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

Data-Driven Distributionally Robust Coverage Control by Mobile Robots

Dimitris Boskos Jorge Cortés Sonia Martínez

Abstract— **This paper provides a data-driven solution to the problem of coverage control by which a team of robots aims to optimally deploy in a spatial region where certain event of interest may occur. This event is random and described by a probability density function, which is unknown and can only be learned by collecting data. In this work, we hedge against this uncertainty by designing a distributionally robust algorithm that optimizes the locations of the robots against the worst-case probability density from an ambiguity set. This ambiguity set is constructed from data initially collected by the agents, and contains the true density function with prescribed confidence. However, the objective function that the robots seek to minimize is non-smooth. To address this issue, we employ the so-called gradient sampling algorithm, which approximates the Clarke generalized gradient by sampling the derivative of the objective function at nearby locations and stabilizes the choice of descent directions around points where the function may fail to be differentiable. This enables us to prove that the algorithm converges to a stationary point from any initial location of the robots, in analogy to the well-known Lloyd algorithm for differentiable costs when the spatial density is known.**

I. Introduction

The deployment of multi-robot systems in realistic environments will enable the realization of multiple tasks in a variety of scenarios. Yet, this requires overcoming prominent challenges; such as that of providing these systems with the capability of operating in unknown environments.

Take the paradigmatic example of multi-robot coverage control [5], by which a team of robots aims to navigate to locations that can provide e.g. optimal assistance in emergency situations. Here, the optimal locations maximize an expected utility of coverage with respect to a spatial density function. Ideally, the task is to be accomplished by robots by means of distributed algorithms. More critically, the utility depends on a model of the environmental events, which is typically not available to robots.

Our goal in this paper is to obtain a data-driven and optimal coverage control algorithm with rigorous guarantees for unknown spatial density functions. To do this, robots have access to a finite and independent collection of samples taken from it. The guarantees that are sought are both of performance—optimal positions via sampling should provide quality solutions with respect to the true coverage control objective with quantifiable guarantees—and of convergence—the algorithm should be stable.

Literature review: There has been extensive work on coverage control during the last two decades. A continuous-time distributed Lloyd algorithm was introduced in [11], which also established convergence of the algorithm to stationary points using tools from the theory of dynamical systems. Generalizations of this algorithm were derived in [10], which considers limited interactions between the robots, and [9], which provides distributed coordination protocols to solve non-smooth locational optimization problems. Other extensions of coverage control include the consideration of visibility constraints [17] and time-varying densities [13]. Further results also consider inference tools to progressively learn the spatial density in a data-driven manner using basis functions [24] or interpolation methods [20]. However, these works cannot directly handle uncertainty about the unknown spatial density, which may not be correctly inferred when the amount of data is not sufficiently large.

To hedge against distributional uncertainty in stochastic decision making, distributionally robust optimization (DRO) makes use of ambiguity sets of probability distributions that contain multiple candidate models of the unknown uncertainty [12], [21]. This enables the designer to make robust decisions, which guarantees that optimality is not significantly jeopardized when the uncertainty model turns out to be the worst-case element from the ambiguity set. The distributions in an ambiguity set are typically grouped using moment constraints [22], relative entropy constraints [1], and optimal transport metrics [3], [15] a.k.a. Wasserstein distances. The latter have emerged as a popular choice for data-driven problems. Among the reasons for this is that they enjoy statistical guarantees of containing the true probabilistic model [14] or an appropriate replacement of it [2] with prescribed confidence. Typically, Wasserstein ambiguity sets are centered at the empirical distribution of collected data and therefore contain several distributions that are not absolutely continuous with respect to the Lebesgue measure. Nevertheless, in [4], we leverage results from wavelet density estimation [25] to build ambiguity sets that only contain densities and control their size in terms of the maximum optimal transport discrepancy between their members.

Statement of contributions: Our main contribution is the development of a distributionally robust coverage control algorithm for an *unknown spatial density*, which is data-driven and enjoys rigorous optimality guarantees. To achieve this result, we exploit Haar-wavelet ambiguity sets, which are data-driven, contain only densities, and have the further benefit of containing an appropriate approximation of the true distribution with high probability. To build the ambiguity sets, we first construct a wavelet density estimator from the collected samples and then consider all densities whose wavelet coefficients are sufficiently close to those of the

This work was partially supported by ONR Award N00014-23-1-2353.

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

estimator. Since the resulting optimization problem is non-smooth, we design a variant of the so-called gradient sampling algorithm, which guarantees convergence to a Clarke stationary point. To this end, our second contribution is the generalization of the gradient sampling algorithm to nested optimization problems, where the objective function does not have a closed-form formula and is only approximated by the solution of an inner optimization problem. In particular, we prove that our modified gradient sampling algorithm shares the same convergence guarantees as the original one. Due to space constraints, the proofs are omitted and will appear elsewhere.

