Data-Driven Ambiguity Sets for Linear Systems Under Disturbances and Noisy Observations

Dimitris Boskos Jorge Cortés Sonia Martínez

Abstract— **This paper studies the characterization of Wasserstein ambiguity sets for dynamic random variables when noisy partial observations are progressively collected from their evolving distribution. The ambiguity sets are accompanied by quantitative guarantees about the true distribution of the data, which renders them appropriate for the formulation of robust stochastic optimization problems. To describe the evolution of the variable, we consider a linear discrete-time dynamic model with random initial conditions, stochastic uncertainty in the dynamics, and partial noisy measurements. The probability distribution of all the involved random elements is supposed to be unknown. To make inferences about the distribution of the state vector, we collect several output samples from multiple realizations of the process. We use a classical Luenberger observer to obtain full-state estimators for the independent realizations and exploit these further to build the centers of the ambiguity sets.**

I. Introduction

Making decisions under uncertainty is an unavoidable task for a wide range of today's engineering applications, where the complexity of the encountered systems does not allow the deterministic modeling of all their components. Thus, the designer seeks to infer about the properties of such elements via the collection of data. This leads to the formulation of data-driven stochastic optimization problems to provide quantitative decisions in the face of uncertainty, finding applications in numerous domains such as finance, networked control systems, and machine learning. The articulation of these problems via distributionally robust optimization, seeks to provide guaranteed results against credible variations of the data. This is accomplished by evaluating the optimal worst-case performance over an ambiguity set of probability distributions that contains the true one with high confidence. Typical Distributionally Robust Optimization (DRO) formulations are based on the assumption that the samples generated from the distribution are measured in a direct manner and can all be obtained right away. This paper instead looks at scenarios where the random variable evolves dynamically and partial measurements of the data are collected, possibly corrupted by noise. Our goal is to build reliable ambiguity sets for such cases by leveraging the underlying dynamics. We also study how the probabilistic properties of the noise affect the ambiguity set size while maintaining the same guarantees.

Literature review: Stochastic optimization is a well established research area, with several techniques available to provide optimal decisions in the face of uncertainty [1]. Typical objectives include expected-cost minimization, and chance-constrained optimization, among others. For stochastic programs, model imperfections occur naturally due to a lack of knowledge on random elements and have led to distributionally robust variants of the problems at hand [2], [3], [4]. This is typical in data-driven cases, where the probability distribution of the random variables is inferred in an approximate manner using a finite amount of collected data [5]. To hedge against data uncertainty, optimal transport ambiguity sets have emerged as a promising way to consider all distributions up to some distance from the sample average approximation in the Wasserstein metric [6], using recent concentration of measure results [7]. The work [8] introduces tractable reformulations of DRO problems with Wasserstein ambiguity sets accompanied by finite-sample performance guarantees, which have later been exploited for chance-constrained programs [9], [10]. The work [11] develops distributed optimization algorithms using Wasserstein balls, while optimal transport ambiguity sets have also been recently connected to regularization for machine learning [12], [13]. The paper [14] exploits Wasserstein balls to robustify data-driven online optimization algorithms, and [15] leverages them for the design of distributionally robust Kalman filters. Time-varying aspects of such ambiguity sets are considered in [16] for dynamic traffic models, and in [17], which constructs Wasserstein ambiguity balls using progressively assimilated dynamic data for processes with random initial conditions. Compared to [17], the present paper addresses nontrivial generalizations, because (i) the state distribution does no longer evolve under deterministic dynamics, due to internal noise and, most notably, (ii) noisy partial measurement are considered, which generate an additional stochastic element that requires quantification for the ambiguity set guarantees.

Statement of contributions: We build ambiguity sets for dynamically evolving random variables of an unknown distribution using progressively collected samples from multiple realizations of the process. Our contributions are focused on establishing quantifiable characterizations of Wasserstein ambiguity sets with rigorous probabilistic guarantees under several nonidealities associated to the data assimilation. We construct ambiguity sets for dynamic processes with random initial conditions and further uncertainty in the dynamics. The probabilistic model of the random elements is unknown and we make inferences about the process distribution along time using incrementally assimilated samples. Our main

The authors are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, {dboskos,cortes,soniamd}@ucsd.edu.

contribution is the construction of ambiguity sets for the evolving random variable under partial-state measurements that are corrupted by noise. This is accomplished by using an observer to approximate the dynamics' state value from the output samples, and exploiting concentration of measure results to quantify reliable closeness to the empirical distribution, while accounting for the estimation error. The results are validated in simulation for an optimal sensor placement example. We note that our objective is fundamentally different from classical Kalman filtering, where the initial state and dynamics noise distributions are known and Gaussian, and hence, the state distribution over time is a known Gaussian random variable too. Here, instead, we are interested to infer about the unknown state distribution from data collected by multiple executions of the dynamical system. Thus, for each such execution we use a suboptimal state estimator, precisely due to the fact that we have no concrete knowledge of the state and noise random models. Due to space constraints, all proofs are omitted and will appear elsewhere

