Uncertain uncertainty in data-driven stochastic optimization: towards structured ambiguity sets

Chaouach, L.; Boskos, D.; Oomen, T.A.E.

DOI
10.1109/CDC51059.2022.9992405

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the IEEE 61st Conference on Decision and Control (CDC 2022)

2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992405

Uncertain uncertainty in data-driven stochastic optimization: towards structured ambiguity sets

Lotfi M. Chaouach Dimitris Boskos Tom Oomen

Abstract— **Ambiguity sets of probability distributions are a prominent tool to hedge against distributional uncertainty in stochastic optimization. The aim of this paper is to build tight Wasserstein ambiguity sets for data-driven optimization problems. The method exploits independence between the distribution components to introduce structure in the ambiguity sets and speed up their shrinkage with the number of collected samples. Tractable reformulations of the stochastic optimization problems are derived for costs that are expressed as sums or products of functions that depend only on the individual distribution components. The statistical benefits of the approach are theoretically analyzed for compactly supported distributions and demonstrated in a numerical example.**

I. Introduction

Uncertainty is abundant in today's real-world systems. Their increasing complexity, sophistication, and connectivity generate further sources of uncertainty that need to be taken into account when optimizing their performance. These uncertainties can be handled through deterministic or stochastic approaches. While we account for the worst-case scenario in the deterministic framework, the probabilistic approach provides the flexibility of excluding events that are highly unlikely to occur, thus, reducing the potential conservativeness of worst-case designs. Probabilistic models usually assume the existence of a distribution that fully describes their stochastic attributes. In practice, this distribution is often not known and needs to be approximated by leveraging information regarding the system of interest. Typically, it is inferred using collected data from the system to formulate data-driven stochastic optimization problems that optimize the system's performance.

To hedge against uncertainty about the probabilistic model, stochastic optimization problems are reinforced via distributionally robust formulations that optimize the worst-case performance over an ambiguity set of plausible distribution models [31]. This distibutionally robust optimization (DRO) framework is attracting increasing attention to solve stochastic optimization problems across fields such as operations research [1], statistical learning [17], and control [20].

There are several methods to build ambiguity sets that robustify the decisions of optimization problems. Among the most prominent tools to group distributions into an ambiguity set are statistical divergences [7], [16], moment constraints [10], [23], total variation metrics [27], and optimal transport metrics [21], such as the Wasserstein distance [29]. Optimal transport ambiguity sets are widely used for data-driven problems. Among the main reasons are that they are accompanied by statistical guarantees [12] and that they facilitate tractable reformulations of the associated DRO problems [3], [14], [19].

Applications of DRO formulations are widespread across control engineering and related fields. The work [28] develops a distributionally robust LQR framework. Data-driven aspects of Wasserstein distributionally robust stochastic control are found in [32], while [25] considers the problem of Kalman filtering under distributional uncertainty. Dynamic aspects of data-driven ambiguity sets with probabilistic guarantees are considered in [5] and [6], which accounts for data assimilation nonidealities. Further applications of DRO include economic dispatch in power systems [22], congestion avoidance in traffic control [18], and motion planning in dynamic environments [15].

One type of statistical guarantees for Wasserstein ambiguity sets is to ensure that they contain the data-generating distribution with prescribed probability. This approach suffers from the curse of dimensionality since the size of Wasserstein ambiguity sets exhibits very slow decay rates with respect to the number of samples for high-dimensional data [11], [12], [30]. To ameliorate this drawback, a recent line of work informs the ambiguity sets by the specific optimization problem, restoring favorable decay rates [2], [4], [13], [24]. Still, the curse of dimensionality persists when solving multiple optimization problems under the same uncertainty as for instance in model predictive control [8].

Although it is important to construct optimal transport ambiguity sets with probabilistic guarantees of containing the distribution of the data, existing approaches for this purpose provide conservative characterizations for high-dimensional uncertainty. In this paper we tackle the curse of dimensionality by leveraging independence between lower dimensional components of the random variables.

We develop structured Wasserstein ambiguity sets that only contain product distributions and deduce confidence guarantees from their constituent components. These ambiguity sets shrink at much faster rates compared to their monolithic counterparts while containing the data-generating distribution with the same confidence level. For certain classes of cost functions we provide tractable dual reformulations of the associated DRO problems. A simulation example is also provided to demonstrate the effectiveness of the presented approach. Due to space constraints, the proofs

All the authors are with the Delft Center for Systems and Control, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology. Tom Oomen is also with the Control Systems Technology Group, Department of Mechanical Engineering, Eindhoven University of Technology {L.Chaouach,D.Boskos}@tudelft.nl,T.A.E.Oomen@tue.nl.

are omitted and will appear elsewhere.

II. Preliminaries and notation

Throughout this paper, we use the following notation. We denote by $\|\cdot\|_p$ the pth norm in $\mathbb{R}^d$, $p \in [1, \infty]$ and omit the index in the Euclidean case $p = 2$. The diameter of $S \subset \mathbb{R}^d$ is $\operatorname{diam}(S) := \sup\{\|x - y\|_\infty \,|\, x, y \in S\}$.