II. Preliminaries on Haar wavelets and non-smooth analysis

We denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^n$. We use the notation $[n_1 : n_2]$ for the set of integers $\{n_1, n_1+1, \dots, n_2\} \subset \mathbb{N} \cup \{0\} =: \mathbb{N}_0$ and denote $\bar{\mathbb{R}}_{\geq 0} := \mathbb{R}_{\geq 0} \cup \{+\infty\}$. Given $d \in \mathbb{N}$ and the index vector $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d) \in \mathbb{N}^d$, we denote $\mathbb{Z}_{\boldsymbol{\ell}} := \prod_{l=1}^{d} [0 : \ell_l]$. We denote by $B(x, \varepsilon)$ the ball with center $x \in \mathbb{R}^n$ and radius $\varepsilon > 0$. Given two sets $A, B \subset \mathbb{R}^n$ we denote the smallest distance between their elements by $\operatorname{dist}(A, B) := \inf\{\|x - y\| \,|\, x \in A, y \in B\}$ and also use the notation $\operatorname{dist}(x, A) := \operatorname{dist}(\{x\}, A)$ when considering single-element sets. The convex hull and closure of a set $A \subset \mathbb{R}^n$ are denoted by $\operatorname{conv}(A)$ and $\operatorname{cl}(A)$, respectively. Vectors will be interpreted as column vectors in linear algebra operations unless indicated by a transpose.

Haar wavelets: Wavelets are used to construct bases of function spaces that are suitable to approximate functions at varying resolution levels. Here, we consider Haar wavelets on $\mathbb{R}^2$ to approximate functions on bounded rectangular domains, following the exposition in [8]. Throughout the paper, we use boldface to compactly denote vectors of indices and parameters. Consider the families of dyadic squares

$$I_{j,\boldsymbol{k}} := [k_1 2^{-j}, (k_1+1)2^{-j}) \times [k_2 2^{-j}, (k_2+1)2^{-j}),$$

in $\mathbb{R}^2$, where $j \in \mathbb{N}_0$, $\boldsymbol{k} := (k_1, k_2) \in \mathbb{Z}^2$, and let $\varphi := \mathbf{1}_{[0,1)}$ and $\psi := \mathbf{1}_{[0,1/2)} - \mathbf{1}_{[1/2,1)}$. Define

$$\varphi_{j,\boldsymbol{k}}(x) := 2^j \varphi(2^j x_1 - k_1)\varphi(2^j x_2 - k_2),$$

where $x := (x_1, x_2)$. The function $\varphi_{j,\mathbf{0}}$ is called the scaling function and we can equivalently define $\varphi_{j,\boldsymbol{k}} = 2^j \mathbf{1}_{I_{j,\boldsymbol{k}}}$. Consider also the wavelets

$$\psi^r_{j,\boldsymbol{k}} \equiv \psi^{\boldsymbol{\epsilon}}_{j,\boldsymbol{k}}(x) := 2^j \psi^{\epsilon_1}(2^j x_1 - k_1)\psi^{\epsilon_2}(2^j x_2 - k_2),$$

where $\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2) \in \{0,1\}^2 \setminus \mathbf{0}$, $r \in [1:3]$ and $\psi^0 \equiv \varphi$, $\psi^1 \equiv \psi$. Figure 1 shows the scaling function and wavelets for $j = 0$.

Consider the rectangular domain $Q_{\boldsymbol{\ell}} := [0, \ell_1] \times [0, \ell_2]$ with $\boldsymbol{\ell} := (\ell_1, \ell_2) \in \mathbb{N}^2$, a resolution index $J \in \mathbb{N}_0$ and let $2^J \boldsymbol{\ell} := (2^J \ell_1, 2^J \ell_2)$. The functions $\{\varphi_{0,\boldsymbol{k}}\}_{k \in \mathbb{Z}_{\boldsymbol{\ell}}} \cup \{\psi^r_{j,\boldsymbol{k}}\}_{0 \leq j < J-1, \boldsymbol{k} \in \mathbb{Z}_{2^j \boldsymbol{\ell}}, r \in [1:3]}$, span the space

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

comprising of the functions that are constant at scale 2^{-J}. Namely, $V_J^{\boldsymbol{\ell}}$ is spanned by the scaling functions

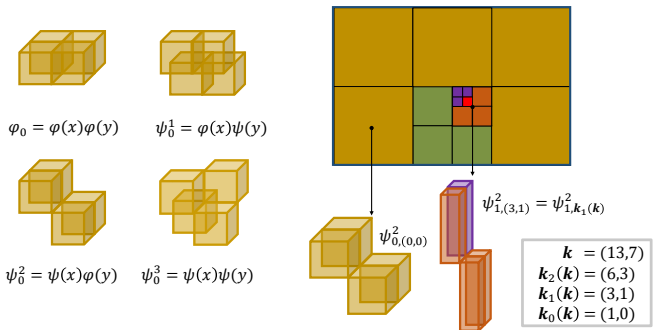

Fig. 1. The left plot shows the scaling function and the the wavelets at the lowest resolution level. The right plot illustrates how to obtain the indices of the squares that intersect the red one at the highest resolution.