II. Preliminaries

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

Notation: We denote by $\|\cdot\|_p$ the p-th norm in $\mathbb{R}^n$, $p \in [1,\infty]$, using also the shorter notation $\|\cdot\| \equiv \|\cdot\|_2$ for the Euclidean norm. Let $B_\infty^n(\rho)$ denote the ball of center zero and radius ρ in $\mathbb{R}^n$ with the norm $\|\cdot\|_\infty$. We use the notation $[n_1 : n_2]$ for the set of integers $\{n_1, n_1+1, \ldots, n_2\} \subset \mathbb{N} \cup \{0\} =: \mathbb{N}_0$. The diameter of a set $S \subset \mathbb{R}^n$ is defined as $\mathrm{diam}(S) := \sup\{\|x-y\|_\infty \,|\, x, y \in S\}$ and for $z \in \mathbb{R}^n$, $S + z := \{x+z \,|\, x \in S\}$. We denote the induced Euclidean norm of a matrix $A \in \mathbb{R}^{m\times n}$ by $\|A\| := \max_{\|x\|=1} \|Ax\|/\|x\|$. Given $B \subset \Omega$, we denote by $\mathbf{1}_B$ the indicator function of B on Ω, with $\mathbf{1}_B(x) = 1$ for $x \in B$ and $\mathbf{1}_B(x) = 0$ for $x \notin B$.

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 $p \geq 1$, $\mathcal{P}_p(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) \,|\, \int_{\mathbb{R}^d} \|x\|^p d\mu < \infty\}$ is the set of probability measures in $\mathcal{P}(\mathbb{R}^d)$ with finite p-th moment. Given $p \geq 1$, 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. For any $\mu \in \mathcal{P}(\mathbb{R}^d)$, its support is the closed set $\mathrm{supp}(\mu) := \{x \in \mathbb{R}^d \,|\, \mu(U) > 0 \text{ for each neighborhood } U \text{ of } x\}$, or equivalently, the smallest closed set with measure one. Given a measurable space $(\Omega, \mathcal{F})$, an exponent $p \geq 1$, the convex function $\psi_p(x) := e^{x^p} - 1$, and the linear space of scalar random variables $L_{\psi_p} := \{X \,|\, \mathbb{E}[\psi_p(|X|/t)] < \infty \text{ for some } t > 0\}$ on $(\Omega, \mathcal{F})$, the ψ_p-Orlicz norm (cf. [18, Section 2.7.1]) of $X \in L_{\psi_p}$ is

$$\|X\|_{\psi_p} := \inf\{t > 0 \,|\, \mathbb{E}[\psi_p(|X|/t)] \leq 1\}.$$

When $p = 1$ and $p = 2$, each random variable in L_{ψ_p} is sub-exponential and sub-Gaussian, respectively. We also denote by $\|X\|_p \equiv \big(\mathbb{E}\big[|X|^p\big]\big)^{\frac{1}{p}}$ the norm of a scalar random variable with finite p-th moment, i.e., the classical norm in $L^p(\Omega)$. The interpretation of $\|\cdot\|_p$ as the p-norm of a vector in $\mathbb{R}^n$ or a random variable in L^p should be clear from the context throughout the paper. Given a set $\{X_i\}_{i\in I}$ of random variables, we denote by $\sigma(\{X_i\}_{i\in I})$ the σ-algebra generated by them. We conclude with a useful technical result whose proof follows from Fubini's theorem [19, Theorem 2.6.5].

Lemma 2.1: ***(Expectation inequality).*** Consider the independent random vectors X and Y, taking values in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively, and let $g(X,Y)$ be integrable. Assume that $\mathbb{E}[g(x,Y)] \geq k(x)$ for some integrable function k and all $x \in K \subset \mathrm{supp}(X) \subset \mathbb{R}^{n_1}$. Then, $\mathbb{E}[g(X,Y)] \geq \mathbb{E}[k(X)]$.

III. Problem formulation

Consider a stochastic optimization problem where the objective function $x \mapsto f(x,\xi)$ depends on a random variable ξ whose distribution P_ξ is *unknown*. To hedge this uncertainty, rather than using the empirical distribution

$$P_\xi^N := \frac{1}{N} \sum_{i=1}^N \delta_{\xi^i}, \tag{1}$$

formed by N i.i.d. samples $\xi^1, \ldots, \xi^N$ of P_ξ to compute a sample average approximation of the expected value of f, one can instead consider the worst-case expectation problem

$$\inf_{x\in\mathcal{X}} \sup_{P\in\mathcal{P}^N} \mathbb{E}_P[f(x,\xi)],$$

over some *ambiguity set* $\mathcal{P}^N$ of probability measures. Different approaches exist to construct these ambiguity sets so that they contain the true distribution P_ξ with high confidence. We are interested in approaches that employ data, and in particular the empirical distribution P_ξ^N, to construct them. In the present setup, the data is generated by a dynamical system subject to disturbances, and we only collect partial (instead of full) measurements that are distorted by noise. Therefore, it is no longer obvious how to build a candidate state distribution as in (1) from the collected samples. Further, we seek to address this in a distributionally robust way, i.e., finding a suitable replacement $\widehat{P}_\xi^N$ for (1) together with an associated ambiguity set, by exploiting the dynamics of the underlying process.

