High-Confidence Data-Driven Ambiguity Sets for Time-Varying Linear Systems

Dimitris Boskos Jorge Cortés Sonia Martínez

***Abstract*—This paper builds Wasserstein ambiguity sets for the unknown probability distribution of dynamic random variables leveraging noisy partial-state observations. The constructed ambiguity sets contain the true distribution of the data with quantifiable probability and can be exploited to formulate robust stochastic optimization problems with out-of-sample guarantees. We assume the random variable evolves in discrete time under uncertain initial conditions and dynamics, and that noisy partial measurements are available. All random elements have unknown probability distributions and we make inferences about the distribution of the state vector using several output samples from multiple realizations of the process. To this end, we leverage an observer to estimate the state of each independent realization and exploit the outcome to construct the ambiguity sets. We illustrate our results in an economic dispatch problem involving distributed energy resources over which the scheduler has no direct control.**

***Index Terms*—Distributional uncertainty, Wasserstein ambiguity sets, stochastic systems, state estimation**

I. Introduction

Decisions under uncertainty are ubiquitous in a wide range of engineering applications. Faced with complex systems that include components with probabilistic models, such decisions seek to provide rigorous solutions with quantifiable guarantees in hedging against uncertainty. In practice, the designer makes inferences about uncertain elements based on collected data and exploits them to formulate data-driven stochastic optimization problems. This decision-making paradigm has found applications in finance, communications, control, medicine, and machine learning. Recent research focuses on how to retain high-confidence guarantees for the optimization problems under plausible variations of the data. To this end, distributionally robust optimization (DRO) formulations evaluate the optimal worst-case performance over an ambiguity set of probability distributions that contains the true one with high confidence. Such ambiguity sets are typically constructed under the assumption that data are generated from a static distribution and can be measured in a direct manner. In this paper we significantly expand on the class of scenarios for which reliable ambiguity sets can be constructed. We consider scenarios where the random variable is dynamic and partial measurements, corrupted by noise, are progressively collected from its evolving distribution. In our analysis, we exploit the underlying dynamics and study how the probabilistic properties of the noise affect the ambiguity set size while maintaining the same guarantees.

Literature review: Optimal decision problems in the face of uncertainty, like expected-cost minimization and chance-constrained optimization, are the cornerstones of stochastic programming [42]. Distributionally robust versions of stochastic optimization problems [2], [5], [41] carry out a worst-case optimization over all possibilities from an ambiguity set of probability distributions. This is of particular importance in data-driven scenarios where the unknown distributions of the random variables are inferred in an approximate manner using a finite amount of data [3]. To hedge this uncertainty, optimal transport ambiguity sets have emerged as a promising tool. These sets typically group all distributions up to some distance from the empirical approximation in the Wasserstein metric [45]. There are several reasons that make this metric a popular choice among the distances between probability distributions, particularly, for data-driven problems. Most notably, the Wasserstein metric penalizes horizontal dislocations between distributions and provides ambiguity sets that have finite-sample guarantees of containing the true distribution and lead to tractable optimization problems. This has rendered the convergence of empirical measures in the Wasserstein distance an ongoing active research area [16], [17], [19], [27], [46], [47]. Towards the exploitation of Wasserstein ambiguity sets for DRO problems, the work [18] introduces tractable reformulations with finite-sample guarantees, further exploited in [12], [26] to deal with distributionally robust chance-constrained programs. The work [14] develops distributed optimization algorithms using Wasserstein balls, while optimal transport ambiguity sets have recently been connected to regularization for machine learning [4], [21], [39]. The paper [31] exploits Wasserstein balls to robustify data-driven online optimization algorithms, and [40] leverages them for the design of distributionally robust Kalman filters. Further applications of Wasserstein ambiguity sets include the synthesis of robust control policies for Markov decision processes [49] and their data-driven extensions [50], and regularization for stochastic predictive control algorithms [15]. Several recent works have also devoted attention to distributionally robust problems in power systems control, including optimal power flow [24], [28] and economic dispatch [48], [34], [38]. Time-varying aspects of Wasserstein ambiguity sets are considered in our previous work: in [29] for dynamic traffic models, in [30] for online learning of unknown dynamical environments, in [9], which constructs ambiguity balls using progressively assimilated dynamic data for processes with random initial conditions that evolve under deterministic dynamics, and in [11],

A preliminary version of this work appeared as [8] in the American Control Conference. This work was supported by the DARPA Lagrange program through award N66001-18-2-4027.

Dimitris Boskos is with the Delft Center for Systems and Control, Delft University of Technology d.boskos@tudelft.nl

Jorge Cortés and Sonia Martínez are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, {cortes,soniamd}@ucsd.edu.

which studies the propagation of ambiguity bands under hyperbolic PDE dynamics. In contrast, in the present work, the state distribution does not evolve deterministically due to the presence of random disturbances, which together with output measurements that are corrupted by noise, generate additional stochastic elements that make challenging the quantification of the ambiguity set guarantees.

Statement of contributions: Our contributions revolve around building Wasserstein ambiguity sets with probabilistic guarantees for dynamic random variables when we have no knowledge of the probability distributions of their initial condition, the disturbances in their dynamics, and the measurement noise. To this end, our first contribution estimates the states of several process realizations from output samples and exploits these estimates to build a suitable empirical distribution as the center of an ambiguity ball. Our second contribution is the exploitation of concentration of measure results to quantify the radius of this ambiguity ball so that it provably contains the true state distribution with high probability. To achieve this, we break the radius into nominal and noise components. The nominal component captures the deviation between the true distribution and the empirical distribution formed by the state realizations. The noise component captures the deviation between the empirical distribution and the center of our ambiguity ball. To quantify the latter, we carefully evaluate the impact of the estimation error, which due to the measurement noise, does not have a compactly supported distribution like the internal uncertainty and requires a separate analysis. Our third contribution is the extension of these results to obtain *simultaneous* guarantees about ambiguity sets that are built along finite time horizons, instead of at isolated time instances. The fourth contribution is to generalize a concentration inequality around the mean of sufficiently light-tailed independent random variables, which enables us to obtain tighter results when analyzing the effect of the estimation error. Our last contribution is the validation of the results in simulation for a distributionally robust economic dispatch problem, for which we further provide a tractable reformulation. We stress that our general objective revolves around the robust uncertainty quantification (i.e., distributional inference) problem at hand, without having DRO as a necessary end-goal. Further, our approach 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 also a known Gaussian random variable. Here, instead, we are interested to infer the unknown state distribution from data collected by multiple realizations of the dynamics. For each such realization, we use an observer since we have no concrete knowledge of the state and noise random models to directly invoke optimal filtering techniques. In the online version [10] of this manuscript, we provide explicit constants for several of the presented concentration of measure inequalities which, to the best of our knowledge, have not been delineated in the literature. These results are not essential to keep the theoretical presentation self-contained and are omitted due to space constraints.

II. Preliminaries

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

Notation: We denote by $\|\cdot\|_p$ the pth norm in $\mathbb{R}^n$, $p \in [1,\infty]$, using also the notation $\|\cdot\| \equiv \|\cdot\|_2$ for the Euclidean norm. We denote by $B_p^n(\rho)$ the ball of center zero and radius ρ in $\mathbb{R}^n$ with the pth norm, $p \in [1,\infty]$. The inner product of two vectors $a, b \in \mathbb{R}^n$ is denoted by $\langle a, b\rangle$ and the Khatri-Rao product [35] of $\boldsymbol{a} \equiv (a^1,\dots,a^d) \in \mathbb{R}^d$ and $\boldsymbol{b} \equiv (b^1,\dots,b^d) \in \mathbb{R}^{dn}$, with each b^i belonging to $\mathbb{R}^n$, is $\boldsymbol{a}*\boldsymbol{b} := (a^1b^1,\dots,a^db^d) \in \mathbb{R}^{dn}$. We use the notation $[n_1 : n_2]$ for the set of integers $\{n_1, n_1+1,\dots,n_2\} \subset \mathbb{N}\cup\{0\} =: \mathbb{N}_0$. The interpretation of a vector in $\mathbb{R}^n$ as an $n \times 1$ matrix should be clear form the context (this avoids writing double transposes). The diameter of a set $S \subset \mathbb{R}^n$ with the pth norm is defined as $\mathrm{diam}_p(S) := \sup\{\|x-y\|_p \,|\, x, y \in S\}$ 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 $B \subset \Omega$, $\mathbf{1}_B$ 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 $x \notin B$.