$\Phi := \{\varphi_{0,\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}_{\boldsymbol{\ell}}}$ at the lowest resolution and the wavelets $\cup_{j=0}^{J-1} \Psi_j$, $\Psi_j := \{\psi^r_{j,\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}_{2^j \boldsymbol{\ell}}, r \in [1:3]}$, which capture the fluctuations of the functions in V_J at the intermediate scales. The wavelet basis $\Phi \cup \left(\cup_{j=0}^{\infty} \Psi_j\right)$ is the orthonormal Haar system on $Q_{\boldsymbol{\ell}}$ and spans $L^2(Q_{\boldsymbol{\ell}})$. We denote 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

$$f|_{I_{J,\boldsymbol{k}}} = \alpha_{\boldsymbol{k}_0(\boldsymbol{k})} + \sum_{j=0}^{J-1} \sum_{r=1}^{3} 2^j \beta^r_{j,\boldsymbol{k}_j(\boldsymbol{k})} \operatorname{sign}^r_j(\boldsymbol{k}). \tag{1}$$

In (1), $\boldsymbol{k}_j(\boldsymbol{k})$ are the indices of the unique 2^{-j}-resolution square that intersects $I_{J,\boldsymbol{k}}$, and $\operatorname{sign}^r_j(\boldsymbol{k})$ 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 $\{-2^j, 2^j\}$, cf. Figure 1.

Non-smooth analysis: Consider a locally Lipschitz function f on $\mathbb{R}^n$. It is known from Rademacher's theorem that f is differentiable almost everywhere. The Clarke generalized gradient of f at x is defined as $\bar{\partial} f(x) := \operatorname{conv}\{\lim_k \nabla f(x_k) \,|\, x_k \to x, x_k \in A\}$ (following the notation of [7], [18], and [23]), where A is any full-measure subset of a neighborhood of x where f is differentiable. A point $x \in \mathbb{R}^n$ is called Clarke stationary for f if $0 \in \bar{\partial} f(x)$, which generalizes the notion of a stationary point for continuously differentiable functions. The Clarke ε-subdifferential of f at x is defined as $\bar{\partial}_\varepsilon f(x) := \operatorname{conv}(\bar{\partial} f(B(x, \varepsilon)))$. Considering any set $\mathcal{D}_f$ of full measure on $\mathbb{R}^n$ where f is differentiable, its Clarke ε-subdifferential can be approximated by the set

$$G_\varepsilon(x) := \operatorname{cl}(\operatorname{conv}(\nabla f(B(x, \varepsilon) \cap \mathcal{D}_f))),$$

introduced in [7], since $G_\varepsilon(x) \subset \bar{\partial}_\varepsilon f(x)$ and $\bar{\partial}_{\varepsilon_1} f(x) \subset G_{\varepsilon_2}(x)$ for $0 \leq \varepsilon_1 < \varepsilon_2$.

III. Problem formulation

Here, we present the formulation of the coverage optimization problem following [5] and introduce the problem

of interest in this paper. Consider a bounded domain $Q \subset \mathbb{R}^2$ and a probability density function ρ, supported on Q, that captures the probability that a certain event of interest may occur. Consider also a (non-decreasing) performance function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, that encodes the cost for the locational discrepancy of an agent, like the travel time or the required energy consumption to get from one place to another. Our goal is to deploy M agents in Q so that the one that is closest to the place where an event may occur will go there to offer its service. Denoting the agents' positions by $p_1, \ldots, p_M$, the cost of an event occurring at a certain $x \in Q$ is equal to the performance cost of the agent that is closest to it, i.e., $\min_{i=1,\ldots,M} h(\|p_i - x\|)$. Taking into account that the locations of such events are randomly distributed according to the density ρ, we seek to minimize the expected cost function

$$\mathcal{H}_\rho(P) := \int_Q \min_{i=1,\ldots,M} h(\|p_i - x\|)\rho(x)dx, \tag{2}$$

where we use the shorthand notation $P = (p_1, \ldots, p_M)$. Denoting $A_{\text{diff}} := \{(p_1, \ldots, p_M) \in A \,|\, p_i \neq p_j \text{ for all } i \neq j\}$ for each $A \subset \mathbb{R}^{2M}$, for $P \in Q^M_{\text{diff}}$, the cost is equivalently given as

$$\mathcal{H}_\rho(P) := \sum_{i=1}^{M} \int_{V_i} h(\|p_i - x\|)\rho(x)dx,$$

where $V_i \equiv V_i(P)$ is the Voronoi region of agent i defined by $V_i(P) := \{x \in \mathbb{R}^2 \,|\, \|x - p_i\| \leq \|x - p_j\| \text{ for all } j \neq i\}$. Here we use the convention to define the Voronoi regions over the whole space $x \in \mathbb{R}^2$ instead of only focusing on the region Q since the density $\rho(x)$ in the above integrals vanishes when x is outside Q. This allows us later to also consider agent positions that lie outside Q, which facilitates proving convergence of the algorithm.