We consider data generated by a discrete-time system

$$\xi_{k+1} = A\xi_k + Gw_k + r_k, \quad \xi_k \in \mathbb{R}^d, \quad w_k \in \mathbb{R}^q, \tag{2a}$$

with linear output

$$\zeta_k = H\xi_k + v_k, \quad \zeta_k \in \mathbb{R}^r. \tag{2b}$$

The initial condition ξ_0 and the noise w_k and v_k, $k \in \mathbb{N}_0$ in the dynamics and the measurements, respectively, are random variables with an *unknown* distribution, whereas r_k is a deterministic and known input signal. We seek to build an ambiguity set for the state distribution at certain time $\ell \in \mathbb{N}$, by collecting data from multiple independent realizations of the process, denoted by ξ^i, $i \in [1:N]$. This can occur, for instance, when the same process is executed repeatedly, or in multi-agent scenarios where identical entities are subject to the same dynamics. To formally describe the problem, we consider a large enough probability space $(\Omega, \mathcal{F}, \mathbb{P})$ containing all random elements from these realizations, and make the following sampling assumption.

Assumption 3.1: ***(Sampling schedule).*** For each realization i of system (2), output samples $\zeta_0^i, \ldots, \zeta_\ell^i$ are collected over the discrete time instants of the sampling horizon $[0 : \ell]$.

Obtaining quantifiable characterizations for the ambiguity sets requires some further hypotheses on the classes of the distributions P_{ξ_0} of the initial condition, P_{w_k} of the dynamics noise, and P_{v_k} of the measurement errors (cf. Figure 1). These assumptions are made for individual realizations and allow us to consider non-identical observation error distributions—in this way, we allow for the case where each realization is measured by a non-identical sensor of variable precision.

Assumption 3.2: ***(Distribution classes).*** Consider a finite sequence of realizations ξ^i, $i \in [1 : N]$ of (2a) with associated outputs given by (2b), and noise elements w_k^i, v_k^i, $k \in \mathbb{N}_0$.

H1: The distributions $P_{\xi_0^i}$, $i \in [1 : N]$, are identically distributed; further $P_{w_k^i}$, $i \in [1 : N]$, are identically distributed for all $k \in \mathbb{N}_0$.

H2: The sigma fields $\sigma(\{\xi_0^i\} \cup \{w_k^i\}_{k\in\mathbb{N}_0})$, $\sigma(\{v_k^i\}_{k\in\mathbb{N}_0})$, $i \in [1 : N]$ are independent.

H3: The supports of the distributions $P_{\xi_0^i}$ and $P_{w_k^i}$, $k \in \mathbb{N}_0$ are compact, centered at the origin, and have diameters $2\rho_{\xi_0}$ and $2\rho_w$, respectively, for all i.

H4: The components of the random vectors v_k^i have uniformly bounded L^p and ψ_p-Orlicz norms, as follows,

$$0 < m_v \leq \|v_{k,l}^i\|_p \leq M_v, \quad \|v_{k,l}^i\|_{\psi_p} \leq C_v,$$

for all $k \in \mathbb{N}_0$, $i \in [1 : N]$, and $l \in [1 : r]$, where $p \geq 1$.

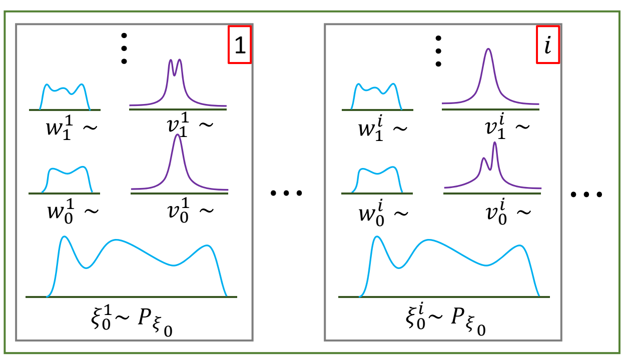


Fig. 1: Illustration of the probabilistic models for the random variables in the dynamics and observations according to Assumption 3.2.

Since the collected measurements give no direct access to the system's state distribution, we aim to leverage the dynamics and estimate the state from the assimilated output values. To guarantee some boundedness notion for the state estimation errors over a sufficient evolution horizon, we make some further assumptions for the dynamics.

Assumption 3.3: ***(Detectability).*** The pair (A, H) is detectable, i.e., $A + KH$ is convergent for some $K \in \mathbb{R}^{d\times r}$.

