Distributionally Robust Optimization via Haar Wavelet Ambiguity Sets

Dimitris Boskos Jorge Cortés Sonia Martínez

Abstract— **This paper introduces a spectral parameterization of ambiguity sets to hedge against distributional uncertainty in stochastic optimization problems. We build an ambiguity set of probability densities around a histogram estimator, which is constructed by independent samples from the unknown distribution. The densities in the ambiguity set are determined by bounding the distance between the coefficients of their Haar wavelet expansion and the expansion of the histogram estimator. This representation facilitates the computation of expectations, leading to tractable minimax problems that are linear in the parameters of the ambiguity set, and enables the inclusion of additional constraints that can capture valuable prior information about the unknown distribution.**

I. Introduction

Uncertainty is ubiquitous across control engineering. Autonomous systems are deployed in unknown environments, sensor networks are subject to variable communication and measurement imperfections, and distributed energy resources are considerably affected by unpredictable weather fluctuations. Probabilistic models provide an expressive tool to quantify this uncertainty and make decisions that are optimal on average. A precise model for the underlying probability distributions may not always be available, so they are typically inferred by collected data. However small data sets and data that are corrupted by noise may lead to unreliable inferences. A strategy to circumvent this problem is to leverage distributionally robust optimization (DRO) formulations, which safeguard against the data variability generated by the stochastic models. DRO robustifies decisions by optimizing the worst-case cost over an *ambiguity set* of probability distributions that contains reasonable candidates for the true distribution. To this end, ambiguity sets that contain the true distribution with high confidence are to be built while excluding irrelevant distributions. Motivated by this, we develop spectral ambiguity sets, which account for prior information about the true distribution and facilitate the formulation of tractable DRO problems.

Literature review: DRO relies on distributional ambiguity sets to hedge against uncertainty about probabilistic models [34]. Ambiguity sets are typically based on moment constraints [32], [14], [10], statistical divergences [7], [1], [38], and optimal transport metrics like the Wasserstein distance [30], [3], [28], [18]. For data-driven problems, Wasserstein ambiguity balls have emerged as a popular choice. The reasons for this include that their size can be tuned by rigorous statistical guarantees [16], [4], they lead to tractable optimization problems, and they can accurately capture the effect of distribution variations on the optimization problems. Applications of Wasserstein ambiguity sets span from power systems [31], to distributed algorithms [8], machine learning [2], [24], traffic control [25], scheduling [23], and motion planning [21]. Dynamic aspects of uncertain distributions that are grouped through Wasserstein balls are also explored in [6] and in [5] for data corrupted by noise. There is an emerging interest to exploit DRO tools in stochastic control with distributional uncertainty, with contributions in linear quadratic regulator problems [10], model predictive control (MPC) [29], [35], [27], distributionally robust dynamic programming [41], [37], and purely data-driven predictive control [12]. DRO is also expected to play an important role in spatially distributed control problems subject to uncertainty, such as optimal sensor placement and coverage control [11].

Results on concentration inequalities for the convergence of empirical distributions in the Wasserstein distance enable the construction of ambiguity balls that contain the true distribution with prescribed probability [16], [39], [13]. These guarantees are especially convenient when solving multiple robust decision problems that are subject to the same uncertainty, such as in MPC [20]. Recent research also includes results where the radius of the ambiguity ball is informed by the optimization problem and exhibits favorable decay rates with the number of samples for problems with high-dimensional uncertainty [2], [33], [17].

These works revolve around approximating the probabilistic model through variations of the empirical distribution of the data, which is atomic and supported on a finite number of samples. On the other hand, in the majority of problems with probabilistic uncertainty, the distribution is characterized by a density. There is a plethora of methods to estimate densities of unknown distributions in nonparametric statistics [36]. Among them, wavelet density estimators are well known for their ability to capture local and heterogeneous effects of densities from general distribution classes [15], [22]. The recent work [40] establishes also new convergence results for wavelet density estimators in the Wasserstein distance.

Statement of contributions: In this paper we provide an alternative construction of ambiguity sets, which retain the benefit of establishing convergence in the Wasserstein distance and are accurately informed by natural prior assumptions for the probability distribution. The ambiguity sets are built around a linear Haar wavelet density estimator that is constructed from the collected samples. We capture the distributions in the set by considering densities whose

This work was partially supported by NSF CMMI-2044900 and ONR N00014-19-1-2471.

The authors are with the Delft Center for Systems and Control, TU Delft and the Department of Mechanical and Aerospace Engineering, University of California, San Diego, d.boskos@tudelft.nl and {cortes,soniamd}@ucsd.edu

wavelet coefficients vary up to some threshold from the coefficients of the estimator. This threshold can be selected to guarantee that the true distribution is properly close to some density from the ambiguity set with high probability. Due to space constraints the proofs are omitted and will appear elsewhere.

II. Preliminaries

Here we present general notation and concepts from probability theory and wavelets that will be used in the paper.