Probability theory: We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-algebra on $\mathbb{R}^d$, and by $\mathcal{P}(\mathbb{R}^d)$ the space of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The Dirac distribution centered at $\xi \in \mathbb{R}^d$ is denoted by δ_ξ. Given $p \geq 1$, we denote by $\mathcal{P}_p(\mathbb{R}^d)$ the set of probability measures in $\mathcal{P}(\mathbb{R}^d)$ with finite pth moment. Given $P,\ Q \in \mathcal{P}_p(\mathbb{R}^d)$, their pth Wasserstein distance is

$$W_p(P, Q) := \left(\inf_{\pi \in \mathcal{M}(P,Q)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \pi(dx, dy) \right\} \right)^{\frac{1}{p}},$$

(cf. [29]). Each $\pi \in \mathcal{M}(P, Q)$ is a transport plan, i.e., a distribution on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q, respectively. Namely, $\pi(A \times \mathbb{R}^d) = P(A)$ for any $A \in \mathcal{B}(\mathbb{R}^d)$ and analogously for Q. We denote by $P \otimes Q$ the product measure of P and Q. For any $P \in \mathcal{P}(\mathbb{R}^d)$, its support is the closed set $\operatorname{supp}(P) := \{x \in \mathbb{R}^d \,|\, P(U) > 0$ for each neighborhood U of $x\}$. Given a function $X : \Omega \to \mathbb{R}$ with the σ-algebra $\mathcal{B}(\mathbb{R})$ we denote by $\sigma(X)$ the σ-algebra generated by X on Ω.

III. Problem formulation

Here we introduce stochastic optimization problems and focus on their robustification in data-driven scenarios.

A. Data-driven stochastic optimization

The central problem in stochastic optimization is to make optimal decisions in problems affected by randomness. A stochastic optimization problem takes the form

$$\inf_{x \in \mathcal{X}} \mathbb{E}_{P_\xi}\big[f(x, \xi)\big] \tag{1}$$

where f is the objective function, $x \in \mathcal{X}$ is the decision variable, and $\xi \in \mathbb{R}^d$ is a random variable with distribution P_ξ. In practice, this distribution is often unknown and there is only access to a finite number of i.i.d. samples $\xi^1, \ldots, \xi^N$ of ξ. A typical approach to solve (1) in this case is to approximate P_ξ by the empirical distribution $P_\xi^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$ of the samples. This problem formulation, also known as the sample average approximation (SAA) converges to the solution of (1) in the asymptotic limit [26], and provides reliable decisions for large amounts of data.

B. Distributionally robust optimization

When the amount of available data is limited, the SAA approach may lead to inaccurate approximations of the true distribution. In this case, the empirical distribution P_ξ^N may exhibit significant deviations from the true distribution P_ξ and lead to overly optimistic values of the SAA. To address this issue, uncertainty in the distribution is incorporated into problem (1) under the robust formulation

$$\inf_{x \in \mathcal{X}} \sup_{P_\xi \in \mathcal{P}^N} \mathbb{E}_{P_\xi}\big[f(x, \xi)\big]. \tag{2}$$

In this distributionally robust optimization (DRO) problem, $\mathcal{P}^N$ is an ambiguity set of distributions that contains plausible models for the true distribution and can be inferred from the collected samples.

A suitable way to build data-driven ambiguity sets for problem (2) is to group all distributions up some distance ε from the empirical distribution P_ξ^N in the Wasserstein metric. Namely, $\mathcal{P}^N$ in (2) is selected to be the ball

$$\mathcal{B}_p(P_\xi^N, \varepsilon) := \{P \in \mathcal{P}_p(\mathbb{R}^d) \,|\, W_p(P_\xi^N, P) \leq \varepsilon\},$$

for certain $p \geq 1$, with center P_ξ^N and radius ε. Among the benefits of this choice are that Wasserstein distances penalize horizontal distribution variations, and hence, their effect on the optimization problem, and that Wasserstein balls lead to tractable DRO problems [19]. In addition, these ambiguity balls do not rely on absolute continuity conditions between the associated distributions and they have finite-sample guarantees of containing the true distribution. In particular, using concentration of measure results [12], we can tune the radius of $\mathcal{B}_p(P_\xi^N, \varepsilon)$ so that it contains the true distribution with prescribed confidence. This way, the value of (2) provides an upper bound for the expected cost (1) with prescribed confidence.

C. Structured ambiguity sets

The size of the ambiguity set $\mathcal{P}^N$ has a direct effect on the solution of (2) since ambiguity balls of larger sizes may lead to conservative upper bounds for (1). This motivates the consideration of appropriate structure for the ambiguity sets to mitigate their potential conservativeness.

Problem formulation: Given a fixed amount of data, determine an ambiguity set of appropriate structure so that it contains the true distribution with prescribed confidence and its size is as small as possible.