Problem statement: Under Assumptions 3.1 and 3.2 on measurements and distributions of N realizations of the system (2), we seek to construct an estimator $\widehat{\xi}_\ell^i(\zeta_0^i, \ldots, \zeta_\ell^i)$ for the state of each realization to build an ambiguity set for the state distribution at time ℓ with probabilistic guarantees. Further, under Assumption 3.3 on system detectability, we aim to characterize the effect of the estimation precision on the quality of the ambiguity sets.

We proceed to address this problem in Section IV by exploiting a Luenberger observer to estimate the states of the collected data and using them to replace the classical empirical distribution (1) in the ambiguity-set construction. To obtain the guarantees we leverage concentration inequalities to bound the distance between the updated empirical distribution and the true state-distribution with high confidence. The precise effect of measurement noise in the ambiguity radius is quantified in Section V, where we also study the increase of ambiguity radius due to the noise and the positive effect on it of detectability for longer evolution horizons.

IV. STATE-ESTIMATOR BASED AMBIGUITY SETS

We turn to address the question of how to construct an ambiguity set at certain time instant ℓ, when samples are collected according to Assumption 3.1 from (2). If we had access to N independent full-state samples $\xi^1, \ldots, \xi^N$ from the distribution of ξ at ℓ, we could construct an ambiguity ball in the Wasserstein metric W_p, that is centered at the empirical distribution (1) with $\xi^i \equiv \xi_\ell^i$ and contains the true distribution with high confidence. In particular, for any confidence $1 - \beta > 0$, it is possible, cf. [8, Theorem 3.5], to specify an ambiguity ball radius $\varepsilon_N(\beta)$ so that the true distribution of ξ_ℓ is in this ball with confidence $1 - \beta$, i.e.,

$$\mathbb{P}(W_p(P_{\xi_\ell}^N, P_{\xi_\ell}) \leq \varepsilon_N(\beta)) \geq 1 - \beta.$$

Instead, since we only can collect noisy partial measurements of the state, we use a Luenberger observer to estimate ξ. The dynamics of the observer, initiated from zero, is given by

$$\widehat{\xi}_{k+1} = A\widehat{\xi}_k + K(H\widehat{\xi}_k - \zeta_k) + r_k, \qquad \widehat{\xi}_0 = 0, \qquad (3)$$

where K is a constant gain matrix. We thus define the *(dynamic) estimator-based empirical distribution*

$$\widehat{P}_{\xi_k}^N := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_k^i}, \qquad (4)$$

formed by using the corresponding estimates from system (3) for the independent realizations of (2a). Denoting by $e_k := \xi_k - \widehat{\xi}_k$ the error between (2a) and the observer (3), the error dynamics is

$$e_{k+1} = Fe_k + Gw_k + Kv_k, \qquad e_0 = \xi_0,$$

where $F := A + KC$ and ξ_0 is the initial condition of (2a). In particular,

$$e_k = F^k\xi_0 + \sum_{\kappa=0}^{k-1}\left(F^\kappa Gw_{k-\kappa-1} + F^\kappa Kv_{k-\kappa-1}\right) \qquad (5)$$

for all $k \geq 1$. To build the ambiguity set at time ℓ, we set its center at the estimator-based empirical distribution $\widehat{P}_{\xi_\ell}^N$ given by (4). In what follows, we leverage concentration of measure results to quantify an ambiguity radius $\psi_N(\beta)$ so that the resulting Wasserstein ball contains the true distribution with a given confidence $1-\beta$.

Note that the random variable ξ_k^i of a system realization at time k is a function $\xi_k^i(\xi_0^i, \boldsymbol{w}_k^i)$ of the random initial condition ξ_0^i and the dynamics noise $\boldsymbol{w}_k^i \equiv (w_0^i, \ldots, w_{k-1}^i)$. Analogously, the estimated state $\widehat{\xi}_k^i$ of each observer realization is a stochastic variable $\widehat{\xi}_k^i(\xi_0^i, \boldsymbol{w}_k^i, \boldsymbol{v}_k^i)$ with additional randomness induced through the output noise $\boldsymbol{v}_k^i \equiv (v_0^i, \ldots, v_{k-1}^i)$. Using the compact notation $\boldsymbol{\xi}_0 \equiv (\xi_0^1, \ldots, \xi_0^N)$, $\boldsymbol{w}_k \equiv (\boldsymbol{w}_k^1, \ldots, \boldsymbol{w}_k^N)$, and $\boldsymbol{v}_k \equiv (\boldsymbol{v}_k^1, \ldots, \boldsymbol{v}_k^N)$ for the corresponding initial conditions, dynamics noise, and output noise of all realizations, respectively, we can denote the true- and estimator-based-empirical distributions at time ℓ as $P_{\xi_\ell}^N(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell)$ and $\widehat{P}_{\xi_\ell}^N(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell, \boldsymbol{v}_\ell)$. If we view the initial conditions and the corresponding internal noise of the realizations ξ^i over the whole time horizon as deterministic quantities, we use the alternative notation $P_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega})$ and $\widehat{P}_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{v}_\ell)$ for the corresponding distributions, where $\boldsymbol{z} = (z^1, \ldots, z^N)$, $z^1 \equiv \xi_0^1, \ldots, z^N \equiv \xi_0^N$, and $\boldsymbol{\omega} = (\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^N)$, $\boldsymbol{\omega}^1 \equiv \boldsymbol{w}_\ell^1, \ldots, \boldsymbol{\omega}^N \equiv \boldsymbol{w}_\ell^N$. We also denote by P_{ξ_ℓ} the true distribution of the data at discrete time ℓ, where due to (2a),