Notation: We denote by $\|\cdot\|_p$ the pth norm in $\mathbb{R}^n$ and omit the index in the Euclidean case $p = 2$. We use the notation $[n_1 : n_2]$ for the set of integers $\{n_1, n_1 + 1, \dots, n_2\} \subset \mathbb{N} \cup \{0\} =: \mathbb{N}_0$. Given $d \in \mathbb{N}$ and the index vector $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d) \in \mathbb{N}^d$, we denote $\mathbb{Z}_{\boldsymbol{\ell}} := \prod_{l=1}^d [0 : \ell_l]$. For an integrable function f on $\mathbb{R}^d$, we denote its L^1 norm as $\|f\|_{L^1} := \int_{\mathbb{R}^d} |f(x)|dx$.

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 probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. For any real number $p \geq 1$, $\mathcal{P}_p(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} \|x\|^p d\mu < \infty\}$ is the set of probability measures in $\mathcal{P}(\mathbb{R}^d)$ with finite pth moment. The Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is

$$W_p(\mu, \nu) := \Big(\inf_{\pi \in \mathcal{H}(\mu,\nu)} \Big\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \pi(dx, dy) \Big\} \Big)^{1/p},$$

where $\mathcal{H}(\mu, \nu)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, respectively. Given $B \subset \Omega$, $\mathbf{1}_B$ is the indicator function of B on Ω, with $\mathbf{1}_B(x) = 1$ for $x \in B$ and $\mathbf{1}_B(x) = 0$ for $x \notin B$.

Haar wavelets: Wavelets provide a well-established framework to approximate functions at varying resolution levels and have found tremendous success in signal processing [26], scientific computing [9], and statistics [22]. They enable multi-scale decompositions where functions are expanded as the sum of a coarse approximation and successive refinements of increasing detail. In these expansions, the wavelets capture the fluctuations of a function across the successive scales. Here, we focus exclusively on Haar wavelets to approximate functions on bounded rectangular domains, following the exposition in [9]. Throughout the paper, we use boldface to compactly denote vectors of indices and parameters. Consider the families of dyadic squares

$$I_{j,\boldsymbol{k}} := \prod_{l=1}^d [k_l 2^{-j}, (k_l + 1)2^{-j}),$$

in $\mathbb{R}^d$, where $j \in \mathbb{N}_0$, $\boldsymbol{k} := (k_1, \dots, k_d) \in \mathbb{Z}^d$, and let $\varphi := \mathbf{1}_{[0,1)}$ and $\psi := \mathbf{1}_{[0,1/2)} - \mathbf{1}_{[1/2,1)}$. Define

$$\varphi_{j,\boldsymbol{k}}(x) := 2^{dj/2} \varphi(2^j x_1 - k_1) \cdots \varphi(2^j x_d - k_d),$$

where $x := (x_1, \dots, x_d)$. The function $\varphi_{j,\mathbf{0}}$ is called the scaling function and we can equivalently define $\varphi_{j,\boldsymbol{k}} = 2^{dj/2} \mathbf{1}_{I_{j,\boldsymbol{k}}}$. Consider also the wavelets

$$\psi^r_{j,\boldsymbol{k}} \equiv \psi^{\boldsymbol{\epsilon}}_{j,\boldsymbol{k}}(x) := 2^{dj/2} \psi^{\epsilon_1}(2^j x_1 - k_1) \cdots \psi^{\epsilon_d}(2^j x_d - k_d),$$

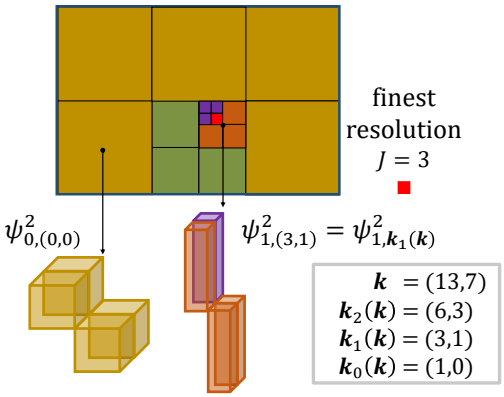

Fig. 1. The plot illustrates the squares that intersect the highest resolution square in red and their respective indices.

where $\boldsymbol{\epsilon} := (\epsilon_1, \dots, \epsilon_d) \in \{0, 1\}^d \setminus \mathbf{0}$, $r \in [1 : 2^d - 1]$ and $\psi^0 \equiv \varphi$, $\psi^1 \equiv \psi$ [9]. Consider the rectangular domain $Q_{\boldsymbol{\ell}} := \prod_{l=1}^d [0, \ell_l]$ with $\boldsymbol{\ell} := (\ell_1, \dots, \ell_d) \in \mathbb{N}^d$, a resolution index $J \in \mathbb{N}_0$ and let $2^J \boldsymbol{\ell} := (2^J \ell_1, \dots, 2^J \ell_d)$. The functions $\varphi_{J,\boldsymbol{k}}$, $\boldsymbol{k} \in \mathbb{Z}_{2^J \boldsymbol{\ell}}$ restricted to $Q_{\boldsymbol{\ell}}$ span the space