To address the problem, we make the following assumption regarding the class of the random variable.

Assumption 3.1: ***(Independence of random variable components).*** The components of $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n} \equiv \mathbb{R}^d$ are independent random variables.

This assumption is reasonable in several problems such as in networked systems, where random inputs at different network locations do not essentially affect each other, or the deployment of multi-robot systems where the sensing capabilities of the individual agents are subject to independent disturbances.

Due to Assumption 3.1, the probability distribution of ξ is expressed as the product measure

$$P_\xi = P_{\xi_1} \otimes \ldots \otimes P_{\xi_n}, \tag{3}$$

with $P_{\xi_k},\ k = 1, \ldots, n$ denoting the distributions of its components. Thus, instead of looking for plausible probabilistic descriptions of P_ξ in an ambiguity ball, the idea is to represent these descriptions through an ambiguity set that only contains product measures. Since this set contains a restricted class of distributions, it should yield less conservative solutions for (2) under the same confidence.

D. Ambiguity radius

Using tools from concentration of measure it is possible to tune the ambiguity radius so that it contains the true distribution with prescribed probability. These results leverage prior assumptions about the class where the unknown distribution belongs. Such assumptions are the size of its support (e.g., [12, Proposition 10], [30]), its tail decay rate (e.g., [12, Theorem 2 cases (1) and (2)]), or bounds on its moments (e.g., [12, Theorem 2 case (3)], [9]). Based on these results, for any confidence $1-\beta$ and number of samples N, we can select the ambiguity radius $\varepsilon(N,\beta)$ so that

$$\mathbb{P}(P_\xi \in \mathcal{B}_p(P_\xi^N,\varepsilon)) \geq 1-\beta. \qquad (4)$$

When we are interested in highlighting the decrease rate of the ambiguity radius with the number of samples, we will consider compactly supported distributions, which facilitate the derivation of convenient concentration bounds. In particular, choosing the ambiguity radius

$$\varepsilon(N,\beta,\rho) := \begin{cases} \left(\frac{\ln(C\beta^{-1})}{c}\right)^{\frac{1}{2p}} \frac{\rho}{N^{\frac{1}{2p}}}, & \text{if } p > d/2, \\ h^{-1}\left(\frac{\ln(C\beta^{-1})}{cN}\right)^{\frac{1}{p}}, & \text{if } p = d/2, \\ \left(\frac{\ln(C\beta^{-1})}{c}\right)^{\frac{1}{d}} \frac{\rho}{N^{\frac{1}{d}}}, & \text{if } p < d/2, \end{cases} \qquad (5)$$

guarantees that (4) holds, where h^{-1} is the inverse of $h(x) = \frac{x^2}{(\ln(2+1/x))^2}$, $x > 0$, and ρ is the diameter of the support of P_ξ (cf. [5, Corollary 3.3]). For high-dimensional random variables, (5) implies that the radius decreases with the slow rate of the order of $N^{-\frac{1}{d}}$. As a consequence, the exploitation of more data does not guarantee any significant improvement of the closeness between the true distribution and its empirical approximation, and hence, also of the size of the ambiguity ball. This brings us to the quantitative question that we address under Assumption 3.1: *Exploiting independence of the components of ξ determine an ambiguity set structure that does not suffer from the curse of dimensionality with respect to d.*

IV. Ambiguity hyperrectangles

To address the conservative radius decrease of high-dimensional ambiguity balls, we exploit independence of lower-dimensional components of the random variable $\xi = (\xi_1,\ldots,\xi_n)$. Using N i.i.d. samples $\xi^1,\ldots,\xi^N$, we first build a lower-dimensional ambiguity ball $\mathcal{B}_p(P_{\xi_k}^N,\varepsilon_k)$ for each component of ξ, where $P_{\xi_k}^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi_k^i}$ denotes its corresponding empirical distribution. From these balls, we construct the *ambiguity (Wasserstein) hyperrectangle*

$$\begin{aligned} \mathcal{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon}) := \{&P'_{\xi_1}\otimes\cdots\otimes P'_{\xi_n} \mid \\ &P'_{\xi_k} \in \mathcal{B}_p(P_{\xi_k}^N,\varepsilon_k),\ k=1,\ldots,n\} \qquad (6a) \\ \boldsymbol{P}_\xi^N := P_{\xi_1}^N &\otimes\ldots\otimes P_{\xi_n}^N, \quad \boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n), \quad (6b) \end{aligned}$$

by taking the product measures across the individual distributions from the balls. We will also refer to $\boldsymbol{P}_\xi^N$ in (6b) as the *product empirical distribution*.

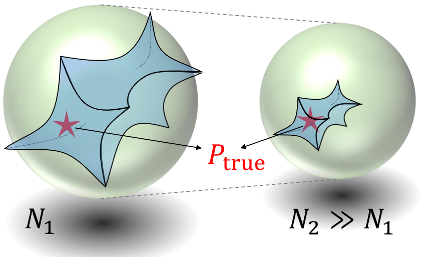

Fig. 1. High-dimensional hyperrectangles (in blue) shrink much faster with the number of samples compared to Wasserstein ambiguity balls while containing the true distribution with the same confidence.