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 $p \geq 1$, $\mathcal{P}_p(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) \,|\, \int_{\mathbb{R}^d} \|x\|^p d\mu < \infty\}$ is the set of probability measures in $\mathcal{P}(\mathbb{R}^d)$ with finite 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 couplings between μ and ν, i.e., 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. For a random variable X with distribution μ we also denote $\mathrm{supp}(X) \equiv \mathrm{supp}(\mu)$. We denote the product of the distributions μ in $\mathbb{R}^d$ and ν in $\mathbb{R}^r$ by the distribution $\mu \otimes \nu$ in $\mathbb{R}^d \times \mathbb{R}^r$. The convolution $\mu \star \nu$ of the distributions μ and ν on $\mathbb{R}^d$ is the image of the measure $\mu \otimes \nu$ on $\mathbb{R}^d \times \mathbb{R}^d$ under the mapping $(x,y) \mapsto x + y$; equivalently, $\mu \star \nu(B) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbf{1}_B(x+y)\mu(dx)\nu(dy)$ for any $B \in \mathcal{B}(\mathbb{R}^d)$ (c.f. [7, Pages 207, 208]). Given a measurable space $(\Omega, \mathcal{F})$, an exponent $p \geq 1$, the convex function $\mathbb{R} \ni x \mapsto \psi_p(x) := e^{x^p} - 1$, 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. [44, 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 pth moment, i.e., the classical norm in $L^p(\Omega) \equiv L^p(\Omega; P_X)$, where P_X is the distribution of X. The interpretation of $\|\cdot\|_p$ as the pth 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 which follows from Fubini's theorem [1, 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 $(x,y) \mapsto g(x,y)$ be integrable. Assume that $\mathbb{E}[g(x,Y)] \geq k(x)$ for some integrable function k and all $x \in K$ with $\mathrm{supp}(X) \subset K \subset \mathbb{R}^{n_1}$. 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 with an *unknown* distribution P_ξ. 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 $\xi^1,\dots,\xi^N$ of P_ξ to optimize a sample average approximation of the expected value of f, one can instead consider the DRO problem

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

of evaluating the worst-case expectation over some *ambiguity set* $\mathcal{P}^N$ of probability measures. This helps the designer robustify the decision against plausible variations of the data, which can play a significant role when the number of samples is limited. Different approaches exist to construct the ambiguity set $\mathcal{P}^N$ so that it contains 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.

To make things precise, consider data generated by a discrete-time system

$$\xi_{k+1} = A_k\xi_k + G_k w_k, \quad \xi_k \in \mathbb{R}^d, \quad w_k \in \mathbb{R}^q, \tag{3a}$$

with linear output

$$\zeta_k = H_k\xi_k + v_k, \quad \zeta_k \in \mathbb{R}^r. \tag{3b}$$

The initial condition ξ_0 and the noises w_k and v_k, $k\in\mathbb{N}_0$ in the dynamics and the measurements, respectively, are random variables with an *unknown* distribution. We seek to build an ambiguity set for the state distribution at certain time $\ell \in \mathbb{N}$, by collecting data up to time ℓ 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, see e.g. [51]. The time-dependent matrices in the dynamics (3) widen the applicability of the results, since they can capture the linearization of nonlinear systems along trajectories or the sampled-data analogues of continuous-time systems under irregular sampling, even if the latter are linear and time invariant. To formally describe the problem, we consider a 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 (3), output samples $\zeta_0^i,\dots,\zeta_\ell^i$ are collected over the discrete time instants of the sampling horizon $[0:\ell]$.

According to this assumption, the measurements of all realizations are collected over the same time window $[0:\ell]$. To obtain quantifiable characterizations of the ambiguity sets, we require 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 (3a) with associated outputs given by (3b), and noise elements w_k^i, v_k^i, $k\in\mathbb{N}_0$. We assume the following:

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 $p\geq 1$.

Remark 3.3: ***(Bounded ψ_p-Orlicz/L_p-norm ratio).*** By definition, ψ_p-Orlicz norms can become significantly larger than L_p norms for random variables with heavier tails. Thus, over an infinite sequence of random variables $\{X_k\}$, the ratio $\|X_k\|_{\psi_p}/\|X_k\|_p$ may grow unbounded. We exclude this by assuming that C_v and m_v are either positive or zero simultaneously, in which case we set $C_v/m_v := 0$. □

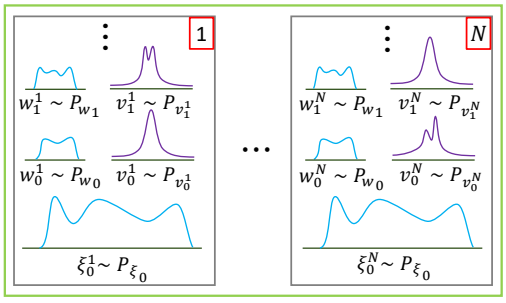


Fig. 1. Illustration of the probabilistic models for the random variables in the dynamics and observations according to Assumption 3.2.

A direct approach to build the ambiguity set using the measurements of the trajectories at time ℓ would be severely limited, since the output map is in general not invertible. In such case, the inverse image of each measurement is the translation of a subspace, whose location is further obscured by the measurement noise. As a consequence, candidate states for a generated output sample may lie at an arbitrary distance apart, which could only be bounded by making additional hypotheses about the support of the state distribution. Instead, despite the lack of full-state information, we aim to leverage the system dynamics to estimate the state from the whole assimilated output trajectory. To guarantee some boundedness notion for the state estimation errors over arbitrary evolution horizons, we make the following assumption.

Assumption 3.4: ***(Detectability/uniform observability).*** System (3) satisfies one of the following properties:
(i) It is time invariant and the pair (A, H) (with $A \equiv A_k$ and $H \equiv H_k$) is detectable.
(ii) It is uniformly observable, i.e., for some $t \in \mathbb{N}$, the observability Gramian

$$\mathcal{O}_{k+t,k} := \sum_{i=k}^{k+t} \Phi_{i,k}^\top H_i^\top H_i \Phi_{i,k}$$

satisfies $\mathcal{O}_{k+t,k} \succeq bI$ for certain $b > 0$ and all $k \in \mathbb{N}_0$, where we denote $\Phi_{k+s,k} := A_{k+s-1} \cdots$
$A_{k+1}A_k$. Further, all system matrices are uniformly bounded and the singular values of A_k and the norms of $\|H_k\|$ are uniformly bounded below.

Problem statement: Under Assumptions 3.1 and 3.2 on the measurements and distributions of N realizations of the system (3), we seek to construct an estimator $\widehat{\xi}_\ell^i(\zeta_0^i, \ldots, \zeta_\ell^i)$ for the state of each realization and build an ambiguity set for the state distribution at time ℓ with probabilistic guarantees. Further, under Assumption 3.4 on the system's detectability/uniform observability properties, we aim to characterize the effect of the estimation precision on the accuracy of the ambiguity sets.

We proceed to address the 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 construction of the ambiguity set. To obtain the probabilistic guarantees, we leverage concentration inequalities to bound the distance between the updated empirical distribution and the true state distribution with high confidence. To this end, we further quantify the increase of the ambiguity radius due to the noise. We also study the beneficial effect on the ambiguity radius of detectability/uniform observability for arbitrarily long evolution horizons in Section V.

IV. State-estimator based ambiguity sets

We address here the question of how to construct an ambiguity set at certain time instant ℓ, when samples are collected from (3) according to Assumption 3.1. If we had access to N independent full-state samples $\xi_\ell^1, \ldots, \xi_\ell^N$ from the distribution of ξ at ℓ, we could construct an ambiguity ball in the Wasserstein metric W_p centered at the empirical distribution (1) with $\xi^i \equiv \xi_\ell^i$ and containing the true distribution with high confidence. In particular, for any confidence $1 - \beta > 0$, it is possible, cf. [18, 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 at time ℓ. The dynamics of the observer, initialized at zero, is given by

$$\widehat{\xi}_{k+1} = A_k \widehat{\xi}_k + K_k(H_k \widehat{\xi}_k - \zeta_k), \qquad \widehat{\xi}_0 = 0, \tag{4}$$

where each K_k is a nonzero gain matrix. Using the corresponding estimates from system (4) for the independent realizations of (3a), we define the *(dynamic) estimator-based empirical distribution*

$$\widehat{P}_{\xi_k}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\widehat{\xi}_k^i}, \tag{5}$$

Denoting by $e_k := \xi_k - \widehat{\xi}_k$ the error between (3a) and the observer (4), the error dynamics is $e_{k+1} = F_k e_k + G_k w_k + K_k v_k$, $e_0 = \xi_0$, where $F_k := A_k + K_k H_k$ and ξ_0 is the initial condition of (3a). In particular,

$$\begin{aligned} e_k = \Psi_k \xi_0 + \sum_{\kappa=1}^k \big(& \Psi_{k,k-\kappa+1} G_{k-\kappa} w_{k-\kappa} \\ & + \Psi_{k,k-\kappa+1} K_{k-\kappa} v_{k-\kappa} \big) \end{aligned} \tag{6}$$

for all $k \geq 1$, where $\Psi_{k+s,k} := F_{k+s-1} \cdots F_{k+1} F_k$, $\Psi_{k,k} := I$ and $\Psi_k := \Psi_{k,0}$. 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 (5). In what follows, we leverage concentration of measure results to identify an ambiguity radius $\psi_N(\beta)$ so that the resulting Wasserstein ball contains the true distribution with a given confidence $1-\beta$. Note that even if a distributionally robust framework is not employed, replacing the empirical distribution by the estimator empirical distribution in (5) does no longer guarantee consistency, in the sense that the estimator empirical distribution *does not necessarily converge (weakly) to the true distribution*. Hence, there is also no indication that the solution to the associated estimator Sample Average Approximation (SAA) problem, i.e., to

$$\inf_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N f(x, \xi_\ell^i)$$

with ξ_ℓ^i replaced by $\widehat{\xi}_\ell^i$, will be a consistent estimator of the solution to the nominal stochastic optimization problem. This is a fundamental limitation that is justified by the fact that, in general, the estimation error is dependent on the state realization, i.e., it has a variable distribution when conditioned on the state and the internal noise, and so its effect cannot be easily reversed (this may only be possible in rather degenerate cases, e.g., one has access to full-sate samples and the measurement noise is known).