$$V_J^{\boldsymbol{\ell}} := \{f \in L^2(Q_{\boldsymbol{\ell}}) \mid f \text{ is constant on } I_{J,\boldsymbol{k}}, \boldsymbol{k} \in \mathbb{Z}_{2^J \boldsymbol{\ell}}\},$$

comprising of the functions that are constant at scale 2^{-J}. Alternatively, for any $0 \leq j_0 < J$, V_J is spanned by $\{\varphi_{j_0,\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}_{2^{j_0} \boldsymbol{\ell}}} \cup \{\psi^r_{j,\boldsymbol{k}}\}_{j_0 \leq j < J-1, \boldsymbol{k} \in \mathbb{Z}_{2^j \boldsymbol{\ell}}, r \in [1:2^d-1]}$, namely the basis $\{\varphi_{j_0,\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}_{2^{j_0} \boldsymbol{\ell}}}$ at resolution 2^{j_0} and the wavelets $\{\psi^r_{j,\boldsymbol{k}}\}_{j_0 \leq j < J-1, \boldsymbol{k} \in \mathbb{Z}_{2^j \boldsymbol{\ell}}, r \in [1:2^d-1]}$, that capture the fluctuations of the functions in V_J at the intermediate scales. For $j_0 = 0$, which is the case we will consider here, the wavelet basis $\Phi \cup \left(\cup_{j=0}^\infty \Psi_j \right)$ with

$$\Phi := \{\varphi_{j,\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}_{\boldsymbol{\ell}}} \tag{1a}$$

$$\Psi_j := \{\psi^r_{j,\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}_{2^j \boldsymbol{\ell}}, r \in [1:2^d-1]}, \tag{1b}$$

is the orthonormal Haar system on $Q_{\boldsymbol{\ell}}$ and spans $L^2(Q_{\boldsymbol{\ell}})$. We denote by Π_j the orthogonal projection in $L^2(Q_{\boldsymbol{\ell}})$ to the subspace $V_j^{\boldsymbol{\ell}}$ and by $D(V_j^{\boldsymbol{\ell}})$ the set of probability densities on $V_j^{\boldsymbol{\ell}}$. Each function $f \in L^2(Q_{\boldsymbol{\ell}})$ can be expressed as

$$f(x) = \sum_{\varphi \in \Phi} \alpha_\varphi \varphi(x) + \sum_{j=0}^\infty \sum_{\psi \in \Psi_j} \beta_\psi \psi(x).$$

When $f \in V_J^{\boldsymbol{\ell}}$, its constant value at each fine-grained interval $I_{J,\boldsymbol{k}}$, $\boldsymbol{k} \in \mathbb{Z}_{2^J \boldsymbol{\ell}}$ is evaluated through its nonzero wavelet coefficients as

$$f|_{I_{J,\boldsymbol{k}}} = \alpha_{\boldsymbol{k}_0(\boldsymbol{k})} + \sum_{j=0}^{J-1} \sum_{r=1}^{2^d-1} 2^{dj/2} \beta^r_{j,\boldsymbol{k}_j(\boldsymbol{k})} \times \operatorname{sign}^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}(\boldsymbol{k}_{j+1}(\boldsymbol{k})). \tag{2}$$

In (2), $\boldsymbol{k}_j(\boldsymbol{k})$ are the indices of the unique 2^{-j}-resolution square that intersects $I_{J,\boldsymbol{k}}$, and $\operatorname{sign}^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}(\boldsymbol{k}_{j+1}(\boldsymbol{k}))$ is the sign of the wavelet $\psi^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}$ on the square $I_{j,\boldsymbol{k}_{j+1}(\boldsymbol{k})}$, which takes values in $\{-2^{dj/2}, 2^{dj/2}\}$ (cf. Figure 1).

III. Problem formulation

Stochastic optimization is focused on taking optimal decisions in problems affected by uncertainty. A typical instance of a stochastic optimization problem takes the form

$$\min_{u\in\mathcal{U}} \mathbb{E}[g(u,X)] \equiv \min_{u\in\mathcal{U}} \int_Q g(u,x)\rho(x)dx, \tag{3}$$

where the objective function g depends on the decision variable $u \in \mathcal{U} \subset \mathbb{R}^n$ and the random variable X. In this paper we assume that X has a density ρ that is supported on the compact rectangular set $Q \subset \mathbb{R}^d$, and hence, the expected cost is expressed as in the right-hand side of (3). Without loss of generality we henceforth assume that $Q \equiv Q_{\boldsymbol{\ell}} := \prod_{l=1}^d [0,\ell_l]$ as in the bounded domain of Section II.