We next establish probabilistic guarantees for the Wasserstein hyperrectangles, which ensure that they contain the distribution of ξ with prescribed confidence. We also exploit them to alleviate the curse of dimensionality regarding the convergence of Wasserstein balls for specific distribution classes, cf. Figure 1. The following result establishes the guarantees that an ambiguity hyperrectangle inherits from its lower-dimensional constituent ambiguity balls.

Theorem 4.1: ***(Probabilistic guarantees for Wasserstein hyperrectangles).*** Assume that the random variable ξ satisfies the independence Assumption 3.1 and that $P_\xi \in \mathcal{P}_p(\mathbb{R}^d)$. Given i.i.d. samples $\xi^1,\ldots,\xi^N$ of ξ, let $P_{\xi_1}^N,\ldots,P_{\xi_n}^N$ be the empirical distributions of the individual components. Assume also that each Wasserstein ball $\mathcal{B}_p(P_{\xi_k}^N,\varepsilon_k)$ contains P_{ξ_k} with confidence $1-\beta_k$. Then, the hyperrectangle $\mathcal{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})$ given by (6) contains P_ξ with confidence $\prod_{k=1}^n 1-\beta_k$.

The proof of Theorem 4.1 is based on the following lemma.

Lemma 4.2: ***(Independent Wasserstein distances across empirical distributions).*** Assume that the random variable ξ satisfies the independence Assumption 3.1 and that $P_\xi \in \mathcal{P}_p(\mathbb{R}^d)$. Given i.i.d. samples $\xi^1,\ldots,\xi^N$ of ξ, let $P_{\xi_1}^N,\ldots,P_{\xi_n}^N$ be the empirical distributions of the individual components. Then for any $\varepsilon_1,\ldots,\varepsilon_n \geq 0$ the events $\{W_p(P_{\xi_k}^N,P_{\xi_k}) \leq \varepsilon_k\}$, $k=1,\ldots,n$ are independent.

We next quantify the size reduction of Wasserstein hyperrectangles compared to Wasserstein balls that are constructed using the same samples and the same confidence. The results are focused on compactly supported distributions that facilitate the exact computations of the ambiguity radii. We will use the following probabilistic bounds for the Wasserstein distance between the true and the empirical distribution.

Proposition 4.3: ***(Ambiguity radius [6, Proposition 24]).*** Assume that the probability distribution P_ξ is supported on $\Xi \subset \mathbb{R}^d$ with $\rho := \mathrm{diam}(\Xi) < \infty$. Assume also that $d \geq 2p+1$ and let $\xi^1,\ldots,\xi^N$ be i.i.d. samples of ξ. Then the ambiguity radius

$$\varepsilon \equiv \varepsilon(N,\beta,\rho,p,d) := \rho\varepsilon_\star(\beta,p,d)N^{-\frac{1}{d}}, \qquad (7)$$

where

$$\begin{aligned} \varepsilon_\star(\beta,p,d) &:= \sqrt{d}2^{\frac{1}{2p}}(C(d,p)+(\ln\beta^{-1})^{\frac{1}{2p}}) \\ C(d,p) &:= 2^{(d-2)/2p}\Big(\frac{1}{2^{1/2}-1}+\frac{1}{2^{1/2}-2^{1/2-p}}\Big)^{\frac{1}{p}}, \end{aligned}$$

and $1-\beta$ is a desired confidence level, guarantees that

$$\mathbb{P}(P_\xi \in \mathcal{B}_p(P_\xi^N, \varepsilon)) \geq 1-\beta.$$

We use this ambiguity radius characterization and the guarantees of Theorem 4.1 to determine a ball around the product empirical distribution that contains the hyperrectangle with prescribed probability.

Proposition 4.4: ***(Ambiguity rectangle containment).*** Assume that the random variable ξ is supported on $\Xi \subset \mathbb{R}^d$ with $\rho := \operatorname{diam}(\Xi) < \infty$ and satisfies Assumption 3.1. For any confidence $1-\beta$, consider the Wasserstein hyperrectangle $\mathcal{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ given by (6) with each $\varepsilon_k = \varepsilon(N, \beta_k, \rho, p, d_k)$ as in (7) and $\beta_k = \beta\frac{d_k}{d}$. Then the hyperrectangle contains P_ξ with confidence $1-\beta$ and satisfies

$$\mathcal{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon}) \subset \mathcal{B}_p(\boldsymbol{P}_\xi^N, \varepsilon'), \tag{8}$$

where

$$\varepsilon' = cn^{1/p+\max\{0,1/2-1/p\}}\rho\varepsilon_\star(\beta,p,d)N^{-\frac{1}{d_{\max}}},$$

$c = (\sqrt{5}+1)/\big(2e^{\frac{(\sqrt{5}+1)^2}{4}}\big) \approx 1.1043$, $d_{\max} := \max_{k=1,\ldots,n} d_k$, and $\varepsilon_\star$ is defined in Proposition 4.3.