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 by the output noise $\boldsymbol{v}_k^i \equiv (v_0^i, \dots, v_{k-1}^i)$. Using the compact notation $\boldsymbol{\xi}_0 \equiv (\xi_0^1, \dots, \xi_0^N)$, $\boldsymbol{w}_k \equiv (\boldsymbol{w}_k^1, \dots, \boldsymbol{w}_k^N)$, and $\boldsymbol{v}_k \equiv (\boldsymbol{v}_k^1, \dots, \boldsymbol{v}_k^N)$ for the corresponding initial conditions, dynamics noise, and output noise of all realizations, respectively, we can denote the empirical 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 $\boldsymbol{z} = (z^1, \dots, z^N)$, $z^1 \equiv \xi_0^1, \dots, z^N \equiv \xi_0^N$, and $\boldsymbol{\omega} = (\boldsymbol{\omega}^1, \dots, \boldsymbol{\omega}^N)$, $\boldsymbol{\omega}^1 \equiv \boldsymbol{w}_\ell^1, \dots, \boldsymbol{\omega}^N \equiv \boldsymbol{w}_\ell^N$. We also denote by P_{ξ_ℓ} the true distribution of the data at discrete time ℓ, where from (3a),

$$\xi_\ell = \Phi_\ell \xi_0 + \sum_{k=1}^{\ell} \Phi_{\ell,\ell-k+1} G_{\ell-k} w_{\ell-k}, \tag{7}$$

where $\Phi_\ell := \Phi_{\ell,0}$ and $\Phi_{\ell,\ell} := I$ (and with $\Phi_{k+\delta k,k}$ defined in Assumption 3.4). 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

$$\mathbb{P}(W_p(\widehat{P}_{\xi_\ell}^N, P_{\xi_\ell}) \le \psi_N(\beta)) \ge 1 - \beta. \tag{8}$$

To achieve this, we decompose the confidence as the product of two factors:

$$1 - \beta = (1 - \beta_{\text{nom}})(1 - \beta_{\text{ns}}). \tag{9}$$

The first factor (the nominal component "nom") is exploited to control the Wasserstein distance between the empirical distribution and the true state distribution P_{ξ_ℓ}. The purpose of the second factor (the noise component "ns") is to bound the Wasserstein distance between the empirical- and the estimator-based-empirical distributions, which is affected by the measurement noise. Using this decomposition, our strategy to get (8) builds on further breaking the ambiguity radius as

$$\psi_N(\beta) := \varepsilon_N(\beta_{\text{nom}}) + \widehat{\varepsilon}_N(\beta_{\text{ns}}). \tag{10}$$

We exploit what is known [9] for the no-noise case to bound the *nominal ambiguity radius* $\varepsilon_N(\beta_{\text{nom}})$ with confidence $1 - \beta_{\text{nom}}$. Moreover, we bound the *noise ambiguity radius* $\widehat{\varepsilon}_N(\beta_{\text{ns}})$ with confidence $1-\beta_{\text{ns}}$. This latter radius corresponds to the impact on distributional uncertainty of the internal and measurement noise. In the next sections we present the precise individual bounds for these terms and then combine them to obtain the overall ambiguity radius.

A. Nominal ambiguity radius

According to Assumption 3.2, the initial condition and internal noise distributions are compactly supported, and hence, the same holds also for the state distribution along time. We will therefore use the following result, that is focused on compactly supported distributions and bounds the distance between the true and empirical distribution for any fixed confidence level.

Proposition 4.1: ***(Nominal ambiguity radius [9, Corollary 3.3]).*** Consider a sequence $\{X_i\}_{i\in\mathbb{N}}$ of i.i.d. $\mathbb{R}^d$-valued random variables with a compactly supported distribution μ. Then for any $p \ge 1$, $N \ge 1$, and confidence $1-\beta$ with $\beta \in (0,1)$, we have $\mathbb{P}(W_p(\mu^N, \mu) \le \varepsilon_N(\beta, \rho)) \ge 1 - \beta$, where

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

$\mu^N := \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$, $\rho := \frac{1}{2}\text{diam}_\infty(\text{supp}(\mu))$, $h(x) := \frac{x^2}{(\ln(2+1/x))^2}$, $x > 0$, and the constants C and c depend only on p and d.

This result shows how the nominal ambiguity radius depends on the size of the distribution's support, the confidence level, and the number of samples, and is based on recent concentration of measure inequalities from [19].

Remark 4.2: The determination of the constants C and c in (11) for the whole spectrum of data dimensions d and Wasserstein exponents p is a particularly cumbersome task. Nevertheless, in the online version [10, Section 8.2], we provide some alternative concentration of measure results and use them to obtain explicit formulas for these constants when $d > 2p$. In particular, the constants in the third expression in (11) can be chosen as $C := \frac{C_\star^d}{2\sqrt{d}^d}$ and $c := \frac{1}{2^d\sqrt{d}^d}$, where

$$C_\star := \sqrt{d} 2^{(d-2)/(2p)} \left(\frac{1}{1-2^{p-d/2}} + \frac{1}{1-2^{-p}}\right)^{1/p}.$$

Recent work [4], [20], [6] informs the ambiguity radius by the optimization problem at hand to ameliorate its slow decay with the number of samples. However, the resulting ambiguity balls often contain the true distribution with low probability, which may fail to provide guarantees when solving multiple DRO problems using the same data, as is done for instance in model predictive control [37], [25]. □

B. Noise ambiguity radius

In this section, we quantify the noise ambiguity radius $\widehat{\varepsilon}_N(\beta_{\text{ns}})$ for any prescribed confidence $1 - \beta_{\text{ns}}$. We first give a result that uniformly bounds the distance between the empirical and estimator-based-empirical distributions with prescribed confidence for all values of the initial condition and the internal noise from the set $B_\infty^{Nd}(\rho_{\xi_0}) \times B_\infty^{N\ell q}(\rho_w)$, which contains the support of their joint distribution (and hence all their possible realizations). For the results of this section, the initial condition and the internal noise are interpreted as deterministic quantities, as discussed above.

Lemma 4.3: ***(Distance between empirical & 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 $P_{\xi_\ell}^N \equiv P_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega})$ and the empirical distribution $\widehat{P}_{\xi_\ell}^N \equiv \widehat{P}_{\xi_\ell}^N(\boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{v}_\ell)$, where $\boldsymbol{v}_\ell$ is the measurement noise of the realizations. Then,

$$W_p(\widehat{P}_{\xi_\ell}^N, P_{\xi_\ell}^N) \le 2^{\frac{p-1}{p}} \mathfrak{M}_w + 2^{\frac{p-1}{p}} \Big(\frac{1}{N}\sum_{i=1}^N (\mathfrak{E}^i)^p\Big)^{\frac{1}{p}}, \tag{12a}$$

where

$$\mathfrak{M}_w := \sqrt{d}\|\Psi_\ell\|\rho_{\xi_0} + \sqrt{q}\sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}G_{\ell-k}\|\rho_w, \quad (12b)$$

$$\mathfrak{E}^i \equiv \mathfrak{E}(\boldsymbol{v}^i) := \sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|\|v^i_{\ell-k}\|_1. \quad (12c)$$

The next result gives bounds for the norms of the random variables $\mathfrak{E}^i$ in Lemma 4.3.

Lemma 4.4: ***(Orlicz- & L^p-norm bounds for $\mathfrak{E}^i$).*** The random variables $\mathfrak{E}^i$ in (12c) satisfy

$$\|\mathfrak{E}^i\|_p \le \mathfrak{M}_v := M_v r\sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|, \quad (13a)$$

$$\|\mathfrak{E}^i\|_{\psi_p} \le \mathfrak{C}_v := C_v r\sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|, \quad (13b)$$

$$\|\mathfrak{E}^i\|_p \ge \mathfrak{m}_v := m_v r^{\frac{1}{p}}\Big(\sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|^p\Big)^{\frac{1}{p}}, \quad (13c)$$

with m_v, M_v, and C_v as given in **H4**.