Throughout the paper, we take[1] Q to be the rectangle $Q_{\boldsymbol{\ell}} = [0, \ell_1] \times [0, \ell_2]$ defined by positive integers $\boldsymbol{\ell} = (\ell_1, \ell_2)$. The challenge we address here is having the agents optimize $\mathcal{H}_\rho$ when the density ρ is unknown. Instead, we only have access to N i.i.d. samples $X_1, \ldots, X_N$ taken from ρ. Since the number of these samples is typically limited in many practical scenarios, it is not possible to accurately infer the distribution of the data, resulting in model misspecification. We assume the unknown density is bounded as follows.

Assumption 3.1: ***(Upper and lower density bounds).*** There exist $\rho_{\text{low}} : Q \to \mathbb{R}_{\geq 0}$, $\rho_{\text{up}} : Q \to \bar{\mathbb{R}}_{\geq 0}$ with

$$0 \leq \rho_{\text{low}}(x) \leq \rho(x) \leq \rho_{\text{up}}(x) \quad \forall x \in Q. \tag{3}$$

This assumption enables us to embed prior knowledge when inferring the unknown density from data. For instance, we may know beforehand that the (unknown) probability does not exceed a threshold $p^\star > 0$ over a subset A of Q and that no point of A is $c \geq 1$ times more likely to be sampled than any other point in A, where c encodes how far from uniform the distribution is on this set. Then, we can pick $\rho_{\text{up}}(x) := \frac{cp^\star}{\text{area}(A)}$ for $x \in A$ and $\rho_{\text{up}}(x) := +\infty \in \bar{\mathbb{R}}_{\geq 0}$ to indicate that there are no density constraints outside A.

To hedge against uncertainty about the density, instead of (2), we solve the distributionally robust coverage problem

$$\min_{P \in Q^M} \max_{\rho \in \mathcal{P}} \int_Q \min_{i=1,\ldots,M} h(\|p_i - x\|)\rho(x)dx, \tag{4}$$

where $\mathcal{P}$ is a data-driven ambiguity set of probability densities that contains the true density with high probability and respects Assumption 3.1. The construction of $\mathcal{P}$ is a main goal of this work and is essential to obtain a tractable algorithm to solve (4). To this end, we use the Haar wavelet ambiguity sets introduced in [4]. These sets are data-driven and, unlike the commonly used Wasserstein balls for such problems, they only contain densities and can incorporate assumptions like Assumption 3.1 in a direct way.

IV. Wavelet estimator-based coverage DRO

In this section, we follow the wavelet estimator construction of [4] to build an ambiguity set of probability densities for the true density ρ. Using the N independent samples $X_1, \ldots, X_N$, 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

$$\widehat{\alpha}_\varphi := \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i), \qquad \varphi \in \Phi, \tag{5a}$$

$$\widehat{\beta}_\psi := \frac{1}{N} \sum_{i=1}^{N} \psi(X_i), \qquad \psi \in \cup_{j=0}^{J-1} \Psi_j. \tag{5b}$$

The Haar wavelet basis $\Phi \cup \{\Psi_j\}_{j=0}^{\infty}$ is the one described in Section II. 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. We compactly denote by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$ the coefficients of the scaling functions and the wavelets at each scale j, and $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}_j$ the corresponding coefficients of the estimator. Given the radii $\boldsymbol{\varepsilon} = (\varepsilon_0, \ldots, \varepsilon_J)$, the ambiguity set is determined through the wavelet coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{J-1}) \in \mathbb{R}^K$, $K := 4^J \ell_1 \ell_2$, that satisfy

$$\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|^2 \leq \varepsilon_0, \ \|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\|^2 \leq \varepsilon_{j+1}, \ j \in [0 : J-1] \tag{6}$$

and the following constraints:

- *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

$$\sum_{\boldsymbol{k} \in \mathbb{Z}_{\boldsymbol{\ell}}} \alpha_{\boldsymbol{k}} = 1. \tag{7}$$

- *Upper and lower density bounds.* Since the true density should satisfy the bounds of Assumption 3.1, these are

[1] Although our analysis can be generalized to higher dimensions, we focus on $\mathbb{R}^2$ to simplify the exposition.

captured at resolution 2^{-J} by the linear constraints

$$\min_{x\in I_{J,\boldsymbol{k}}} \rho_{\text{low}}(x) \le \alpha_{\boldsymbol{k}_0(\boldsymbol{k})} + \sum_{j=0}^{J-1}\sum_{r=1}^{3} 2^j \beta^r_{j,\boldsymbol{k}_j(\boldsymbol{k})} \operatorname{sign}^r_j(\boldsymbol{k}) \le \max_{x\in I_{J,\boldsymbol{k}}} \rho_{\text{up}}(x) \quad \forall \boldsymbol{k} \in \mathbb{Z}_{2^J\boldsymbol{\ell}}, \tag{8}$$

with $\boldsymbol{k}_j(\boldsymbol{k})$ and $\operatorname{sign}^r_j(\boldsymbol{k})$ as given in (1).