Remark 4.5: ***(Ambiguity rectangle vs ball shrinkage).*** Under the assumptions of Proposition 4.4, we can compare the size of a hyperrectangle and a monolithic ball that contain P_ξ with the same confidence. The radius ε' of the Wasserstein ball that encloses the hyperrectangle is guaranteed to be strictly smaller than the radius ε of the monolithic ball when $N \geq \big(cn^{1/p+\max\{0,1/2-1/p\}}\big)^{\frac{1}{d_{\max}}-\frac{1}{d}}$, and decreases much faster for larger N (cf. Figure 2(a)). The "centroid" of the hyperrectangle, namely, the center $\boldsymbol{P}_\xi^N$ of its enclosing ball is different from the center P_ξ^N of the monolithic ball, since the latter is the empirical distribution $\frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$, whereas the former is the product empirical distribution

$$P_{\xi_1}^N \otimes \cdots \otimes P_{\xi_n}^N = \frac{1}{N^n} \sum_{(i_1,\ldots,i_n)\in\{1,\ldots,N\}^n} \delta_{(\xi_1^{i_1},\ldots,\xi_n^{i_n})}.$$

(cf. Figure 2(b)). This also implies that under Assumption 3.1, an ambiguity ball that is centered at the product empirical distribution $\boldsymbol{P}_\xi^N$ will contain the true distribution with significantly higher probability compared to when it is centered at the empirical distribution P_ξ^N. As a result, when the components of the random variable are independent, shifting the center of the ambiguity ball to the product empirical distribution $\boldsymbol{P}_\xi^N$ provides an ambiguity set that is better informed about the true distribution.

V. DRO Reformulations over Hyperrectangles

In this section, we provide tractable reformulations of the DRO problem (2) when the ambiguity set $\mathcal{P}^N$ is a Wasserstein hyperrectangle. Namely, we provide tractable equivalent forms of the problem

$$\inf_{x\in\mathcal{X}} \sup_{P_\xi \in \mathcal{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})} \mathbb{E}_{P_\xi}\big[f(x,\xi)\big]. \tag{9}$$

As common in DRO, computational tractability relies on the reformulation of the inner maximization problem. To facilitate notation, we fix the decision variable x and denote $g(\xi) := f(x,\xi)$. Thus, we are interested in reformulating the inner problem

$$\sup_{P_\xi \in \mathcal{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})} \mathbb{E}_{P_\xi}[g(\xi)]. \tag{10}$$

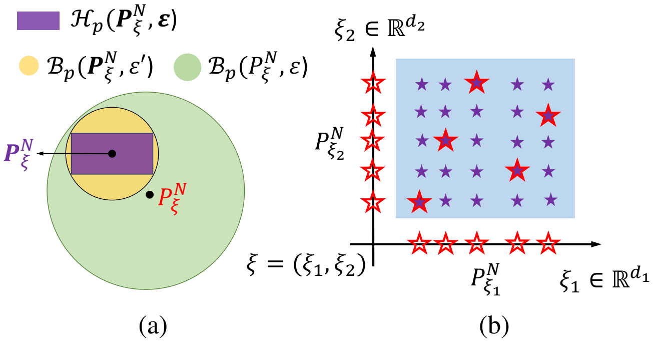
$\mathcal{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$
$\mathcal{B}_p(\boldsymbol{P}_\xi^N, \varepsilon')$
$\mathcal{B}_p(P_\xi^N, \varepsilon)$
$\boldsymbol{P}_\xi^N$
P_ξ^N
$\xi_2 \in \mathbb{R}^{d_2}$
$P_{\xi_2}^N$
$\xi = (\xi_1, \xi_2)$
$P_{\xi_1}^N$
$\xi_1 \in \mathbb{R}^{d_1}$
(a)
(b)

Fig. 2. (a) shows the Wasserstein hyperrectangle $\mathcal{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$, its enclosing ball $\mathcal{B}_p(\boldsymbol{P}_\xi^N, \varepsilon')$ around the product empirical distribution $\boldsymbol{P}_\xi^N$, and the monolithic Wasserstein ball around the empirical distribution P_ξ^N for a random variable with two components. (b) depicts the empirical distribution P_ξ^N from (a) with the filled red stars, which correspond to the N samples, and the product empirical distribution $\boldsymbol{P}_\xi^N$ with the N^2 purple stars. The purple stars are formed by taking the product of the marginal empirical distributions $P_{\xi_1}^N$ and $P_{\xi_2}^N$ of the independent components of ξ, depicted by the hollow red stars.

We make the following assumption for g.