The proofs of both results above are given in the Appendix. We further rely on the following concentration of measure result around the mean of nonnegative independent random variables, whose proof is also in the Appendix, to bound the term $\big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\big)^{\frac{1}{p}}$, and control the Wasserstein distance between the empirical and the estimator-based-empirical distribution.

Proposition 4.5: ***(Concentration around pth mean).*** Let $X_1,\dots,X_N$ be scalar, nonnegative, independent random variables with finite ψ_p norm and $\mathbb{E}[X_i^p]=1$. Then,

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N X_i^p\Big)^{\frac{1}{p}} - 1 \ge t\bigg) \le 2\exp\Big(-\frac{c'N}{R^2}\alpha_p(t)\Big), \quad (14)$$

for every $t\ge 0$, with $c'=1/10$, $R := \max_{i\in[1:N]}\|X_i\|_{\psi_p} + 1/\ln 2$, and

$$\alpha_p(s) := \begin{cases} s^2, \text{ if } s\in[0,1], \\ s^p, \text{ if } s\in(1,\infty). \end{cases} \quad (15)$$

Combining the results above, we obtain the main result of this section regarding the ambiguity center difference.

Proposition 4.6: ***(Distance guarantee between empirical & estimator-based-empirical distribution).*** Consider a confidence $1-\beta_{\text{ns}}$ and let

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

with $\mathfrak{M}_w$, $\mathfrak{M}_v$ given by (12b), (13a),

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

and $\mathfrak{C}_v$, $\mathfrak{m}_v$ as in (13b), (13c). Then, for all $(\boldsymbol{z},\boldsymbol{\omega}) \in B^{Nd}_\infty(\rho_{\xi_0})\times B^{N\ell q}_\infty(\rho_w)$, we have

$$\mathbb{P}\big(W_p(\widehat{P}^N_{\xi_\ell}(\boldsymbol{z},\boldsymbol{\omega},\boldsymbol{v}_\ell), P^N_{\xi_\ell}(\boldsymbol{z},\boldsymbol{\omega})) \le \widehat{\varepsilon}_N(\beta_{\text{ns}})\big) \ge 1-\beta_{\text{ns}}. \quad (18)$$

Proof: For each i, the random variable $X_i := \mathfrak{E}^i/\|\mathfrak{E}^i\|_p$ satisfies $\|X_i\|_p = 1$. Thus, we obtain from Proposition 4.5 that

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N\Big(\frac{\mathfrak{E}^i}{\|\mathfrak{E}^i\|_p}\Big)^p\Big)^{\frac{1}{p}} - 1 \ge t\bigg) \le 2\exp\Big(-\frac{c'N}{R^2}\alpha_p(t)\Big),$$

where $R = \max_{i\in[1:N]}\big\|\mathfrak{E}^i/\|\mathfrak{E}^i\|_p\big\|_{\psi_p} + 1/\ln 2$. From (13b), (13c), and (17), we deduce $\mathfrak{R}\ge R$, and thus,

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N\Big(\frac{\mathfrak{E}^i}{\|\mathfrak{E}^i\|_p}\Big)^p\Big)^{\frac{1}{p}} - 1 \ge t\bigg) \le 2\exp\Big(-\frac{c'N}{\mathfrak{R}^2}\alpha_p(t)\Big).$$

Now, it follows from (13a) that

$$\mathfrak{M}_v\bigg(\frac{1}{N}\sum_{i=1}^N\Big(\frac{\mathfrak{E}^i}{\|\mathfrak{E}^i\|_p}\Big)^p\bigg)^{\frac{1}{p}} - \mathfrak{M}_v \ge \Big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}} - \mathfrak{M}_v.$$

Thus, we deduce

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}} - \mathfrak{M}_v \ge \mathfrak{M}_v t\bigg) \le 2\exp\Big(-\frac{c'N}{\mathfrak{R}^2}\alpha_p(t)\Big),$$

or, equivalently, that

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}} \ge \mathfrak{M}_v + s\bigg) \le 2\exp\Big(-\frac{c'N}{\mathfrak{R}^2}\alpha_p\Big(\frac{s}{\mathfrak{M}_v}\Big)\Big). \quad (19)$$

To establish (18), it suffices by Lemma 4.3 to show that

$$\mathbb{P}\bigg(2^{\frac{p-1}{p}}\mathfrak{M}_w + 2^{\frac{p-1}{p}}\Big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}} \le \widehat{\varepsilon}_N(\beta_{\text{ns}})\bigg) \ge 1-\beta_{\text{ns}}.$$

By the definition of $\widehat{\varepsilon}_N$ and exploiting that it is strictly decreasing with β_{ns}, it suffices to prove

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}} < \mathfrak{M}_v + \mathfrak{M}_v\alpha_p^{-1}\Big(\frac{\mathfrak{R}^2}{c'N}\ln\frac{2}{\beta_{\text{ns}}}\Big)\bigg) \ge 1-\beta_{\text{ns}}.$$

Setting $\tau = \alpha_p^{-1}\big(\frac{\mathfrak{R}^2}{c'N}\ln\frac{2}{\beta_{\text{ns}}}\big)$, we equivalently need to show

$$\mathbb{P}\bigg(\Big(\frac{1}{N}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}} \ge \mathfrak{M}_v + \tau\mathfrak{M}_v\bigg) \le \beta_{\text{ns}},$$

which follows by (19) with $s=\tau\mathfrak{M}_v$. ∎

C. Overall ambiguity set

Here we combine the results from Sections IV-A and IV-B to obtain the ambiguity set of the state distribution in the following result.

Theorem 4.7: ***(Ambiguity set under noisy dynamics & observations).*** Consider data collected from N realizations of system (3) in accordance to Assumptions 3.1 and 3.2, a confidence $1-\beta$, and let $\beta_{\text{nom}}, \beta_{\text{ns}} \in (0,1)$ satisfying (9). Then, the guarantee (8) holds, where $\psi_N(\beta)$ is given in (10)

and its components $\varepsilon_N(\beta_{\text{nom}}) \equiv \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})$ and $\widehat{\varepsilon}_N(\beta_{\text{ns}})$ are given by (11) and (16), respectively, with

$$\rho_{\xi_\ell} := \sqrt{d}\|\Phi_\ell\|\rho_{\xi_0} + \sqrt{q}\sum_{k=1}^{\ell}\|\Phi_{\ell,\ell-k+1}G_{\ell-k}\|\rho_w. \quad (20)$$

Proof: Due to (10) and the triangle inequality for W_p,

$$\{W_p(\widehat{P}^N_{\xi_\ell}, P_{\xi_\ell}) \le \psi_N(\beta)\} \supset \{W_p(\widehat{P}^N_{\xi_\ell}, P^N_{\xi_\ell}) \le \widehat{\varepsilon}_N(\beta_{\text{ns}})\} \cap \{W_p(P^N_{\xi_\ell}, P_{\xi_\ell}) \le \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})\}.$$

Thus, to show (8), it suffices to show that

$$\mathbb{E}\Big[\mathbf{1}_{\{W_p(\widehat{P}^N_{\xi_\ell}, P^N_{\xi_\ell}) - \widehat{\varepsilon}_N(\beta_{\text{ns}}) \le 0\}} \times \mathbf{1}_{\{W_p(P^N_{\xi_\ell}, P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \le 0\}}\Big] \ge 1 - \beta. \quad (21)$$

We therefore exploit Lemma 2.1 with the random variable $X \equiv (\boldsymbol{\xi}_0, \boldsymbol{w}_\ell)$, taking values in the compact set $K \equiv B^{Nd}_\infty(\rho_{\xi_0}) \times B^{N\ell q}_\infty(\rho_w)$, the random variable $Y \equiv \boldsymbol{v}_\ell \in \mathbb{R}^{N\ell r}$, and $g(X, Y) \equiv g(\boldsymbol{\xi_0}, \boldsymbol{w}_\ell, \boldsymbol{v}_\ell)$, where

$$g(\boldsymbol{\xi_0}, \boldsymbol{w}_\ell, \boldsymbol{v}_\ell) := \mathbf{1}_{\{W_p(P^N_{\xi_\ell}(\boldsymbol{\xi_0}, \boldsymbol{w}_\ell), P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \le 0\}} \times \mathbf{1}_{\{W_p(\widehat{P}^N_{\xi_\ell}(\boldsymbol{\xi_0}, \boldsymbol{w}_\ell, \boldsymbol{v}_\ell), P^N_{\xi_\ell}(\boldsymbol{\xi_0}, \boldsymbol{w}_\ell)) - \widehat{\varepsilon}_N(\beta_{\text{ns}}) \le 0\}}.$$