$$\xi_\ell = A^\ell \xi_0 + \sum_{k=0}^{\ell-1} A^k G w_{\ell-k-1} + \sum_{k=0}^{\ell-1} A^k r_{\ell-k-1}. \quad (6)$$

Then, it follows from **H1** and **H2** in Assumption 3.2 that the random states ξ_ℓ^i of the system realizations are independent and identically distributed. Leveraging this, our goal is to associate to each confidence $1-\beta$, an ambiguity radius $\psi_N(\beta)$ so that

$$\mathbb{P}(W_p(\widehat{P}_{\xi_\ell}^N, P_{\xi_\ell}) \leq \psi_N(\beta)) \geq 1-\beta, \quad (7)$$

where $\widehat{P}_{\xi_\ell}^N \equiv \widehat{P}_{\xi_\ell}^N(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell, \boldsymbol{v}_\ell)$ is a random measure as above, and P_{ξ_ℓ} is the true distribution. To achieve this, we decompose the confidence as the product of two factors:

$$1-\beta = (1-\beta_{\text{nom}})(1-\beta_{\text{ns}}). \quad (8)$$

The first factor (the nominal component "nom") is devoted to control the Wasserstein distance between the true empirical distribution and the true state distribution P_{ξ_ℓ}. While the second factor (the noise component "ns") is aimed at capturing the Wasserstein distance between the true- and the estimator-based-empirical distributions, which differ due to measurement noise. Assuming a uniform bound for the latter of these two Wasserstein distances (an assumption whose justification we address in the next section), we obtain the following desired ambiguity sets, which is our central result.

Theorem 4.1: ***(Ambiguity radius under noisy dynamics and observations).*** Consider data collected from N realizations of system (2) in accordance to Assumptions 3.1 and 3.2, and a confidence $1-\beta$. Let $\beta_{\text{nom}}, \beta_{\text{ns}} \in (0,1)$ satisfying (8) and assume that there is an $\widehat{\varepsilon}_N(\beta_{\text{ns}})$ so that

$$\mathbb{P}\big(W_p(\widehat{P}_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{v}_\ell), P_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega})) \leq \widehat{\varepsilon}_N(\beta_{\text{ns}})\big) \geq 1-\beta_{\text{ns}}, \quad (9)$$

for all $(\boldsymbol{z}, \boldsymbol{\omega}) \in B_\infty^{Nd}(\rho_{\xi_0}) \times B_\infty^{N\ell q}(\rho_w)$, where $P_{\xi_\ell}^N$ and $\widehat{P}_{\xi_\ell}^N$ are the true- and estimator-based-empirical distribution, respectively. Then, (7) is fulfilled with the ambiguity radius

$$\psi_N(\beta) := \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) + \widehat{\varepsilon}_N(\beta_{\text{ns}}), \quad (10)$$

$$\varepsilon_N(\beta, \rho) := \begin{cases} \left(\frac{\ln(C\beta^{-1})}{c}\right)^{\frac{1}{2p}} \frac{\rho}{N^{\frac{1}{2p}}}, & \text{if } p > d/2, \\ h^{-1}\left(\frac{\ln(C\beta^{-1})}{cN}\right)^{\frac{1}{p}} \rho, & \text{if } p = d/2, \\ \left(\frac{\ln(C\beta^{-1})}{c}\right)^{\frac{1}{d}} \frac{\rho}{N^{\frac{1}{d}}}, & \text{if } p < d/2, \end{cases} \quad (11)$$

where $h(x) := \frac{x^2}{(\ln(2+1/x))^2}$, $x > 0$, the constants C and c depend only on p and d, and

$$\rho_{\xi_\ell} := \sqrt{d}\|A^\ell\|\rho_{\xi_0} + \sqrt{q}\sum_{k=0}^{\ell-1}\|A^k G\|\rho_w. \quad (12)$$

V. Deviation between true and estimated ambiguity center

In this section, we show that for any confidence $1-\beta_{\text{ns}}$, there exists a Wasserstein distance $\widehat{\varepsilon}_N(\beta_{\text{ns}})$ so that the hypothesis (9) of Theorem 4.1 is valid. This needs to hold for all initial condition and noise values in the set $B_\infty^{Nd}(\rho_{\xi_0}) \times B_\infty^{N\ell q}(\rho_w)$, which contains the support of the corresponding random elements' distributions. To characterize the ambiguity radius $\widehat{\varepsilon}_N$ which guarantees (9) we exploit the subsequent intermediate results.

