization: A review on theory and applications," *Numerical Algebra, Control & Optimization*, vol. 12, no. 1, p. 159, 2022.

[17] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of *Grundlehren Der Mathematischen Wissenschaften*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[18] H. Rahimian, G. Bayraksan, and T. Homem-de-Mello, "Identifying effective scenarios in distributionally robust stochastic programs with total variation distance," *Mathematical Programming*, vol. 173, pp. 393–430, Jan. 2019.

[19] M. Schuurmans and P. Patrinos, "Distributionally Robust Optimization using Cost-Aware Ambiguity Sets," Mar. 2023, arXiv: 2303.09408.

[20] T. Anderson, "Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function," Technical Report AD0696676, Stanford University CA Dept. of Statistics, Oct. 1969.

[21] P. Coppens and P. Patrinos, "Robustified Empirical Risk Minimization with Law-Invariant, Coherent Risk Measures," Mar. 2023, arXiv: 2303.09196.

[22] S. S. Wilks, *Mathematical Statistics*. A Wiley Publication in Mathematical Statistics, New York: Wiley, 2. print ed., 1963.

[23] P. Sopasakis, M. Schuurmans, and P. Patrinos, "Risk-averse risk-constrained optimal control," in *2019 18th European Control Conference (ECC)*, pp. 375–380, June 2019.

[24] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics, Jan. 2001.

[25] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK ; New York: Cambridge University Press, 2004.

[26] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the L1 Deviation of the Empirical Distribution," tech. rep., Information Theory Research Group, HP Laboratories Palo Alto, Palo Alto, California, 2003.

[27] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer, 2000.

[28] A. L. Gibbs and F. E. Su, "On Choosing and Bounding Probability Metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.