Due to (18), $\mathbb{E}\Big[\mathbf{1}_{\{W_p(\widehat{P}^N_{\xi_\ell}(\boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{v}_\ell), P^N_{\xi_\ell}(\boldsymbol{z}, \boldsymbol{\omega})) - \widehat{\varepsilon}_N(\beta_{\text{ns}}) \le 0\}}\Big] \ge 1 - \beta_{\text{ns}}$ for any $x = (\boldsymbol{z}, \boldsymbol{\omega}) \in K$ and thus $\mathbb{E}[g(x, Y)] \ge \mathbf{1}_{\{W_p(P^N_{\xi_\ell}(x), P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \le 0\}} \times (1 - \beta_{\text{ns}}) =: k(x)$, for all $x \in K$. Hence, since $X \equiv (\boldsymbol{\xi}_0, \boldsymbol{w}_\ell)$ and $Y \equiv \boldsymbol{v}_\ell$ are independent by **H2**, we deduce from Lemma 2.1 that

$$\begin{aligned}&\mathbb{E}[g(X, Y)]\\ &\ge \mathbb{E}\Big[\mathbf{1}_{\{W_p(P^N_{\xi_\ell}(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell), P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \le 0\}}(1 - \beta_{\text{ns}})\Big]\\ &= (1 - \beta_{\text{ns}})\mathbb{P}(W_p(P^N_{\xi_\ell}(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell), P_{\xi_\ell}) \le \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})).\end{aligned}$$

From (7) and **H3** in Assumption 3.2, it follows that P_{ξ_ℓ} is supported on the compact set $B^d_\infty(\rho_{\xi_\ell})$ with $\operatorname{diam}_\infty(B^d_\infty(\rho_{\xi_\ell})) = 2\rho_{\xi_\ell}$ and ρ_{ξ_ℓ} given in (20). In addition, due to **H1** and **H2** in Assumption 3.2 the random states ξ^i_ℓ in the empirical distribution $P^N_{\xi_\ell}(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell) = \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i_\ell}$ are i.i.d.. Thus, we get from Proposition 4.1 that $\mathbb{P}(W_p(P^N_{\xi_\ell}(\boldsymbol{\xi}_0, \boldsymbol{w}_\ell), P_{\xi_\ell}) \le \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})) \ge 1 - \beta_{\text{nom}}$, which implies $\mathbb{E}[g(X, Y)] \ge (1 - \beta_{\text{ns}})(1 - \beta_{\text{nom}}) = 1 - \beta$. Finally, (21) follows from this and the definition of g. ∎

With this result at hand, we deduce from the expressions (11) and (18) for the components of the ambiguity radius that it decreases as we exploit a larger number N of independent trajectories and relax our confidence choices, i.e., reduce $1 - \beta_{\text{nom}}$ and $1 - \beta_{\text{ns}}$. Notice further that no matter how many trajectories we use, the noise ambiguity radius decreases to a strictly positive value. It is also worth to observe that ψ_N generalizes the nominal ambiguity radius ε_N in the DRO literature (even when dynamic random variables are considered [9]) and reduces to ε_N in the noise-free case where $\widehat{\varepsilon}_N = 0$.

Drawing conclusions about how the ambiguity radius behaves as we simultaneously allow the horizon $[0 : \ell]$ and the number N of sampled trajectories to increase is a more delicate matter. The value of the nominal component depends essentially on N and the support of the distribution at ℓ, with the latter in turn depending on the system's stability properties and the support of the initial condition and internal noise distributions. On the other hand, the noise component depends on N and the quality of the estimation error. We quantify in the next section how the latter guarantees uniform boundedness of the noise radius under detectability-type assumptions.

Remark 4.8: ***(Positive lower bound of the noise radius).*** The positive lower bound $2^{\frac{p-1}{p}}(\mathfrak{M}_w + \mathfrak{M}_v)$ on the noise radius in (16) represents in general a fundamental limitation for the ambiguity set accuracy, which is independent of the number N of estimated state samples. This is because the bound is related to the size of the state estimation error, which persists under the presence of noise and may further grow in time if there is no system detectability. □

Remark 4.9: ***(Optimal radius selection).*** Once a desired confidence level $1 - \beta$ and the number of independent trajectories N are fixed, we can optimally select the ambiguity radius by minimizing the function

$$\beta_{\text{nom}} \mapsto \psi_N(\beta_{\text{nom}}) \equiv \varepsilon_N(\beta_{\text{nom}}) + \widehat{\varepsilon}_N((\beta - \beta_{\text{nom}})/(1 - \beta_{\text{nom}})),$$

where we have taken into account the constraint (9) between the nominal and the noise confidence. This function is nonconvex, but one-dimensional, and its minimizer is in the interior of the interval $(0, \beta)$, so its optimal value can be approximated with high accuracy. □

D. Uncertainty quantification over bounded time horizons

In this section we discuss how the guarantees can be extended to scenarios where an ambiguity set is built over a finite-time horizon instead of a single instance ℓ. In this case we assume that samples are collected over the time window $[0 : \ell_2]$ and we seek to build an ambiguity set about the state distribution along $[\ell_1 : \ell_2]$, with $0 \le \ell_1 \le \ell_2$. We distinguish between two ambiguity set descriptions depending on the way the associated probabilistic guarantees are obtained. In the first, we directly build an ambiguity set for the probability distribution of the random vector $\boldsymbol{\xi_\ell} := (\xi_{\ell_1}, \dots, \xi_{\ell_2}) \in \mathbb{R}^{\bar{\ell}d}$ with $\boldsymbol{\ell} := (\ell_1, \dots, \ell_2)$ and $\bar{\ell} = \ell_2 - \ell_1 + 1$, comprising of all states over the interval of interest and using the concentration of measure result of Proposition 4.1 for $\bar{\ell}d$-dimensional random variables. This has the drawback that the ambiguity radius decays slowly with the number of trajectories due to the high dimension of $\boldsymbol{\xi_\ell}$. The other description derives an ambiguity set about the state distribution $P_{\xi_{\ell_1}}$ at time ℓ_1 with prescribed confidence, and propagates it under the dynamics while taking into account the possible values of the internal noise. We also present sharper results for the case when the internal noise sequence is known. The first ambiguity set description is provided by the following analogue of Theorem 4.7.

Theorem 4.10: ***(Ambiguity set over a bounded time horizon).*** Consider output data collected from N realizations of system (3) over the interval $[0 : \ell_2]$ and let Assumption 3.1 hold. Pick a confidence $1 - \beta$, let $\beta_{\text{nom}}, \beta_{\text{ns}} \in (0, 1)$ satisfying (9), and consider the bounded-horizon estimator empirical distribution

$$\widehat{P}^N_{\boldsymbol{\xi_\ell}} := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\boldsymbol{\xi}}^i_{\boldsymbol{\ell}}}$$

over the horizon $[\ell_1 : \ell_2]$, where $\widehat{\boldsymbol{\xi}}^i_{\boldsymbol{\ell}} := (\widehat{\xi}^i_{\ell_1}, \dots, \widehat{\xi}^i_{\ell_2}) \in \mathbb{R}^{\widetilde{\ell} d}$ and each $\widehat{\xi}^i_\ell$ is given by the observer (4). Then

$$\mathbb{P}(W_p(\widehat{P}^N_{\boldsymbol{\xi}_{\boldsymbol{\ell}}}, P_{\boldsymbol{\xi}_{\boldsymbol{\ell}}}) \le \psi_N(\beta)) \ge 1 - \beta \tag{22}$$

holds, where $\boldsymbol{\xi}_{\boldsymbol{\ell}} := (\xi_{\ell_1}, \dots, \xi_{\ell_2})$ and $\psi_N(\beta)$ is given in (10). The nominal component $\varepsilon_N(\beta_{\text{nom}}) \equiv \varepsilon_N(\beta_{\text{nom}}, \rho_{\boldsymbol{\xi}_{\boldsymbol{\ell}}})$ is given by (11) (with d in the expression substituted by ℓd)

$$\rho_{\boldsymbol{\xi}_{\boldsymbol{\ell}}} := \max_{\ell \in [\ell_1 : \ell_2]} \left\{ \sqrt{d} \|\Phi_\ell\| \rho_{\xi_0} + \sqrt{q} \sum_{k=1}^{\ell} \|\Phi_{\ell, \ell-k+1} G_{\ell-k}\| \rho_w \right\}, \tag{23}$$

whereas $\widehat{\varepsilon}_N(\beta_{\text{ns}})$ is given as

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

$$\widetilde{\mathfrak{M}}_w := \sum_{\ell=\ell_1}^{\ell_2} \mathfrak{M}_w(\ell),\ \widetilde{\mathfrak{M}}_v := \sum_{\ell=\ell_1}^{\ell_2} \mathfrak{M}_v(\ell),$$

$$\widetilde{\mathfrak{R}} := \frac{\widetilde{\mathfrak{C}}_v}{\widetilde{\mathfrak{m}}_v} + \frac{1}{\ln 2},\ \widetilde{\mathfrak{C}}_v := \sum_{\ell=\ell_1}^{\ell_2} \mathfrak{C}_v(\ell),\ \widetilde{\mathfrak{m}}_v := \sum_{\ell=\ell_1}^{\ell_2} \mathfrak{m}_v(\ell),$$

and $\mathfrak{M}_w(\ell) \equiv \mathfrak{M}_w$, $\mathfrak{M}_v(\ell) \equiv \mathfrak{M}_v$, $\mathfrak{C}_v(\ell) \equiv \mathfrak{C}_v$, and $\mathfrak{m}_v(\ell) \equiv \mathfrak{m}_v$, as given by (12b), (13a), (13b), and (13c), respectively.