Assumption 5.1: ***(Sum/product decomposition).*** The objective function can be expressed as the sum of upper semicontinuous functions or the product of nonnegative upper semicontinuous functions that depend only on the respective components of the random variable. Namely,

$$g(\xi) = \sum_{k=1}^n g_k(\xi_k) \tag{11a}$$

$$\text{or} \quad g(\xi) = \prod_{k=1}^n g_k(\xi_k), \qquad g_k(\xi_k) \geq 0. \tag{11b}$$

Notice that this assumption is satisfied if and only if f can also be written as the sum or product of appropriate functions, respectively, that depend only on the individual components of ξ. To reformulate (10) we will leverage the following strong duality result for the maximization over Wasserstein balls.

Proposition 5.2: ***(DRO dual over Wasserstein balls [3, Theorem 1 & Remark 1]).*** Consider the i.i.d. samples $\xi^1, \ldots, \xi^N$ of the random variable ξ, the Wasserstein ball $\mathcal{B}_p(P_\xi^N, \varepsilon)$, and the upper semicontinuous function g. Then

$$\begin{aligned}&\sup_{P_\xi \in \mathcal{B}_p(P_\xi^N, \varepsilon)} \mathbb{E}_{P_\xi}[g(\xi)] \\ &= \inf_{\lambda\geq 0}\left\{\frac{1}{N}\sum_{i=1}^N \sup_{\xi\in\mathbb{R}^d}\{g(\xi) + \lambda(\varepsilon^p - \|\xi-\xi^i\|^p)\}\right\}. \end{aligned} \tag{12}$$

We next establish strong duality for DRO problems with Wasserstein hyperrectangles when the objective function satisfies Assumption 5.1.

Proposition 5.3: ***(DRO dual over Wasserstein hyperrectangles).*** Consider the optimization problem (10) and let the objective function g satisfy Assumption 5.1. Then (10) admits the equivalent dual formulations

$$\inf_{\boldsymbol{\lambda}\geq 0}\sum_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}\sup_{\xi_k\in\mathbb{R}^{d_k}}\{g_k(\xi_k)+\lambda_k(\varepsilon_k^p-\left\|\xi_k-\xi_k^i\right\|^p)\} \tag{13a}$$

$$\inf_{\boldsymbol{\lambda}\geq 0}\prod_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}\sup_{\xi_k\in\mathbb{R}^{d_k}}\{g_k(\xi_k)+\lambda_k(\varepsilon_k^p-\left\|\xi_k-\xi_k^i\right\|^p)\}, \tag{13b}$$

corresponding to Assumptions (11a) and (11b), respectively, where $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_n)$ and $\boldsymbol{\lambda}\geq 0$ holds componentwise.

Remark 5.4: The additive duality result of Proposition 5.3 can be applied to optimization problems that have separable additive costs with respect to the uncertainty. In an analogous manner, the multiplicative version can be applied to problems with nonnegative multiplicative costs, such as uncertainty quantification problems which seek to determine the product of probabilities across independent events.

VI. Simulation Example

In this section, we present an uncertainty quantification problem that illustrates the advantage of using ambiguity hyperrectangles compared to monolithic ambiguity balls. We consider four drones that need to reach a region within a specific deadline to assist a team with a collaborative search-and-rescue mission. The probability that they reach the region before the deadline determines if a fallback plan will be used for the mission or not. The drones are operating in different locations and are simultaneously informed by distinct operators to reach the region as fast as possible. Each operator sends to the team information about the distance and maximum velocity of the drones, which are unknown, independent across the drones, and inferred by a limited amount of data. The goal is to build an ambiguity set from these data and determine a lower bound for the probability that all drones can reach the region before the deadline. Namely, we seek to determine the worst-case (smallest) probability that this happens among all distributions in the ambiguity set.

Denote by τ the deadline and by r_k and v_k the initial distance and maximum velocity of each drone. A necessary and sufficient condition for the drones to reach the region is

$$\tau v_k - r_k \geq 0 \iff a_k\xi\leq 0 \quad \forall k=1,\ldots,4,$$

where $a_k\in\mathbb{R}^{1\times 8}$, with $a_k(j)=1$, if $j=2k-1$, $a_k(j)=-\tau$, if $j=2k$, $a_k(j)=0$, otherwise, and $\xi=(\xi_1,\ldots,\xi_4)\equiv(r_1,v_1,\ldots,r_4,v_4)$ is the random variable of our problem. Denoting by $A\in\mathbb{R}^{4\times 8}$ the matrix with rows a_k and

$$B_k := \{\xi \,|\, a_k\xi\leq 0\} \tag{14a}$$

$$B := \cap_{k=1}^4 B_k \equiv \{\xi \,|\, A\xi\leq 0\}, \tag{14b}$$

we seek to determine the probability bound

$$\min_{P_\xi\in\mathcal{P}^N} P_\xi(B) = \min_{P_\xi\in\mathcal{P}^N}\mathbb{E}_{P_\xi}[\mathbf{1}_B(\xi)], \tag{15}$$

where the ambiguity set $\mathcal{P}^N$ is built by N independent samples of ξ. We will compare these probability bounds when $\mathcal{P}^N$ is either an ambiguity ball or a hyperrectangle.