We consider the case when the density ρ is not known and we only have access to N i.i.d. samples $X_1,\dots,X_N$ from it. To hedge against the lack of information about the density due to the finite number of available samples, we follow the distributionally robust optimization (DRO) paradigm. Instead of (3), we set out to solve

$$\min_{u\in\mathcal{U}} \max_{\rho'\in\mathcal{P}} \int_{Q_{\boldsymbol{\ell}}} g(u,x)\rho'(x)dx, \tag{4}$$

over an ambiguity set $\mathcal{P}$ of distributions. This set should be large enough to contain the unknown distribution ρ with high probability, yet as small as possible to avoid overconservative optimizers. We also want to inform the ambiguity set by prior assumptions about the density, such as the following one.

Assumption 3.1: ***(Upper and lower density bounds).*** There exist functions ρ_{low} and ρ_{up} with

$$0 \le \rho_{\text{low}}(x) \le \rho(x) \le \rho_{\text{up}}(x) \quad \forall x \in Q_{\boldsymbol{\ell}}. \tag{5}$$

Both functions are measurable and may take the value ∞.

Such upper bounds can for instance be motivated by assumptions like "$A \subset Q_{\boldsymbol{\ell}} \subset \mathbb{R}^2$ contains up to 0.4 of the probability mass and the probability of sampling any point in A is no more than twice any other point therein", which would result in the upper density bound $\rho_{\text{up}}(x) = 0.8/\text{area}(A)$ for all $x \in A$.

Problem formulation: Construct an ambiguity set of probability densities that explicitly takes into account Assumption 3.1, contains the true density with prescribed probability, and enables the derivation of tractable DRO problems.

Despite the benefits of Wasserstein ambiguity sets, cf. Section I, they contain atomic distributions, like the empirical distribution of the samples, which are clearly not densities. Most important, it is not straightforward how they can effectively capture constraints of the form (5). To address these issues, here we introduce wavelet-based ambiguity sets and capture the densities that they contain through the variation of their wavelet coefficients.

IV. Wavelet estimator ambiguity sets

In this section we use a wavelet estimator to build an ambiguity set of probability densities for the true density ρ. As ρ is supported on the domain $Q_{\boldsymbol{\ell}}$, it can be expressed as

$$\rho(x) = \sum_{\varphi\in\Phi} \alpha_\varphi \varphi(x) + \sum_{j=0}^{\infty} \sum_{\psi\in\Psi_j} \beta_\psi \psi(x),$$

with the Haar wavelet basis $\Phi \cup \{\Psi_j\}_{j=0}^{\infty}$ given in Section II. To infer a data-driven model of this unknown density using the N independent samples $X_1,\dots,X_N$, we select a resolution threshold 2^{-J} and build the wavelet density estimator

$$\widehat{\rho}(x) = \sum_{\varphi\in\Phi} \widehat{\alpha}_\varphi \varphi(x) + \sum_{j=0}^{J-1} \sum_{\psi\in\Psi_j} \widehat{\beta}_\psi \psi(x),$$

with

$$\widehat{\alpha}_\varphi := \frac{1}{N}\sum_{i=1}^N \varphi(X_i), \qquad \varphi \in \Phi, \tag{6a}$$

$$\widehat{\beta}_\psi := \frac{1}{N}\sum_{i=1}^N \psi(X_i), \qquad \psi \in \cup_{j=0}^{J-1}\Psi_j. \tag{6b}$$

Note that the estimator $\widehat{\rho}$ is equal to the histogram

$$h(x) := \frac{1}{N}\sum_{i=1}^N 2^{dJ/2}\mathbf{1}_{I_{J,\boldsymbol{k}}}(X_i), x \in I_{J,\boldsymbol{k}}, \boldsymbol{k} \in \mathbb{Z}_{2^J\boldsymbol{\ell}}.$$

To define the ambiguity set, we consider all densities in $V_J^{\boldsymbol{\ell}}$ whose wavelet coefficients are within prescribed bounds from the coefficients of the estimator. If the true distribution were also an element of $V_J^{\boldsymbol{\ell}}$, our goal would be to establish that it belongs to the ambiguity set with high probability. As we are choosing a specific resolution and ρ may lie outside $V_J^{\boldsymbol{\ell}}$, the best guarantee that we can have is that its projection will be in the ambiguity set with prescribed probability. To capture the multi-scale property of wavelet bases, we provide separate bounds for the coefficient discrepancies across different scales. We will use the compact notation $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$ for the coefficients of the scaling functions and the wavelets at each scale j, and $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}_j$ for the corresponding coefficients of the estimator. Given $s \in [1,2]$ and the radii $\boldsymbol{\varepsilon} = (\varepsilon_0,\dots,\varepsilon_J)$, the ambiguity set is determined through the wavelet coefficients $(\boldsymbol{\alpha},\boldsymbol{\beta}_0,\dots,\boldsymbol{\beta}_{J-1}) \in \mathbb{R}^K$, $K := 2^{dJ}\prod_{l=1}^d \ell_l$, that satisfy

$$\|\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}\|_s^s \le \varepsilon_0,\ \|\boldsymbol{\beta}_j-\widehat{\boldsymbol{\beta}}_j\|_s^s \le \varepsilon_{j+1},\ j \in [0:J-1] \tag{7}$$

and the following constraint:

- *Unit mass.* Each density from the ambiguity set should integrate to one. Equivalently, the coefficients $\alpha_{\boldsymbol{k}}$ of the scaling functions need to satisfy

$$\sum_{\boldsymbol{k}\in\mathbb{Z}_{\boldsymbol{\ell}}} \alpha_{\boldsymbol{k}} = 1. \tag{8}$$

Additionally, we consider either of the following constraints:

- *Nonnegative densities.* As they are constant at resolution 2^{-J}, their coefficients need to satisfy the constraint

$$\alpha_{\boldsymbol{k}_0(\boldsymbol{k})} + \sum_{j=0}^{J-1}\sum_{r=1}^{2^d-1} 2^{dj/2}\beta^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}$$

$$\times \operatorname{sign}^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}(\boldsymbol{k}_{j+1}(\boldsymbol{k})) \geq 0 \quad \forall \boldsymbol{k} \in \mathbb{Z}_{2^J\boldsymbol{\ell}}, \tag{9a}$$

with $\boldsymbol{k}_j(\boldsymbol{k})$ and $\operatorname{sign}^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}$ as in (2).

- *Upper and lower density bounds.* The true density should satisfy the bounds of Assumption 3.1. These are captured by the alternative set of linear constraints

$$\begin{aligned}
&\min_{x\in I_{J,\boldsymbol{k}}} \rho_{\text{low}}(x) \\
&\leq \alpha_{\boldsymbol{k}_0(\boldsymbol{k})} + \sum_{j=0}^{J-1}\sum_{r=1}^{2^d-1} 2^{dj/2}\beta^r_{j,\boldsymbol{k}_j(\boldsymbol{k})} \operatorname{sign}^r_{j,\boldsymbol{k}_j(\boldsymbol{k})}(\boldsymbol{k}_{j+1}(\boldsymbol{k})) \\
&\leq \max_{x\in I_{J,\boldsymbol{k}}} \rho_{\text{up}}(x) \quad \forall \boldsymbol{k} \in \mathbb{Z}_{2^J\boldsymbol{\ell}}.
\end{aligned} \tag{9b}$$

The constraints (9) are facilitated by the choice of Haar wavelets, which provide piecewise constant approximations of the density. In particular, the piecewise constant approximation of a density that satisfies the constraints of Assumption 3.1 will also satisfy the constraints (9b), which can be checked at a finite number of representative points. We succinctly denote the ambiguity set as

$$\begin{aligned}
\mathcal{P} := \Big\{ \rho' \in \mathcal{D}(V_J^{\boldsymbol{\ell}}) \,|\, \rho' = \sum_{\varphi\in\Phi} \alpha_\varphi \varphi + \sum_{j=0}^{J-1}\sum_{\psi\in\Psi_j} \beta_\psi \psi \text{ and} \\
(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{J-1}) \text{ satisfy (7), (8), and (9a) or (9b)} \Big\}.
\end{aligned}$$

We refer to thresholds $\varepsilon_0, \ldots, \varepsilon_J$ in (7) as the *ambiguity radii*.

Remark 4.1: ***(Constraint feasibility).*** The constraints in (9b) may not be feasible for the wavelet coefficients of the density estimator. As a consequence, they may also not be feasible for any distribution of the ambiguity set if its radii are not sufficiently large. We later focus on ambiguity sets where the radii take the form $\boldsymbol{\varepsilon} = r^\star(c_0, \ldots, c_J)$ for some $r^\star > 0$ and fixed constants c_j. In this case, we can check constraint feasibility by solving the convex optimization problem

$$\begin{aligned}
&\min\ r \\
&\text{s.t. } \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_s^s \leq r c_0 \\
&\quad \|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\|_s^s \leq r c_{j+1} \quad j \in [0:J-1] \\
&\quad (8),\ (9b).
\end{aligned}$$

If $r < r^\star$, then the constraint is feasible. Otherwise, larger radii must be selected. □

The ambiguity set facilitates the formulation of DRO problems that are linear in the wavelet coefficients. We use the compact notation $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{J-1})$ to denote the Haar coefficients of a distribution in $V_J^{\boldsymbol{\ell}}$ and let Θ be the set of parameters $\boldsymbol{\theta}$ which satisfy the constraints (7), (8), and either (9a) or (9b). Note that $\boldsymbol{\theta} \in \Theta$ parameterizes the ambiguity set. As a consequence, the DRO problem (4) is equivalently written

$$\min_{u\in\mathcal{U}} \max_{\boldsymbol{\theta}\in\Theta} \int_{Q_{\boldsymbol{\ell}}} g(u,x)\rho_{\boldsymbol{\theta}}(x)dx, \tag{10}$$

with $\rho_{\boldsymbol{\theta}}$ the parameterized distributions. The following result establishes a *linear in* $\boldsymbol{\theta}$ reformulation of the DRO problem.