The proof of this result follows the argumentation employed for the proof of Theorem 4.7 (a sketch can be found in the online version [10]). For the second ambiguity set description we use a pointwise-in-time approach. To this end, we build a family of ambiguity balls so that under the same confidence level the state distribution at each time instant of the horizon lies in the associated ball, i.e.,

$$\mathbb{P}\big(P_{\xi_\ell} \in \mathcal{B}_{\psi_{N,\ell}}(\widetilde{P}^N_{\xi_\ell})\ \forall \ell \in [\ell_1 : \ell_2]\big) \ge 1 - \beta, \tag{24}$$

where $\mathcal{B}_{\psi_{N,\ell}}(\widetilde{P}^N_{\xi_\ell}) := \{P \in \mathcal{P}_p(\mathbb{R}^d) \,|\, W_p(P, \widetilde{P}^N_{\xi_\ell}) \le \psi_{N,\ell}\}$ and $\widetilde{P}^N_{\xi_\ell}$ is the center of the ball. This is well suited for stochastic optimization problems that have a separable structure with respect to the stochastic argument across different time instances, i.e., problems of the form

$$\inf_{x \in \mathcal{X}} \mathbb{E}\big[f_1(x, \xi_{\ell_1}) + \cdots + f_{\widetilde{\ell}}(x, \xi_{\ell_2})\big].$$

The pointwise ambiguity sets are quantified in the following result.

Theorem 4.11: ***(Pointwise ambiguity sets over a bounded time horizon).*** Let the assumptions of Theorem 4.10 hold, assume that the internal noise sequence w_ℓ is independent (also of the initial state), $P_{w_\ell} \in \mathcal{P}_p(\mathbb{R}^d)$ for $\ell \in [\ell_1 : \ell_2]$, i.e., it is not necessarily compactly supported, and consider either of the following two cases for its distribution when $\ell \in [\ell_1 : \ell_2]$:

(i) P_{w_ℓ} is not known and $\mathbb{E}\big[\|w_\ell\|^p\big]^{\frac{1}{p}} \le q_w$.
(ii) P_{w_ℓ} is known.

Then, for any confidence $1 - \beta$, and $\beta_{\text{nom}}, \beta_{\text{ns}} \in (0, 1)$ satisfying (9), (24) holds, with $\widetilde{P}^N_{\xi_{\ell_1}} := \widehat{P}^N_{\xi_{\ell_1}}$ and ψ_{N,ℓ_1} as given by Theorem 4.7 (for $\ell \equiv \ell_1$), and $\widetilde{P}^N_{\xi_\ell}$, $\psi_{N,\ell}$, $\ell \in [\ell_1 + 1 : \ell_2]$ defined as follows for the respective two cases above:

(i) The ambiguity set center is $\widetilde{P}^N_{\xi_\ell} := \frac{1}{N} \sum_{i=1}^N \delta_{\widetilde{\xi}^i_\ell}$ with $\widetilde{\xi}^i_\ell := \Phi_{\ell,\ell_1} \widehat{\xi}_{\ell_1}$ and the radius is given recursively by $\psi_{N,\ell} := \|A_{\ell-1}\| \psi_{N,\ell-1} + q_w$.
(ii) The ambiguity set center is $\widetilde{P}^N_{\xi_\ell} := \big((A_{\ell-1})_\# \widetilde{P}^N_{\xi_{\ell-1}}\big) \star P_{w_{\ell-1}}$ and the radius is $\psi_{N,\ell} := \|A_{\ell-1}\| \cdots \|A_{\ell_1}\| \psi_{N,\ell_1}$.

Note that when the internal noise distribution is known, all individual ambiguity balls of Theorem 4.11 shrink at the exact same decay rate with the number of samples, which overcomes the slow decay rate of the ambiguity radius of Theorem 4.10 for larger time horizons. In our technical approach, we use the next result, whose proof can be found in the online version [10]. The result examines what happens to the Wasserstein distance between the distributions of two random variables when other random variables are added.

Lemma 4.12: ***(Wasserstein distance under convolution).*** Given $p \ge 1$ and distributions $P_1, P_2, Q \in \mathcal{P}_p(\mathbb{R}^d)$, it holds that $W_p(P_1, P_2) \le W_p(P_1 \star Q, P_2 \star Q)$. Also, if it holds that $\big(\int_{\mathbb{R}^d} \|x\|^p Q(dx)\big)^{\frac{1}{p}} \le q$, then $W_p(P_1, P_2 \star Q) \le W_p(P_1, P_2) + q$.

Proof of Theorem 4.11: The proof is carried out by induction on $\ell \in [\ell_1 : \ell_2]$. In particular, it suffices to establish that

$$\begin{aligned} &W_p\big(P_{\xi_{\ell_1}}, \widetilde{P}^N_{\xi_{\ell_1}}\big) \le \psi_{N,\ell_1} \Longrightarrow \\ &W_p\big(P_{\xi_{\ell'}}, \widetilde{P}^N_{\xi_{\ell'}}\big) \le \psi_{N,\ell'}\ \forall \ell' \in [\ell_1 : \ell]. \end{aligned} \tag{25}$$

Note that from Theorem 4.7, $\mathbb{P}\big(P_{\xi_{\ell_1}} \in \mathcal{B}_{\psi_{N,\ell_1}}(\widetilde{P}^N_{\xi_{\ell_1}})\big) \ge 1 - \beta$. From (25), this also implies that $\mathbb{P}\big(P_{\xi_{\ell'}} \in \mathcal{B}_{\psi_{N,\ell'}}(\widetilde{P}^N_{\xi_{\ell'}})\ \forall \ell' \in [\ell_1 : \ell]\big) \ge 1 - \beta$, establishing validity of the result for $\ell \equiv \ell_2$.

For $\ell \equiv \ell_1$, the induction hypothesis (25) is a tautology. Next, assuming that it is true for certain $\ell \in [\ell_1 : \ell_2 - 1]$, we show that it also holds for $\ell + 1$. Hence it suffices to show that if $W_p\big(P_{\xi_\ell}, \widetilde{P}^N_{\xi_\ell}\big) \le \psi_{N,\ell}$ then also $W_p\big(P_{\xi_{\ell+1}}, \widetilde{P}^N_{\xi_{\ell+1}}\big) \le \psi_{N,\ell+1}$ for both cases (i) and (ii). Consider first (i) and note that then the ambiguity set center at $\ell + 1$ satisfies $\widetilde{P}^N_{\xi_{\ell+1}} = \frac{1}{N} \sum_{i=1}^N \delta_{\widetilde{\xi}^i_{\ell+1}} = \frac{1}{N} \sum_{i=1}^N \delta_{A_\ell \widetilde{\xi}^i_\ell} = (A_\ell)_\# \widetilde{P}^N_{\xi_\ell}$, where we have exploited that $\widetilde{\xi}^i_k \equiv \Phi_{k,\ell_1} \widehat{\xi}^i_{\ell_1}$ and the definition of Φ_{k,ℓ_1} (for $k = \ell - 1, \ell$) to derive the second equality. Using also the fact that $P_{\xi_{\ell+1}} = \big((A_\ell)_\# P_{\xi_\ell}\big) \star P_{w_\ell}$, we get from the second result of Lemma 4.12 that

$$\begin{aligned} W_p(P_{\xi_{\ell+1}}, \widetilde{P}^N_{\xi_{\ell+1}}) &= W_p\big(\big((A_\ell)_\# P_{\xi_\ell}\big) \star P_{w_\ell}, (A_\ell)_\# \widetilde{P}^N_{\xi_\ell}\big) \\ &\le W_p\big((A_\ell)_\# P_{\xi_\ell}, (A_\ell)_\# \widetilde{P}^N_{\xi_\ell}\big) + q_w \\ &\le \|A_\ell\| W_p\big(P_{\xi_\ell}, \widetilde{P}^N_{\xi_\ell}\big) + q_w \le \|A_\ell\| \psi_{N,\ell} + q_w = \psi_{N,\ell+1}. \end{aligned}$$

Here we also used the fact that $W_p(f_\# P, f_\# Q) \le L W_p(P, Q)$ for any globally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^r$ with Lipschitz constant L in the second to last inequality (see e.g., [45, Proposition 7.16]).