Note that the right-hand-side constraint in (8) becomes trivial when $\max_{x\in I_{J,\boldsymbol{k}}} \rho_{\text{up}}(x) = +\infty$. In addition, (8) always implies the non-negativity constraint for the density ρ, which in the absence of any further density constraints is equivalent to setting $\rho_{\text{low}}(x) \equiv 0$ and $\rho_{\text{up}}(x) \equiv +\infty$. We refer to the thresholds $\varepsilon_0, \dots, \varepsilon_J$ in (6) as the ambiguity radii, which can be tuned so that the ambiguity set contains the projection of the true density to the space $D(V_J^{\boldsymbol{\ell}})$ with prescribed probability [4, Theorem 5.2]. We compactly denote by $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_{J-1})$ the Haar coefficients of a distribution in $V_J^{\boldsymbol{\ell}}$ and by Θ the set of parameters $\boldsymbol{\theta}$ that satisfy the constraints (6), (7), and (8). By parameterizing the ambiguity set through $\boldsymbol{\theta} \in \Theta$, the DRO problem is equivalently written

$$\min_{P\in Q^M} \max_{\boldsymbol{\theta}\in\Theta} \int_Q \min_{i=1,\dots,M} h(\|p_i - x\|)\rho_{\boldsymbol{\theta}}(x)dx = \min_{P\in Q^M_{\text{diff}}} \max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{M} \int_{V_i} h(\|p_i - x\|)\rho_{\boldsymbol{\theta}}(x)dx, \tag{9}$$

with $\rho_{\boldsymbol{\theta}}$ the parameterized distributions. Using the compact notation $F(P,\boldsymbol{\theta}) := \int_Q \min_{i=1,\dots,M} h(\|p_i - x\|)\rho_{\boldsymbol{\theta}}(x)dx$, $f(P) := \max_{\boldsymbol{\theta}\in\Theta} F(P,\boldsymbol{\theta})$, the DRO problem is written as

$$\min_{P\in Q^M} f(P) = \min_{P\in Q^M} \max_{\boldsymbol{\theta}\in\Theta} F(P,\boldsymbol{\theta}). \tag{10}$$

Note that this problem is non-smooth and non-convex. Denoting $\rho_{\boldsymbol{\theta}}(x) \equiv \sum_{k=1}^K \theta_k\phi_k(x)$ with $\phi_k \equiv \varphi$ for some $\varphi \in \Phi$ or $\phi_k \equiv \psi$ for some $\psi \in \Psi$, F can be expressed as

$$F(P,\boldsymbol{\theta}) = \langle \boldsymbol{c}(P), \boldsymbol{\theta}\rangle, \tag{11}$$

for all $P \in Q^M_{\text{diff}}$, where $\boldsymbol{c}(P) := (c_1(P), \dots, c_K(P))$ and

$$c_k(P) := \sum_{i=1}^{M} \int_{V_i} h(\|p_i - x\|)\phi_k(x)dx. \tag{12}$$

Remark 4.1: ***(Closed-form expressions for the integrals in*** (12)***).*** All the Haar wavelets take constant values across the squares at the lowest resolution. Since the Voronoi regions are convex polygons, the integrals (12) are finite sums of the integrals of $h(\|p_i - x\|)$ across polygonal regions and can be computed analytically for polynomial h. •

V. Gradient sampling for coverage control

In this section we provide the optimization algorithm to solve the DRO problem (10). Given its nonsmoothness, we build on a modification of the Gradient Sampling (GS) algorithm, introduced in [7] to optimize locally Lipschitz functions.

A. Modified GS sampling algorithm for nested cost functions

Here, we assume the objective function $f : \mathbb{R}^n \to \mathbb{R}$ takes the form

$$f(x) = \max_{\theta\in\Theta} F(x,\theta), \tag{13}$$

where $\Theta \subset \mathbb{R}^d$, and denote by $\mathcal{D}_f$ the set on which it is *continuously* differentiable. We further assume throughout the section that $\mathcal{D}_f$ is an *open set of full measure* in $\mathbb{R}^n$ (hence also dense)[2]. The main reason why we modify the algorithm is that one cannot directly determine the value of the objective function at any given point, given its definition through the solution of a maximization problem, which can typically only be solved approximately. As a result, we cannot compute the gradients of the function exactly at points where it is differentiable to approximate its Clarke ε-subdifferential. Neither can we actually check differentiablility of f at a specific point, which is required in the original algorithm. The third issue is that not knowing the exact values of the function poses challenges on how to perform line search to determine the stepsize.