Lemma 5.1: ***(Distance between true- and estimator-based-empirical distribution).*** Let $(\boldsymbol{z}, \boldsymbol{\omega}) \in B_\infty^{Nd}(\rho_{\xi_0}) \times B_\infty^{N\ell q}(\rho_w)$ and consider the discrete distribution $P_{\xi_\ell}^N \equiv P_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega}) = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_\ell^i(z^i, \boldsymbol{\omega}^i)}$ and the empirical distribution $\widehat{P}_{\xi_\ell}^N \equiv \widehat{P}_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{v}_\ell) = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_\ell^i(z^i, \boldsymbol{\omega}^i, \boldsymbol{v}_\ell^i)}$, where $\boldsymbol{v}_\ell$ is the measurement noise of the realizations. Then,

$$W_p(\widehat{P}_{\xi_\ell}^N, P_{\xi_\ell}^N) \leq 2^{\frac{p-1}{p}}\mathfrak{M}_w + 2^{\frac{p-1}{p}}\left(\frac{1}{N}\sum_{i=1}^N (\mathfrak{E}^i)^p\right)^{\frac{1}{p}}, \quad (13)$$

where

$$\mathfrak{M}_w := \sqrt{d}\|F^\ell\|\rho_{\xi_0} + \sqrt{q}\sum_{k=0}^{\ell-1}\|F^k G\|\rho_w, \quad (14a)$$

$$\mathfrak{E}^i \equiv \mathfrak{E}^i(\boldsymbol{v}^i) := \sum_{k=0}^{\ell-1}\|F^k K\|\|v_{\ell-k-1}^i\|_1. \quad (14b)$$

We next provide some bounds for the random variables $\mathfrak{E}^i$.

Lemma 5.2: ***(Orlicz and L^p norm bounds for $\mathfrak{E}^i$).*** The random variables $\mathfrak{E}^i$ in (14b) satisfy

$$\|\mathfrak{E}^i\|_{\psi_p} \leq \mathfrak{C}_v := C_v r \sum_{k=0}^{\ell-1}\|F^k K\|, \quad (15a)$$

$$\|\mathfrak{E}^i\|_p \le \mathfrak{M}_v := M_v r \sum_{k=0}^{\ell-1} \|F^k K\|, \tag{15b}$$

$$\|\mathfrak{E}^i\|_p \ge \mathfrak{m}_v := m_v r^{\frac{1}{p}} \left(\sum_{k=0}^{\ell-1} \|F^k K\|^p \right)^{\frac{1}{p}}. \tag{15c}$$

We also rely on the following concentration-of-measure result around the mean of nonnegative independent random variables to bound the term $\left(\frac{1}{N}\sum_{i=1}^N (\mathfrak{E}^i)^p\right)^{\frac{1}{p}}$, and control the Wasserstein distance between the true- and the estimator-based empirical distribution.

Combining the results above with a generalization of the concentration inequality in [18, Theorem 3.1.1] for random variables with finite ψ_p norm, we obtain the main result of this section regarding the ambiguity center difference.

Proposition 5.3: ***(Guarantees for distance between true- and estimator-based-empirical distribution).*** Consider a confidence $1-\beta_{\rm ns}$ and let

$$\widehat{\varepsilon}_N(\beta_{\rm ns}) := 2^{\frac{p-1}{p}} \left(\mathfrak{M}_w + \mathfrak{M}_v + \mathfrak{M}_v \alpha_p^{-1} \left(\frac{\mathfrak{R}^2}{c' N} \ln \frac{2}{\beta_{\rm ns}} \right) \right), \tag{16}$$

with $\mathfrak{M}_w$, $\mathfrak{M}_v$ given by (14a), (15b),

$$\mathfrak{R} := \mathfrak{C}_v / \mathfrak{m}_v + 1/\ln 2, \tag{17}$$

and $\mathfrak{C}_v$, $\mathfrak{m}_v$ as in (15a), (15c). Then, (9) is fulfilled.

The combination of Theorem 4.1 and Proposition 5.3 yields the following result.

Corollary 5.4: ***(Explicit ambiguity radius characterization).*** Consider data collected from N realizations of system (2) in accordance to Assumptions 3.1 and 3.2, and a confidence $1-\beta$. Let $\beta_{\rm nom}, \beta_{\rm ns} \in (0,1)$ satisfying (8) and $\widehat{\varepsilon}_N$ as given by (16). Then, (7) holds with the ambiguity radius (10).