Proposition 4.2: ***(DRO reformulation).*** Let $\boldsymbol{g}(u)$ be the vector comprising of the integrals $\int_{Q_{\boldsymbol{\ell}}} g(u,x)\varphi(x)dx$, $\int_{Q_{\boldsymbol{\ell}}} g(u,x)\psi(x)dx$ ordered as the parameters from $\boldsymbol{\theta}$ associated to the φ's and ψ's. Then the DRO problem (10) admits the reformulation

$$\min_{u\in\mathcal{U}} \max_{\boldsymbol{\theta}\in\Theta} \boldsymbol{\theta}^\top \boldsymbol{g}(u). \tag{11}$$

This parameterizaton of the optimization problem is in practice suitable for problems with low dimensional data since the number of parameters scales exponentially with the dimension d and the finest resolution level J.

V. Probabilistic guarantees

Here we study how to tune the radii of the ambiguity set so that it contains the projection of the true density to $V_J^{\boldsymbol{\ell}}$ with prescribed probability. Our approach draws from the recent work [40], which establishes expected value norm-bounds for the discrepancy vectors between the wavelet coefficients of the estimator and the density across scales. Here we sharpen these expected value bounds by leveraging the specific properties of the Haar system, and exploit them to further obtain concentration results for the coefficients of the true density.

Henceforth, we also denote as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$ the wavelet coefficients of the unknown density ρ. We obtain concentration bounds for the sums $\|\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}\|_s^s$ and $\|\boldsymbol{\beta}_j-\widehat{\boldsymbol{\beta}}_j\|_s^s$, $j \in [0:J-1]$, where $s \in [1,2]$. A reason to consider this spectrum of exponents is that we get close probabilistic guarantees for each of them, which provides the design flexibility to pick a convenient s for the definition of the ambiguity set. To clarify the benefits of this flexibility, note that since e.g., $\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_s \leq \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_1$ for all $s \in [1,2]$, an ambiguity set for $s = 1$ is also an ambiguity set for $s \in (1,2]$. However, the associated constraint set $\{\boldsymbol{\alpha} \in \mathbb{R}^{\prod_{l=1}^d \ell_l} \,|\, \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_1 \leq \varepsilon_0\}$, is not uniformly convex. We can therefore balance both requirements by exploiting the option to pick some other s from $(1,2]$. We next provide bounds for the expected values of the wavelet coefficient discrepancies at each scale.

Proposition 5.1: ***(Expected discrepancies of wavelet coefficients).*** For each $s \in [1,2]$, the expected discrepancies from the empirical wavelet coefficients satisfy

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_s^s) &\leq C_0(s)\frac{1}{N^{s/2}}, \\
\mathbb{E}(\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\|_s^s) &\leq C_{j+1}(s)\frac{1}{N^{s/2}}, \quad j \in [0:J-1],
\end{aligned}$$

with

$$C_0(s) := \left(\frac{K_0-1}{K_0}\right)^{\frac{s}{2}}, \quad C_j(s) := (2^d-1)^{s/2}2^{sd(j-1)/2}, \quad j \in [1:J]. \tag{12}$$

Using this result we derive concentration inequalities for the discrepancy vectors between the true and the empirical wavelet coefficients.

Theorem 5.2: ***(Coefficient concentration).*** Let

$$\begin{aligned}
f_0(X_1, \ldots, X_N) &:= \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_s^s \\
f_j(X_1, \ldots, X_N) &:= \|\boldsymbol{\beta}_{j-1} - \widehat{\boldsymbol{\beta}}_{j-1}\|_s^s,
\end{aligned}$$

for $j \in [1:J]$. Then for any confidence $1-\delta$, $\delta \in (0,1)$,

$$\mathbb{P}(f_j \le \varepsilon_j) \ge 1-\delta, \tag{13a}$$

$$\varepsilon_j \equiv \varepsilon_j(\delta) := C_j \frac{1}{N^{s/2}} + C'_j \left(2\ln\frac{1}{\delta}\right)^{\frac{1}{2}} \frac{1}{N^{1/2}}, \tag{13b}$$

for $j \in [0:J]$, with C_j as given in (12) and

$$C'_0 := s, \qquad C'_j := (2^d-1)s2^{s-1+sd(j-1)/2}. \tag{14}$$

From Theorem 5.2, we obtain the following result, which simultaneously bounds the coefficient discrepancies at all scales for any prescribed confidence.

Corollary 5.3: ***(Simultaneous guarantees across scales).*** Consider a confidence $1-\delta$, $\delta \in (0,1)$, the maximum resolution 2^{-J}, and let $\delta' \equiv \delta'(\delta) := \delta/(J+1)$. Then

$$\mathbb{P}(\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_s^s \le \varepsilon_0, \\ \|\boldsymbol{\beta}_{j-1} - \widehat{\boldsymbol{\beta}}_{j-1}\|_s^s \le \varepsilon_j, j \in [1:J]) \ge 1-\delta,$$

with $\varepsilon_j \equiv \varepsilon_j(\delta')$, $j \in [0:J]$ as given by (13b). In addition, we may choose the uniformly dilated ambiguity radii

$$\boldsymbol{\varepsilon} := \frac{1}{N^{1/2}}(C_0^\star, \ldots, C_J^\star), \quad C_j^\star := C_j + C'_j\left(2\ln\frac{1}{\delta'}\right)^{\frac{1}{2}},$$

and obtain the same guarantees.