For the function f in (13), we denote

$$\theta_\star(x) := \operatorname{argmax}_{\theta\in\Theta} F(x,\theta),$$

for each $x \in \mathbb{R}^n$. We also assume that each $\theta_\star(x)$ is nonempty and consider access to an oracle, representing an optimization algorithm that carries out the maximization of F with respect to θ and returns, for each x and user-defined accuracy δ, a value $\theta_\star$ at a distance $\operatorname{dist}(\theta_\star, \theta_\star(x)) < \delta$. We make the following assumption about the differentiability properties of F.

Assumption 5.1: ***(Regularity of*** F***).*** The function $x \mapsto F(x,\theta)$ is continuously differentiable for all $x \in \mathcal{D}_f$. In addition, for each compact $S \subset \mathbb{R}^n$, there exist $L_F(S), L_{F_x}(S) > 0$, such that for each $x' \in S \cap \mathcal{D}_f$ the functions $\theta \mapsto F(x',\theta)$ and $\theta \mapsto \nabla_x F(x',\theta)$ are globally Lipschitz with respect to θ with constants $L_F(S)$ and $L_{F_x}(S)$, respectively.

Here we provide our modification of the GS algorithm for functions of the form (13) that satisfy Assumption 5.1.

(Modified) Gradient Sampling Algorithm

Step 0: ***(Initialization)***

Select $x^1 \in \mathbb{R}^n$, $\alpha, \beta, \gamma \in (0,1)$, $\varepsilon_1, \nu_1 > 0$, $\mu, \vartheta \in (0,1]$, $m \in \{n+1, n+2, \dots\}$, and approximation parameters $\delta_k \searrow 0$. Set $k := 0$.

Step 1: ***(Approximation of the Clarke*** ε***-subdifferential by gradient sampling)***

Sample $x^{k1}, \dots, x^{km}$ independently and uniformly from $B(x^k, \varepsilon_k)$ and set

$$G_k := \operatorname{conv}(\{\nabla_x F(x^{k1}, \theta^{k1}_\star), \dots, \nabla_x F(x^{km}, \theta^{km}_\star)\}),$$

where

$$\|\theta^{ki}_\star - \theta_\star\| \le \delta_k \text{ for some } \theta_\star \in \theta_\star(x^{ki}),\ i = 1, \dots, m.$$

[2]This is required to prove convergence of the gradient sampling algorithm, as clarified in the recent paper [6].

Step 2: ***(Search direction computation)***

Find the optimizer g^k of the quadratic program

$$\begin{aligned} &\min \|g\|^2 \\ &\text{s.t. } g \in G_k. \end{aligned}$$

Step 3: ***(Sampling radius update)***

If $\|g^k\| \leq \nu_k$, set $t_k := 0$, $\nu_{k+1} := \vartheta\nu_k$, and $\varepsilon_{k+1} := \mu\varepsilon_k$, and **go to** Step 5.

Else, set $\nu_{k+1} := \nu_k$, $\varepsilon_{k+1} := \varepsilon_k$, and $d^k := -g^k/\|g^k\|$.

Step 4: ***(Limited Armijo line search)***

(i) Choose an initial step size $t \equiv t_{k,\text{init}} \geq t_{k,\min} := \gamma\varepsilon_k/3$.

(ii) Set the tolerance level $c_k := \gamma(1-\alpha)\|g^k\|\varepsilon_k/3$, pick $\theta_\star$ satisfying

$$\operatorname{dist}(\theta_\star, \theta_\star(x^k)) \leq \frac{c_k}{4L_F(\{x^k\})},$$

and set $\theta_\star^k := \theta_\star$.

(iii) Pick $\theta'_\star$ satisfying

$$\operatorname{dist}(\theta'_\star, \theta_\star(x^k + td^k)) \leq \frac{c_k}{4L_F(\{x^k + td^k\})},$$

and set $(\theta_\star^k)' := \theta'_\star$.

(iv) **If**

$$F(x^k + td^k, \theta'_\star) \leq F(x^k, \theta_\star) - \beta t_k\|g^k\| + \frac{c_k}{2},$$

set $t_k := t$ and **go to** Step 5.

(v) **If** $\gamma t < t_{k,\min}$, set $t_k := 0$ and **go to** Step 5.

(vi) Set $t := \gamma t$ and **go to** (iii).

Step 5: ***(Update)***

Set $x^{k+1} := x^k + t_k d^k$, $k := k+1$ and **go to** Step 1.