We conclude this section by leveraging the detectability assumption to quantify the size of the ambiguity radius as a function of the estimation error as the sampling horizon increases.

Proposition 5.5: ***(Noise ambiguity boundedness for detectable systems).*** Consider data collected from N realizations of system (2), a confidence $1-\beta$ as in (8), and let all Assumptions 3.1, 3.2, and 3.3 hold. Then, the ambiguity radius component $\widehat{\varepsilon}_N$ in (16) is uniformly bounded with respect to the sampling horizon size. In particular, there exists $\ell_0 \in \mathbb{N}$, so that for each $\ell \ge \ell_0$, the constants $\mathfrak{M}_w \equiv \mathfrak{M}_w(\ell)$, $\mathfrak{M}_v \equiv \mathfrak{M}_v(\ell)$, and $\mathfrak{R} \equiv \mathfrak{R}(\ell)$ in the expression for $\widehat{\varepsilon}_N$, given by (14a), (15b), and (14a), are uniformly upper bounded as

$$\mathfrak{M}_w \le \frac{1}{2}\sqrt{d}\rho_{\xi_0} + 2\sqrt{q} \sum_{k=0}^{\ell_0-1} \|F^k G\| \rho_w,$$

$$\mathfrak{M}_v \le 2 M_v r \sum_{k=0}^{\ell_0-1} \|F^k K\|,$$

$$\mathfrak{R} \le 2 \frac{C_v}{m_v} r^{\frac{p-1}{p}} \frac{\sum_{k=0}^{\ell_0-1} \|F^k K\|}{\left(\sum_{k=0}^{\ell_0-1} \|F^k K\|^p\right)^{\frac{1}{p}}}.$$

VI. Sensor placement for optimal tracking

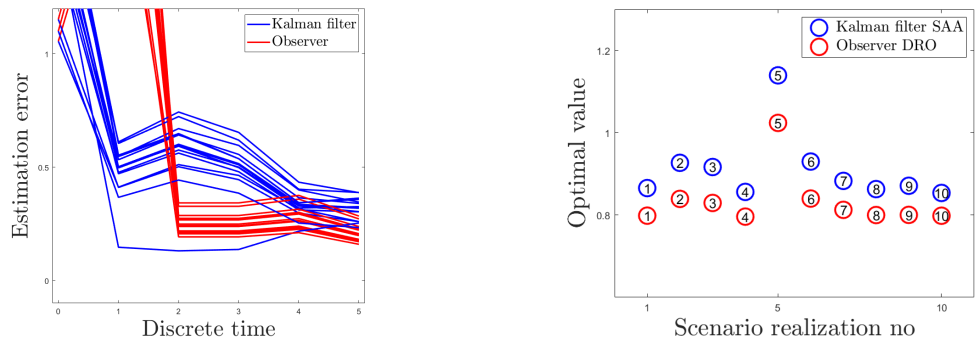
Consider a scenario where identical 1D particles track a known trajectory χ, but start from a different initial condition generated by a random distribution. Their continuous-time dynamics is

$$\dot{\xi}_1(t) = \xi_2(t),$$
$$\dot{\xi}_2(t) = \omega^2(\chi(t) - \xi_1(t)),$$

where $\omega > 0$ is the tracking frequency, and their position is measured at discrete time instants $k\tau$, $k = 0, 1, \ldots$, where τ is the sampling period. Denoting by $\widehat{A} = \left(\begin{smallmatrix} 0 & 1 \\ -\omega^2 & 0 \end{smallmatrix}\right)$ the system matrix and evaluating its exponential $e^{\widehat{A}t}$, we get the sampled version (2a) of the particle dynamics, with $A \equiv e^{\widehat{A}\tau}$, $r_k \equiv \int_0^\tau e^{\widehat{A}(\tau-s)} u(k\tau+s) ds$, $u(t) := (0 \quad \omega^2\chi(t))^\top$, and $w_k \equiv 0$. The output map (2b) is $H \equiv (0 \quad 1)$ and we consider nonzero measurement noise v_k.

We assume that noisy position measurements are available from each particle only during the first three discrete time instants $k = 0, 1, 2$. Our goal is to leverage these measurements and place a sensor at the location x which minimizes the expected squared distance from the position of the particle at $k = 5$. Namely, solve $\min_{x\in\mathbb{R}} \mathbb{E}_{P_{\xi_5}}\left[f(x, \xi_5)\right]$, with $f(x,\xi) = |x-\xi|^2$. For this we use the outputs to estimate the system state at $k = 3$, $\widehat{\xi}_3$, and subsequently its value at $k = 5$ exploiting knowledge on the dynamics, which give $\widehat{\xi}_5 = A^2\widehat{\xi}_3 + Ar_3 + r_4$. The unknown random initial condition is supported on the set $B_\infty^2(\rho_{\xi_0})$. In particular, we consider zero initial velocity and initial position distributed according to $0.1\delta_{-3\rho_{\xi_0}/4} + 0.9\mathcal{U}([\rho_{\xi_0}/2, \rho_{\xi_0}])$, with $\mathcal{U}$ denoting the uniform distribution. The noise distribution is $\mathcal{U}([\rho_v/2, \rho_v])$ and is captured by the same random variable for all measurements of the same particle, to model a random sensor offset for each measurement tuple (note that **H2** does not require independence between the v_k's of the same particle).