We assume that the distribution of ξ is compactly supported and build the ambiguity set using the results from Section IV. Since the confidence bounds of Proposition 4.3 may become conservative (see e.g., [6, Section 6.5]), we use them to tune the relative sizes between a monolithic ambiguity ball and an associated hyperrectangle. In particular, using the exact same reasoning as in the proof of Proposition 4.4, under the same confidence level, it can be shown that the radius ε_k of each hyperrectangle component satisfies

$$\varepsilon_k \leq c\frac{\rho_k}{\rho}N^{-\frac{1}{3}+\frac{1}{8}}\varepsilon,$$

with ρ_k, ρ the diameters of the supports of ξ_k, ξ, and ε the radius of the monolithic ambiguity ball. Thus, after selecting the radius of the monolithic ball, we pick the radii of the hyperrectangles to satisfy the above relation as an equality.

To solve the optimization problem (15) over the monolithic ball $\mathcal{P}^N\equiv\mathcal{B}_p(P_\xi^N,\varepsilon)$, we rewrite it as

$$\min_{P_\xi\in\mathcal{B}_p(P_\xi^N,\varepsilon)} P_\xi(B) = 1-\max_{P_\xi\in\mathcal{B}_p(P_\xi^N,\varepsilon)} P_\xi(B^c), \tag{16}$$

where B^c denotes the complement of the set B. To obtain (15) for the case of the Wasserstein hyperrectangle $\mathcal{P}^N\equiv\mathcal{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})$, we compute (15) by following the reasoning of the proof of Proposition 5.3, namely

$$\begin{aligned}
\min_{P_\xi\in\mathcal{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})} P_\xi(B) &= \min_{P_\xi\in\mathcal{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})}\mathbb{E}_{P_\xi}[\mathbf{1}_B(\xi)]\\
&= \min_{P_\xi\in\mathcal{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})}\mathbb{E}_{P_\xi}\Big[\prod_{k=1}^4\mathbf{1}_{B_k}(\xi_k)\Big]\\
&= \min_{P_{\xi_k}\in\mathcal{B}_p(P_{\xi_k}^N,\varepsilon_k),k=1,\ldots,4}\prod_{k=1}^4\mathbb{E}_{P_{\xi_k}}[\mathbf{1}_{B_k}(\xi_k)]\\
&= \prod_{k=1}^4\Big(1-\max_{P_{\xi_k}\in\mathcal{B}_p(P_{\xi_k}^N,\varepsilon_k)}P_{\xi_k}(B_k^c)\Big).
\end{aligned} \tag{17}$$

Due to (14), both (16) and (17) involve robust uncertainty quantification problems over polytopic sets and can be computed using the reformulations [19, Corollary 5.3] for the 1-Wasserstein distance (i.e., for $p=1$).

For the simulations, the initial distances (in km) of the drones 1–3 and 4 follow the distributions $0.95\mathcal{U}[6,10]+0.05\mathcal{U}[10,11]$ and $0.95\mathcal{U}[9,10]+0.05\mathcal{U}[10,11]$, respectively, where $\mathcal{U}$ denotes the uniform distribution. All velocities (in m/sec) follow the distribution $\mathcal{U}[50,50.5]$ and the deadline is set to $\tau=200$sec. The exact supports of the distributions are assumed known (but not the distributions themselves). Using this information, we selected a radius for the monolithic ball and the relative size of the hyperrectangle as exemplified above. Figure 3 shows the results across 30 realization of the simulations that leverage 100 samples each. The Wasserstein hyperrectangle exhibits superior performance compared to the monolithic ball, since the worst-case values are above the probability threshold (set at 0.45) in 83%

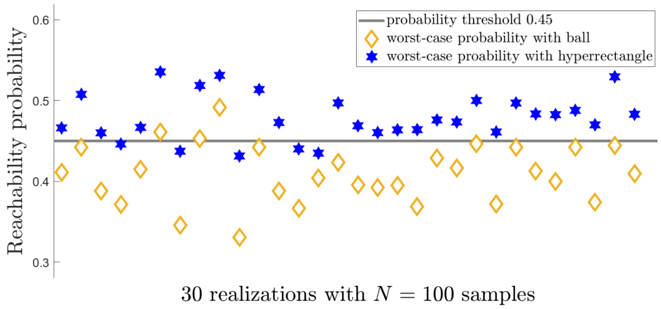

Fig. 3. The figure shows the lower probability bound (15) of reaching the region across 30 realizations. The results obtained by using the monolithic ball are depicted by the diamonds and the results with the hyperrectangle by the stars. In both cases, the ambiguity sets are built using 100 samples. The results obtained by the hyperrectangle outperform those obtained by the monolithic ball, since the lower probability bound is above the desired threshold (solid line) in considerably more occasions.

of the realizations in the former case compared to 10% in the latter. This improvement is also aided by the fact that the hyperrectangle can take advantage of the heterogeneity across the components of the random variable, which in this example is captured by the narrower support of drone 4.