Remark 5.4: ***(Tuning of the ambiguity radii).*** The concentration results of Theorem 5.2 and Corollary 5.3 may overestimate the size of the ambiguity radii, but they still provide a useful indicator about the relative size of the ambiguity set across different scales. Further, the uniformly dilated radii of Corollary 5.3 can be used to check the feasibility of additional density constraints by solving the convex optimization problem in Remark 4.1. □

VI. Approximation analysis

In this section, we show that the ambiguity sets contain a distribution which is close to the true one in the Wasserstein distance with prescribed probability. This closeness level can be made arbitrarily small by selecting an appropriate maximum resolution level for the wavelet decomposition. To show this, we exploit the following result, which quantifies the worst-case Wasserstein distance between a probability density on $Q_{\boldsymbol{\ell}}$ and its projection to $V_J^{\boldsymbol{\ell}}$ in terms of the resolution 2^{-J}. The result follows closely the proof of [40, Proposition 5].

Proposition 6.1: ***(Concentration in the Wasserstein distance).*** Let $X_1, \ldots, X_N$ be N samples of the unknown density ρ. Consider a confidence level $1-\delta$, the maximum resolution 2^{-J}, and let $\delta' \equiv \delta'(\delta) := \delta/(J+1)$. Then the pth Wasserstein distance between the distributions μ_ρ and $\mu_{\widehat{\rho}}$ of the true density and the estimator $\widehat{\rho}$, resp., satisfies

$$\mathbb{P}(W_p^p(\mu_\rho, \mu_{\widehat{\rho}}) \le \varepsilon(\delta')) \ge 1-\delta,$$

where

$$\varepsilon(\delta') := 2^{-Jp} + \left[(\|\boldsymbol{\ell}\|+2)\mathfrak{C}(\delta') + 2J\mathfrak{C}_d(\delta')\right]\frac{1}{N^{1/2}},$$

$$\mathfrak{C}(\delta') := 1 + \left(2\ln\frac{1}{\delta'}\right)^{\frac{1}{2}}, \mathfrak{C}_d(\delta') := 2^{d/2} + \left(2\ln\frac{1}{\delta'}\right)^{\frac{1}{2}} 2^d.$$

VII. Roadside assistance placement

To illustrate our results, we consider the problem of optimally placing a roadside assistance station across a long road segment. The probability of a car needing assistance at each road location is unknown and inferred by a limited amount of historic data. In addition, the cost of offering assistance grows quadratically with the distance between the car and the station. Modeling the road segment by the interval $[0,\ell]$, we seek to minimize the expected cost of assisting a car. Since the probability distribution of this event is unknown, we are interested in solving the DRO problem

$$\min_{u\in[0,\ell]} \max_{\rho'\in\mathcal{P}} \int_{[0,\ell]} |u-x|^2 \rho'(x)dx,$$

where $\mathcal{P}$ denotes the wavelet-based ambiguity set described in Section IV.

We take $\ell = 4$ and the resolution level $J = 2$. The unknown density is uniformly distributed across the left and right half of the lane with probabilities 0.4 and 0.6, respectively. We also have the prior information that the density at the left part of the lane $[0,2]$ does not exceed a uniform density with total mass 0.45. This prior assumption determines the upper bound of the constraint (9b) when building the ambiguity set. For each simulation, we take $N = 100$ samples from the true density ρ and repeat the random experiment 20 times. Each time, we solve the sample average approximation (SAA) and the linear in the wavelet coefficients reformulation of the DRO problem in (11). We tune the ambiguity radius of the DRO problem based on Corollary 5.3 and check the initial feasibility of all the constraints as in Remark 4.1. To solve the resulting saddle point problem (11), we use the Frank-Wolfe algorithm from [19].

The outcome of the simulations is plotted in Figure 2. The plot depicts the expected costs of the true density when using the optimizers obtained from the DRO problem and the same expected costs when using the optimizer from the SAA. The expected costs using the DRO optimizers have a smaller variability and are on average considerably closer to the value obtained using the true optimizer. This improved performance is aided by the fact that the ambiguity set is informed by prior assumptions about the true distribution. Thus, when samples appear on the left side of the road more frequently than what the true distribution dictates, the ambiguity set constraints increase distribution variability on the right side and improve the DRO optimizer.