The problem is solved using two alternative approaches: 1) The particle state is estimated using a classical Kalman filter [20] and a Sample Average Approximation (SAA) of the expected cost is exploited to solve the optimization problem using the estimated states. 2) The particle state is estimated by a dead-beat observer [21, Remark 7.1.4] and a Distributionally Robust Optimization (DRO) problem is solved employing the ambiguity sets of Theorem 4.1 centered at the estimator-based empirical distribution. For each alternative approach, we use either a Kalman filter or an observer to estimate the states of N independent particle trajectories at $k = 5$, and exploit them to solve either the SAA or the DRO problem, respectively. In both cases, the estimation is performed in closed loop up to $k = 3$ using the measurements, and the obtained value is pushed forward two time steps to $k = 5$ by the dynamics, as described above. Since we assume only knowledge of the support of the initial conditions, we selected the covariance of an isotropic normal distribution with more than 0.99 probability inside $B_\infty^2(\rho_{\xi_0})$ for the Kalman filter initialization. Reasoning analogously



(a) (b)

Fig. 2: (a) shows the error evolution using the two alternative state estimators for 15 realizations of one experiment. Graph (b) depicts the expected cost associated to the optimizer found by solving the SAA using a classical Kalman filter (in blue), and the DRO optimization problem using a dead-beat observer (in red), for 10 independent realizations of the whole experiment.

for its noise parameters, we picked a covariance with the same $\|\cdot\|_{\psi_2}$ norm as the worst-case distribution in $[-\rho_v, \rho_v]$. For the simulation results we selected $\chi(t) = 0.2t$, $\omega = 2$, $\tau = \pi/9$, $\rho_\xi = 2$, $\rho_v = 0.2$, and the ambiguity radius $\psi_N = 0.6$, which is compliant with the noise ambiguity bound (16). For the given statistics of the initial condition and the noise, the observer gives a better estimate than the Kalman filter (cf. Figure 2(a)), leading also to an improved expected cost, when using the optimizer of the DRO solution compared to its SAA counterpart. This is verified for 10 realizations of the whole experiment with $N = 30$ independent trajectories sampled for each of them (cf. Figure 2(b)).

VII. Conclusions

We have provided an ambiguity set construction framework using partial-state measurements from dynamically varying random variables with an unknown initial condition distribution. Both the dynamics and the measurements are subject to noise whose probability model is not explicitly known. The ambiguity sets are built using an observer to estimate the full state of each realization and leveraging concentration of measure inequalities. For detectable systems, boundedness of the noise effect on the ambiguity radius is also established. Future research will include the consideration of time-varying and nonlinear dynamics, and improving the noise effect on the ambiguity radius.

VIII. Acknowledgments

This work was supported by the DARPA Lagrange program through award N66001-18-2-4027.

References

[1] A. Shapiro, D. Dentcheva, and A. Ruszczyski. *Lectures on Stochastic Programming: Modeling and Theory*, volume 16. SIAM, Philadelphia, PA, 2014.

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[3] A. Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):22582275, 2017.

[4] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. 44(2):565–600, 2019.

[5] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282, 2018.

[6] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Society, 2003.

[7] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707738, 2015.

[8] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, (1-2):115166, 2018.

[9] Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over Wasserstein balls. *arXiv preprint arXiv:1809.00210*, 2018.

[10] A. Hota, A. Cherukuri, and J. Lygeros. Data-driven chance constrained optimization under Wasserstein ambiguity sets. In *American Control Conference*, pages 1501–1506, Philadelphia, PA, USA, 2019.

[11] A. Cherukuri and J. Cortés. Cooperative data-driven distributionally robust optimization. *IEEE Transactions on Automatic Control*, 65(10), 2020. To appear.

[12] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[13] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

[14] D. Li and S. Martínez. Online data assimilation in distributionally robust optimization. In *IEEE Conf. on Decision and Control*, pages 1961–1966, Miami, FL, USA, 2018.

[15] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. M. Esfahani. Wasserstein distributionally robust Kalman filtering. pages 8474–8483, 2018.

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

[17] D. Boskos, J. Cortés, and S. Martinez. Data-driven ambiguity sets with probabilistic guarantees for dynamic processes. *IEEE Transactions on Automatic Control*, 2019. Submitted. Available at `https://arxiv.org/abs/1909.11194`.

[18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[19] R. B. Ash. *Real Analysis and Probability*. Academic Press, 1972.

[20] B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice Hall, 1979.

[21] E. D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*. Springer, 1998.