Next, we prove the induction hypothesis for (ii). Using Lemma 4.12 and the definition of the ambiguity set center and radius,

$$\begin{aligned} &W_p(P_{\xi_{\ell+1}}, \widetilde{P}^N_{\xi_{\ell+1}}) \\ &\quad = W_p\big(\big((A_\ell)_\# P_{\xi_\ell}\big) \star P_{w_\ell}, \big((A_\ell)_\# \widetilde{P}^N_{\xi_\ell}\big) \star P_{w_\ell}\big) \\ &\quad \le W_p\big((A_\ell)_\# P_{\xi_\ell}, (A_\ell)_\# \widetilde{P}^N_{\xi_\ell}\big) \le \|A_\ell\| W_p\big(P_{\xi_\ell}, \widetilde{P}^N_{\xi_\ell}\big) \\ &\quad \le \|A_\ell\| \psi_{N,\ell} = \|A_\ell\| \|A_{\ell-1}\| \cdots \|A_{\ell_1}\| \psi_{N,\ell_1} = \psi_{N,\ell+1}, \end{aligned}$$

completing the proof. ∎

V. Sufficient conditions for uniformly bounded noise ambiguity radii

In this section we leverage Assumption 3.4 to establish that the noise ambiguity radius remains uniformly bounded as the sampling horizon increases. We first provide uniform bounds for the matrices involved in the system and observer error dynamics.

Proposition 5.1: ***(Bounds on system/observer matrices).*** Under Assumption 3.4, the gain matrices K_k can be selected so that the following properties hold:

(i) There exist $K_\star, K^\star, G^\star > 0$ and $\Psi_s^\star > 0$, $s \in \mathbb{N}_0$, so that $\|G_k\| \leq G^\star$, $K_\star \leq \|K_k\| \leq K^\star$, and $\|\Psi_{k+s,k}\| \leq \Psi_s^\star$ for all and $k \in \mathbb{N}_0$.

(ii) There exists $s_0 \in \mathbb{N}$ so that $\|\Psi_{k+s,k}\| \leq \frac{1}{2}$ for all $k \in \mathbb{N}_0$ and $s \geq s_0$.

Proof: Note that we only need to verify part *(i)* for the time-varying case. Since all G_k are uniformly bounded, we directly obtain the bound $G^\star$. Let

$$K_k := -A_k \Phi_{k,k-t-1} \mathcal{O}_{k,k-t-1}^{-1} \Phi_{k,k-t-1}^\top H_k^\top,$$

(for $k > t+1$) as selected in [36, Page 574] (but with a minus sign at the front to get the plus sign in $F_k = A_k + K_k H_k$) and with the observability Gramian $\mathcal{O}_{k,k-t-1}$ as defined in Assumption 3.4(ii). Then, the upper bound $K^\star$ follows from the fact that the system matrices are uniformly bounded combined with the uniform observability property of Assumption 3.4, which implies that all $\mathcal{O}_{k,k-t-1}^{-1}$ are also uniformly bounded. On the other hand, the lower bound $K_\star$ follows from the assumption that the system matrices are uniformly bounded, which imposes a uniform lower bound on the smallest singular value of $\mathcal{O}_{k,k-t-1}^{-1}$, the uniform lower bound on the smallest singular value of A_k, hence, also on that of $\Phi_{k,k-t-1}$ and $\Phi_{k,k-t-1}^\top$, and the uniform lower bound on $\|H_k\|$ (all found in Assumption 3.4). Finally, the bounds $\Psi_s^\star$ follow from the uniform bounds for all A_k and H_k and the derived bound $K^\star$ for all K_k.

To show part *(ii)*, assume first that Assumption 3.4(i) holds, i.e., the system is time invariant and (A, H) is detectable. Then, we can choose a nonzero gain matrix K so that $F = A + KH$ is convergent (cf. [43, Theorem 31]), namely $\lim_{s\to\infty} \|F^s\| = 0$. Consequently, there is $s_0 \in \mathbb{N}$ with $\|F^s\| \leq \frac{1}{2}$ for all $s \geq s_0$ and the result follows by taking into account that $\Psi_{k+s,k} = F^s$. In case Assumption 3.4(ii) holds, let

$$\widetilde{e}_{k+1} = F_k \widetilde{e}_k \tag{26}$$

be the recursive noise-free version of the error equation (6). Then, from [36, Page 577], there exists a quadratic time-varying Lyapunov function $V(k, \widetilde{e}) := \widetilde{e}^\top Q_k \widetilde{e}$ with each Q_k being positive definite, $a_1, a_2 > 0$, $a_3 \in (0, 1)$, and $m \in \mathbb{N}$ so that

$$a_1 \leq \lambda_{\min}(Q_k) \leq \lambda_{\max}(Q_k) \leq a_2 \tag{27a}$$

$$V(k+m, \widetilde{e}_{k+m}) - V(k, \widetilde{e}_k) \leq -a_3 V(k, \widetilde{e}_k) \tag{27b}$$

for any k and any solution of (26) with state $\widetilde{e}_k$ at time k. Thus, $\Psi_{k+m,m}^\top Q_{k+m} \Psi_{k+m,m} \preceq (1-a_3) Q_k$, and hence, by induction $\Psi_{k+\nu m,m}^\top Q_{k+\nu m} \Psi_{k+\nu m,m} \preceq (1-a_3)^\nu Q_k$, since

$$\begin{aligned}
&\Psi_{k+(\nu+1)m,m}^\top Q_{k+(\nu+1)m} \Psi_{k+(\nu+1)m,m} \\
&= \Psi_{k+m,k}^\top \Psi_{k+(\nu+1)m,k+m}^\top Q_{k+(\nu+1)m} \Psi_{k+(\nu+1)m,k+m} \Psi_{k+m,k} \\
&\preceq (1-a_3)^\nu \Psi_{k+m,k}^\top Q_{k+m} \Psi_{k+m,k} \preceq (1-a_3)^{(\nu+1)} Q_k.
\end{aligned}$$

Next, pick $\widetilde{e}$ with $\|\widetilde{e}\| = 1$ and $\|\Psi_{k+\nu m,m}\widetilde{e}\| = \|\Psi_{k+\nu m,m}\|$. Taking into account that $\widetilde{e}^\top \Psi_{k+\nu m,k}^\top Q_{k+\nu m} \Psi_{k+\nu m,m} \widetilde{e} \leq (1-a_3)^\nu \widetilde{e}^\top Q_k \widetilde{e}$, we get $\lambda_{\min}(Q_{k+\nu m}) \|\Psi_{k+\nu m,k} \widetilde{e}\|^2 \leq (1-a_3)^\nu \lambda_{\max}(Q_k)$. Using (27a),

$$\|\Psi_{k+\nu m,k}\| \leq (1-a_3)^{\frac{\nu}{2}} \Big(\frac{a_2}{a_1}\Big)^{\frac{1}{2}}. \tag{28}$$

Now, select ν so that $(1-a_3)^{\frac{\nu'}{2}} (a_2/a_1)^{\frac{1}{2}} \leq 1/(2 \max_{s\in[1:m]} \Psi_s^\star)$ for all $\nu' \geq \nu$. Let $s_0 := \nu m$ and pick $s \geq s_0$. Then, $s = s_0' + m'$ for some $s_0' = \nu' m$, $\nu' \geq \nu$, and $m' \in [0 : m-1]$ and we get from (28), part (i), and the selection of ν that

$$\begin{aligned}
\|\Psi_{k+sm,k}\| &= \|\Psi_{k+s_0'+m',k+s_0'} \Psi_{k+s_0',k}\| \\
&\leq \|\Psi_{k+s_0'+m',k+s_0'}\| \|\Psi_{k+\nu' m,k}\| \leq \Psi_{m'}^\star \frac{1}{2 \max_{s\in[1:m]} \Psi_s^\star} \leq \frac{1}{2},
\end{aligned}$$

which establishes the result. ∎

Based on this result and Assumption 3.4 about the system's detectability/uniform observability properties, we proceed to provide a uniform bound on the size of the noise radius for arbitrarily long evolution horizons.

Proposition 5.2: ***(Uniform bounds for noise ambiguity radius).*** Consider data collected from N realizations of system (3), a confidence $1-\beta$ as in (9), and let Assumptions 3.1, 3.2, and 3.4 hold. Then, there exist observer gain matrices K_k so that the noise ambiguity radius $\widehat{\varepsilon}_N$ 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 $\ell \geq \ell_0$, $\mathfrak{M}_w \equiv \mathfrak{M}_w(\ell)$, $\mathfrak{M}_v \equiv \mathfrak{M}_v(\ell)$, and $\mathfrak{R} \equiv \mathfrak{R}(\ell)$ given by (12b), (13a), and (17), are uniformly upper bounded as

$$\mathfrak{M}_w \leq \frac{1}{2}\sqrt{d}\rho_{\xi_0} + 3\sqrt{q} \sum_{j=0}^{\ell_0-1} \Psi_j^\star G^\star \rho_w,$$

$$\mathfrak{M}_v \leq 3 M_v r \sum_{j=0}^{\ell_0-1} \Psi_j^\star K^\star, \; \mathfrak{R} \leq 3 \frac{C_v}{m_v} r^{\frac{p-1}{p}} \frac{\sum_{j=0}^{\ell_0-1} \Psi_j^\star K^\star}{K_\star}.$$