VIII. Conclusions

We have provided a spectral ambiguity set characterization by exploiting Haar wavelet expansions for use in distributionally robust optimization (DRO). The ambiguity sets conveniently capture prior information about the unknown distribution and enable the formulation of tractable DRO problems that are linear in the wavelet coefficients of the ambiguous distributions. Our construction is also accompanied by probabilistic guarantees of containing the true distribution. We have shown that each ambiguity set is contained in a Wasserstein ball, whose radius can be made

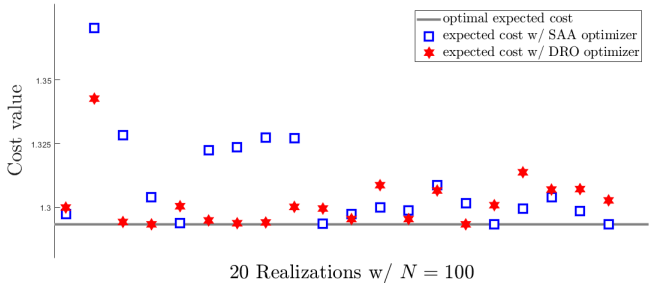

Fig. 2. The plot shows the optimal expected cost and the data sets of the true cost with the DRO optimizer and the SAA optimizer, respectively, across the 20 realizations. The average true cost using the DRO optimizer is clearly closer to the desired optimal expected cost.

arbitrarily small by adjusting the number of samples and the resolution accuracy of the wavelet expansion. Future research will include the generalization of the results for different wavelets and incorporate regularity properties of the densities. We will also study how to exploit sparser wavelet representation while retaining probabilistic guarantees.

REFERENCES

[1] A. Ben-Tal, D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, p. 341–357, 2013.

[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 V. A. Nguyen, "Statistical analysis of Wasserstein distributionally robust estimators," in *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 2021, pp. 227–254.

[5] D. Boskos, J. Cortés, and S. Martinez, "Data-driven ambiguity sets for linear systems under disturbances and noisy observations," in *American Control Conference*, Denver, CO, Jul. 2020, pp. 4491–4496.

[6] ——, "Data-driven ambiguity sets with probabilistic guarantees for dynamic processes," *IEEE Transactions on Automatic Control*, vol. 66, no. 7, pp. 2991–3006, 2021.

[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] A. Cherukuri and J. Cortés, "Cooperative data-driven distributionally robust optimization," *IEEE Transactions on Automatic Control*, vol. 65, no. 10, pp. 4400–4407, 2020.

[9] A. Cohen, *Numerical analysis of wavelet methods*. Elsevier, 2003.

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

[11] J. Cortés, S. Martinez, T. Karatas, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 2, pp. 243–255, 2004.

[12] J. Coulson, J. Lygeros, and F. Dörfler, "Regularized and distributionally robust data-enabled predictive control," in *IEEE Int. Conf. on Decision and Control*, Nice, France, December 2019, pp. 2696–2701.

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

[14] 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.

[15] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[16] 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.

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

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

[19] G. Gidel, T. Jebara, and S. Lacoste-Julien, "Frank-Wolfe algorithms for saddle point problems," in *Artificial Intelligence and Statistics*, 2017, pp. 362–371.

[20] A. Hakobyan, G. C. Kim, and I. Yang, "Risk-aware motion planning and control using CVaR-constrained optimization," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3924–3931, 2019.

[21] A. Hakobyan and I. Yang, "Wasserstein distributionally robust motion planning and control with safety constraints using conditional value-at-risk," *IEEE Int. Conf. on Robotics and Automation*, pp. 490–496, 2020.

[22] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, approximation, and statistical applications*. Springer, 1998, vol. 129.

[23] R. Jiang, M. Ryu, and G. Xu, "Data-driven distributionally robust appointment scheduling over Wasserstein balls," *arXiv preprint arXiv:1907.03219*, 2019.

[24] D. Kuhn, 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.

[25] 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.

[26] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 2009.

[27] C. Mark and S. Liu, "Stochastic MPC with distributionally robust chance constraints," *IFAC Papers Online*, vol. 53, no. 2, pp. 7136–7141, 2020.

[28] 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.

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

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

[31] B. K. Poolla, A. R. Hota, S. Bolognani, D. S. Callaway, and A. Cherukuri, "Wasserstein distributionally robust look-ahead economic dispatch," *arXiv preprint arXiv:2003.04874*, 2020.

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

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

[34] A. Shapiro, "Distributionally robust stochastic programming," *SIAM Journal on Optimization*, vol. 27, no. 4, p. 2258–2275, 2017.

[35] P. Sopasakis, D. Herceg, A. Bemporad, and P. Patrinos, "Risk-averse model predictive control," *Automatica*, vol. 100, pp. 281–288, 2019.

[36] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer, 2009.

[37] 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.

[38] Z. Wang, P. W. Glynn, and Y. Ye, "Likelihood robust optimization for data-driven problems," *Computational Management Science*, vol. 13, no. 2, pp. 241–261, 2016.

[39] 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.

[40] J. Weed and Q. Berthet, "Estimation of smooth densities in Wasserstein distance," in *Conference on Learning Theory*, 2019, pp. 3118–3119.

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