Step 0 of the GS algorithm contains the initialization of the decision variable and the initial tuning of the parameters. These parameters include the tolerances ε_k for the gradient sampling radius and ν_k for the size of the minimum-norm element of G_k. The initial values of these tolerances are set in Step 1 and their subsequent values are obtained using the discount factors μ and ϑ in Step 3 of the algorithm. Step 1 approximates the Clarke ε_k-subdifferential of f through the set G_k generated by approximations of sampled gradients of F, while Step 2 computes the minimum-norm element of this set. Step 3 is responsible for reducing the sampling radius and minimum-norm element tolerance when getting closer to Clarke stationarity. Step 4 performs a line search to determine the gradient step using approximations of the objective function. Finally, Step 5 updates the values of x^k based on the chosen stepsize and search direction.

B. Convergence of the GS algorithm

The following result establishes convergence of the GS algorithm under Assumption 5.1.

Theorem 5.2: ***(Convergence of the GS algorithm).*** Assume f of the form (13) is locally Lipschitz, lower bounded, and continuously differentiable on $\mathcal{D}_f$, which is an open set of full measure in $\mathbb{R}^n$. Assume further that F satisfies Assumption 5.1. Then, with probability one, the GS algorithm does not stop and $\nu_k, \varepsilon_k \searrow 0$. In addition, every accumulation point of $\{x^k\}$ is Clarke stationary for f. •

Remark 5.3: ***(Variants of the GS algorithm).*** Appropriate adjustments of the GS algorithm which account for the issues of not computing the derivative of f and avoiding its differentiability check have appeared in the literature. Specifically, [16] resolves the differentiablility check by randomly perturbing the gradient direction whereas [18] and [19] avoid it by allowing empty steps. which is also the approach that we consider here. A non-derivative version of the algorithm is further considered in [19], which approximates the gradient of f by Steklov averages. Nevertheless, these adjustments are not sufficient to deal with the objective functions we consider here, since they are defined implicitly though the solution of an inner maximization problem, and thus, both their values and their derivatives need to be approximated by different methods. •

C. Application of the GS algorithm to coverage optimization

Here, we show how the distributionally robust coverage control problem (10) fits into the framework described in Section V and is amenable to the Gradient Sampling algorithm, cf. Theorem 5.2. To this end, we henceforth assume that the function h in (4) is continuously differentiable. From [5, Theorem 2.16](i), each function c_k in (12), which is equivalently given by the expression

$$c_k(P) = \int_Q \min_{i=1,\dots,M} h(\|p_i - x\|)\phi_k(x)dx$$

that is valid for all $P \in \mathbb{R}^{2M}$, is globally Lipschitz on Q_c^M for any bounded set $Q_c \subset \mathbb{R}^2$. Thus, the same holds also for each function $F(P, \boldsymbol{\theta}) = \sum_{k=1}^K \theta_k c_k(P)$. In fact, since Θ is bounded, the functions $P \mapsto F(P, \boldsymbol{\theta})$ have a uniformly bounded Lipschitz modulus and it follows from [23, Proposition 9.10, Page 356] that f is globally Lipschitz on Q_c^M. Hence, it is also locally Lipschitz on $\mathbb{R}^{2M}$ and therefore differentiable almost everywhere. We henceforth assume that f is continuously differentiable on a set $\mathcal{D}_{f,\text{init}}$ of full measure. Establishing this fact in general for arbitrary h is challenging, albeit easily satisfied for specific choices (e.g., $h = \text{const}$). This assumption is needed to guarantee convergence of the algorithm but it is not necessary to guarantee differentiability at sampled points of the auxiliary function F with probability one, which we establish next, by showing that F satisfies Assumption 5.1.

To this end, we obtain from [5, Theorem 2.16](ii) that each function $c_k(P)$ is continuously differentiable on $\mathbb{R}^{2M}_{\text{diff}}$, which is open and of full measure (hence also dense) and its partial derivatives are given by

$$\frac{\partial}{\partial p_i} c_k(P) = \int_{V_i} \frac{\partial}{\partial p_i} h(\|p_i - x\|)\phi_k(x)dx.$$

This convenient expression is a result of the fact that the derivatives of the integrals in (12) with respect to changes in the boundaries of the Voronoi cells vanish. The intuition

behind this is that any infinitesimal increase of an agent's integral due to the shift of a boundary face is deducted from the integral of its neighbor along that face. As a result, we get that $\nabla_P F(P, \boldsymbol{\theta}) = \boldsymbol{c}'(P)\boldsymbol{\theta}$ for all $P \in \mathbb{R}^{2M}_{\text{diff}}$, where

$$\boldsymbol{c}'(P) := \begin{pmatrix} \frac{\partial}{\partial p_1} c_1(P) & \cdots & \frac{\partial}{\partial p_1} c_K(P) \\ \vdots & & \vdots \\ \frac{\partial}{\partial p_M} c_1(P) & \cdots & \frac{\partial}{\partial p_M} c_K(P) \end{pmatrix}$$