VII. Conclusion

In this paper, we introduced structured ambiguity sets for data-driven DRO problems where the random variables have independent constituent components. These ambiguity sets are hyperrectangles in the Wasserstein space and can be tuned to contain the true distribution with prescribed confidence. We showed that Wasserstein hyperrectangles exhibit faster decay rates for high-dimensional data compared to monolithic ambiguity balls. We also obtained dual reformulations of the associated DRO problems for specific classes of cost functions. Future research includes the extension of the duality results to general cost functions and the development of tractable optimization algorithms to solve the reformulated DRO problems.

References

[1] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Review*, vol. 53, no. 3, p. 464–501, 2011.

[2] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, vol. 56, no. 3, pp. 830–857, 2019.

[3] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, vol. 44, no. 2, pp. 565–600, 2019.

[4] J. Blanchet, K. Murthy, and N. Si, "Confidence regions in Wasserstein distributionally robust estimation," *Biometrika*, 2021, DOI: https://doi.org/10.1093/biomet/asab026.

[5] D. Boskos, J. Cortés, and S. Martinez, "Data-driven ambiguity sets with probabilistic guarantees for dynamic processes," *IEEE Transactions on Automatic Control*, vol. 66, no. 7, pp. 2991–3006, 2021.

[6] D. Boskos, J. Cortés, and S. Martínez, "High-confidence data-driven ambiguity sets for time-varying linear systems," 2021, arXiv preprint arXiv:2102.01142.

[7] G. C. Calafiore and L. E. Ghaoui, "On distributionally robust chance-constrained linear programs," *Journal of Optimization Theory & Applications*, vol. 130, no. 1, pp. 1–22, 2006.

[8] P. Coppens and P. Patrinos, "Data-driven distributionally robust MPC for constrained stochastic systems," *IEEE Control Systems Letters*, vol. 6, pp. 1274–1279, 2022.

[9] J. Dedecker and F. Merlevède, "Behavior of the empirical Wasserstein distance in R^d under moment conditions," *Electronic Journal of Probability*, vol. 24, 2019.

[10] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, p. 595–612, 2010.

[11] S. Dereich, M. Scheutzow, and R. Schottstedt, "Constructive quantization: Approximation by empirical measures," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 49, no. 4, p. 1183–1203, 2013.

[12] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3-4, p. 707–738, 2015.

[13] R. Gao, "Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality," 2020, arXiv preprint arXiv:2009.04382.

[14] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[15] A. Hakobyan and I. Yang, "Wasserstein distributionally robust motion control for collision avoidance using conditional value-at-risk," *IEEE Transactions on Robotics*, 2021, DOI: 10.1109/TRO.2021.3106827.

[16] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming*, vol. 158, no. 1-2, p. 291–327, 2016.

[17] D. Kuhnl, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.

[18] D. Li, D. Fooladivanda, and S. Martínez, "Data-driven variable speed limit design for highways via distributionally robust optimization," in *European Control Conference*, Napoli, Italy, June 2019, pp. 1055–1061.

[19] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.

[20] B. P. G. V. Parys, D. Kuhn, P. J. Goulart, and M. Morar, "Distributionally robust control of constrained stochastic systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 2, pp. 430–442, 2015.

[21] G. Pflug and D. Wozabal, "Ambiguity in portfolio selection," *Quantitative Finance*, vol. 7, no. 4, pp. 435–442, 2007.

[22] B. K. Poolla, A. R. Hota, S. Bolognani, D. S. Callaway, and A. Cherukuri, "Wasserstein distributionally robust look-ahead economic dispatch," *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 2010–2022, 2020.

[23] I. Popescu, "Robust mean-covariance solutions for stochastic optimization," *Operations Research*, vol. 55, no. 1, pp. 98–112, 2007.

[24] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, "Regularization via mass transportation," *Journal of Machine Learning Research*, vol. 20, no. 103, pp. 1–68, 2019.

[25] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Advances in Neural Information Processing Systems*, 2018, pp. 8474–8483.

[26] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014, vol. 16.

[27] I. Tzortzis, C. D. Charalambous, and T. Charalambous, "Dynamic programming subject to total variation distance ambiguity," *SIAM Journal on Control and Optimization*, vol. 53, no. 4, pp. 2040–2075, 2015.

[28] I. Tzortzis, C. D. Charalambous, and C. N. Hadjicostis, "A distributionally robust LQR for systems with multiple uncertain players," in *IEEE Int. Conf. on Decision and Control*, 2021, pp. 3972–3977.

[29] C. Villani, *Optimal transport: old and new*. Springer, 2008, vol. 338.

[30] J. Weed and F. Bach, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *Bernoulli*, vol. 25, no. 4A, pp. 2620–2648, 2019.

[31] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, pp. 1358–1376, 12 2014.

[32] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3863–3870, 2021.