and therefore that f and each $\boldsymbol{\theta} \to F(P, \boldsymbol{\theta})$ are continuously differentiable on the full-measure set $\mathcal{D}_f := \mathcal{D}_{f,\text{init}} \cap \mathbb{R}^{2M}_{\text{diff}}$. In addition, we get that for any $P' \in \mathcal{D}_f$ it holds that

$$\|\nabla_P F(P', \boldsymbol{\theta}_1) - \nabla_P F(P', \boldsymbol{\theta}_2)\| \leq \|c'(P')\| \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

which implies existence of a Lipschitz constant $L_{F_P}(S)$ for ∇F_P with respect to $\boldsymbol{\theta}$ for every compact $S \subset \mathbb{R}^{2M}$. The same Lipschitz property for $\boldsymbol{\theta} \to F(P, \boldsymbol{\theta})$ follows directly from (11) and continuity of $P \mapsto \boldsymbol{c}(P)$. Hence, Assumption 5.1 is fulfilled for F and it follows from Theorem 5.2 that every accumulation point of the GS algorithm for the coverage objective function is also Clarke stationary. It is worthwhile noting that the inner optimization problem of the distributionally robust coverage algorithm is a linear program and can be numerically solved up to high accuracy using commercial solvers.

Remark 5.4: ***(Coverage cost-function domain).*** Extending the domain of the coverage cost function f from Q^M to $\mathbb{R}^{2M}$ facilitates proving the convergence of the algorithm. Future research will include the verification that executions of the algorithm starting inside Q^M will remain sufficiently close to it for all iterations, and that $\text{dist}(x^k, Q^M)$ will always approach zero for large k. Such an invariance result is justified by the fact that placing agents outside Q^N incurs a higher cost, as all the probability mass is inside the set, which makes the negated gradient of the objective function point towards Q^N. •

VI. Conclusions

This paper proposes a distributionally robust coverage control problem, for the optimal deployment of a group of robots in an uncertain environment. In particular, our formulation employs Haar-wavelet parameterized ambiguity sets that are accompanied by rigorous statistical guarantees and conveniently incorporate prior information about the property of interest. To solve this problem, we propose a variant of the GS algorithm, which can handle objective functions that are implicitly defined through the solution of an inner optimization problem.

Our future work will include the derivation of distributed algorithms for the solution to the inner optimization to solve the distributionally robust coverage control problem in a decentralized manner. We further aim to provide invariance guarantees for the GS algorithm to provably restrict its evolution over bounded domains, to identify costs where continuous differentiablility over an open set of full measure can be rigorously established, and to experimentally validate the approach.

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] D. Boskos, J. Cortés, and S. Martínez, "Distributionally robust optimization via Haar wavelet ambiguity sets," in *IEEE Int. Conf. on Decision and Control*, Cancun, Mexico, Dec. 2022, pp. 4782–4787.

[5] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009.

[6] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. Simões, "Gradient sampling methods for nonsmooth optimization," *Numerical nonsmooth optimization: State of the art algorithms*, pp. 201–225, 2020.

[7] J. V. Burke, A. S. Lewis, and M. L. Overton, "A robust gradient sampling algorithm for nonsmooth, nonconvex optimization," *SIAM Journal on Optimization*, vol. 15, no. 3, pp. 751–779, 2005.

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

[9] J. Cortés and F. Bullo, "Nonsmooth coordination and geometric optimization via distributed dynamical systems," *SIAM Review*, vol. 51, no. 1, pp. 163–189, 2009.

[10] J. Cortés, S. Martinez, and F. Bullo, "Spatially-distributed coverage optimization and control with limited-range interactions," *ESAIM. Control, Optimisation & Calculus of Variations*, vol. 11, no. 4, pp. 691–719, 2005.

[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] 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.

[13] Y. Diaz-Mercado, S. G. Lee, and M. Egerstedt, "Distributed dynamic density coverage for human-swarm interactions," in *American Control Conference*, 2015, pp. 353–358.

[14] 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.

[15] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *Mathematics of Operations Research*, 2022.

[16] E. S. Helou, S. A. Santos, and L. E. A. Simões, "On the differentiability check in gradient sampling methods," vol. 31, no. 5, pp. 983–1007, 2016.

[17] Y. Kantaros, M. Thanou, and A. Tzes, "Distributed coverage control for concave areas by a heterogeneous robot–swarm with visibility sensing constraints," *Automatica*, vol. 53, pp. 195–207, 2015.

[18] K. C. Kiwiel, "Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization," *SIAM Journal on Optimization*, vol. 18, no. 2, pp. 379–388, 2007.

[19] ——, "A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1983–1994, 2010.

[20] S. Martínez, "Distributed interpolation schemes for field estimation by mobile sensor networks," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 2, pp. 491–500, 2010.

[21] 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.

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

[23] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer, 1998.

[24] M. Schwager, D. Rus, and J. J. Slotine, "Decentralized, adaptive coverage control for networked robots," *International Journal of Robotics Research*, vol. 28, no. 3, p. 357–375, 2009.

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