Proof: Consider gain matrices K_k and the time s_0 as given in Proposition 5.1, and let $\ell_0 := s_0$. Then, for any $\ell \geq \ell_0$, $\ell = n\ell_0 + r'$ with $0 \leq r' < \ell_0$ and we have

$$\begin{aligned}
&\sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} G_{\ell-k}\| \leq \sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1}\| G^\star \\
&= \Bigg(\sum_{k=1}^{r'} \|\Psi_{\ell,\ell-k+1}\| + \sum_{k=r'+1}^{\ell} \|\Psi_{\ell,\ell-k+1}\| \Bigg) G^\star \\
&\leq \Bigg(\sum_{s=0}^{r'-1} \Psi_s^\star + \sum_{k=r'+1}^{n\ell_0+r'} \|\Psi_{n\ell_0+r',n\ell_0+r'-k+1}\| \Bigg) G^\star \\
&\quad (k \mapsto (\nu-1)\ell_0 + j + r')
\end{aligned}$$

$$
\begin{aligned}
&= \Bigg(\sum_{s=0}^{r'-1} \Psi_s^\star + \sum_{\nu=1}^{n} \sum_{j=1}^{\ell_0} \|\Psi_{n\ell_0+r',(n-\nu)\ell_0+r'+\ell_0-j+1}\| \Bigg) G^\star \\
&\quad (\ell_0+1-j \mapsto j) \\
&\leq \Bigg(\sum_{s=0}^{r'-1} \Psi_s^\star + \sum_{\nu=1}^{n} \sum_{j=1}^{\ell_0} \|\Psi_{n\ell_0+r',(n-\nu+1)\ell_0+r'}\| \\
&\qquad\qquad \times \|\Psi_{(n-\nu)\ell_0+r'+\ell_0,(n-\nu)\ell_0+r'+j}\| \Bigg) G^\star \\
&= \Bigg(\sum_{s=0}^{r'-1} \Psi_s^\star + \sum_{\nu=1}^{n} \|\Psi_{n\ell_0+r',(n-\nu+1)\ell_0+r'}\| \sum_{j=1}^{\ell_0} \\
&\qquad\qquad \times \|\Psi_{(n-\nu)\ell_0+r'+\ell_0,(n-\nu)\ell_0+r'+j}\| \Bigg) G^\star \\
&\leq \Bigg(\sum_{s=0}^{r'-1} \Psi_s^\star + \sum_{\nu=1}^{n} \Bigg(\prod_{\kappa=1}^{\nu-1} \|\Psi_{(n+1-\kappa)\ell_0+r',(n-\kappa)\ell_0+r'}\| \Bigg) \\
&\qquad\qquad \times \sum_{j=1}^{\ell_0} \Psi_{\ell_0-j}^\star \Bigg) G^\star \\
&\leq \Bigg(\sum_{s=0}^{\ell_0-1} \Psi_s^\star + \sum_{\nu=1}^{n} \Big(\frac{1}{2}\Big)^{\nu-1} \sum_{j=0}^{\ell_0-1} \Psi_j^\star \Bigg) G^\star \leq 3 \sum_{j=0}^{\ell_0-1} \Psi_j^\star G^\star,
\end{aligned}
$$

where we have used $\sum_{\kappa=0}^{-1} \equiv \sum_{\kappa=1}^{0} \equiv 0$ and $\prod_{\kappa=1}^{0} \equiv 1$. From this and the fact that from Proposition 5.1, $\|\Psi_\ell\| \leq \frac{1}{2}$ for all $\ell \geq \ell_0$, we get the upper bound for $\mathfrak{M}_w$. The one for $\mathfrak{M}_v$ is obtained analogously. Finally, for $\mathfrak{R}$, we obtain the same type of upper bound for $\mathfrak{C}_v$ as for $\mathfrak{M}_w$, and exploit Proposition 5.1(i) to get the lower bound $\mathfrak{m}_v = m_v r^{\frac{1}{p}} \big(\sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\|^p \big)^{\frac{1}{p}} \geq m_v r^{\frac{1}{p}} \|\Psi_{\ell,\ell} K_{\ell-1}\| \geq m_v r^{\frac{1}{p}} K_\star$, which is independent of ℓ. ∎

Remark 5.3: ***(Noise ambiguity radius for time-invariant systems).*** For time-invariant systems, it is possible to improve the bounds of Proposition 5.2 for $\mathfrak{M}_w$, $\mathfrak{M}_v$, and $\mathfrak{R}$ by exploiting the fact that the system and observer gain matrices are constant. The precise bounds in this case (see also [8, Proposition 5.5]) are

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

$$
\mathfrak{M}_v \leq 2M_v r \sum_{k=0}^{\ell_0-1} \|\Psi_k K\|,\ \mathfrak{R} \leq 2\frac{C_v}{m_v} r^{\frac{p-1}{p}} \frac{\sum_{k=0}^{\ell_0-1} \|\Psi_k K\|}{\big(\sum_{k=0}^{\ell_0-1} \|\Psi_k K\|^p\big)^{\frac{1}{p}}}
$$

with ℓ_0 as in the time-invariant case of Proposition 5.2, and where G and K denote the constant values of the internal noise and observer gain matrices, resp. The superiority of these bounds can be checked using the matrix bounds in Proposition 5.1(i) and their derivation is based on a simplified version of the arguments employed for the proof of Proposition 5.2.□

VI. Application to economic dispatch with distributed energy resources

In this section, we take advantage of the ambiguity sets constructed with noisy partial measurements, cf. Theorem 4.7, to hedge against the uncertainty in an optimal economic dispatch problem. This is a problem where uncertainty is naturally involved due to (dynamic) energy resources, which the scheduler has no direct access to control or measure, like storage or renewable energy elements. The financial implications of the associated decisions are of utmost importance for the electricity market and justify the use of a reliable decision framework that accounts for the variability of the uncertain factors.

A. Network model and optimization objective

Consider a network with distributed energy resources [13] comprising of n_1 generator units and n_2 storage (battery) units. The network needs to operate as close as possible to a prescribed power demand D at the end of the time horizon $[0:\ell]$, corresponding to a uniform discretization of step-size δt of the continuous-time domain. To this end, each generator and storage unit supplies the network with positive power P^j and S^ι, respectively, at time ℓ. We assume we can control the power of the generators, which additionally needs to be within the upper and lower thresholds $P^j_{\min}$ and $P^j_{\max}$, respectively. Each battery is modeled as an uncertain dynamic element with an unknown initial state distribution and we can decide whether it is connected ($\eta^\iota = 1$) or not ($\eta^\iota = 0$) to the network at time ℓ. Our goal is to minimize the energy cost while remaining as close as possible to the prescribed power demand. Thus, we minimize the overall cost

$$
\begin{aligned}
\mathcal{C}(\boldsymbol{P}, \boldsymbol{\eta}) :=& \sum_{j=1}^{n_1} g^j(P^j) + \sum_{\iota=1}^{n_2} \eta^\iota h^\iota(S^\iota) \\
&+ c\Bigg(\sum_{j=1}^{n_1} P^j + \sum_{\iota=1}^{n_2} \eta^\iota S^\iota - D \Bigg)^2 \qquad (29)
\end{aligned}
$$

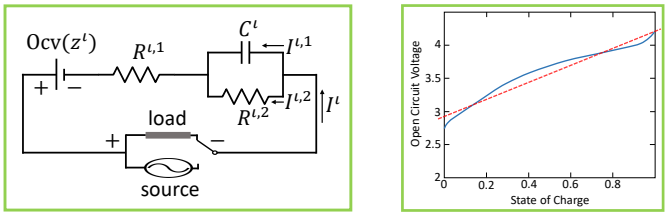
where $\boldsymbol{P} := (P^1, \dots, P^{n_1})$, $\boldsymbol{\eta} := (\eta^1, \dots, \eta^{n_2})$, g^j and h^ι are cost functions for the power provided by generator j and storage unit ι, respectively. We treat the deviation of the injected power from its prescribed demand as a soft constraint by assigning it a quadratic cost with weight c and augmenting the overall cost function (29). Due to the uncertainty about the batteries' state and their injected powers S^ι, the minimization of (29) is a stochastic problem.

B. Battery dynamics and observation model

Each battery is modeled as a single-cell dynamic element and we consider its current I^ι discharging over the operation interval (if connected to the network) as a fixed and a priori known function of time. Its dynamics is conveniently approximated by the equivalent circuit in Figure 2(a) (see e.g., [32], [33]), where z^ι is the state of charge (SoC) of the cell and $\mathrm{Ocv}(z^\iota)$ is its corresponding open-circuit voltage, which we approximate by the affine function $\alpha^\iota z^\iota + \beta^\iota$ in Figure 2(b). The associated discrete-time cell model is

$$
\begin{aligned}
\chi^\iota_{k+1} \equiv \begin{pmatrix} I^{\iota,2}_{k+1} \\ z^\iota_{k+1} \end{pmatrix} &= \begin{pmatrix} a^\iota & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} I^{\iota,2}_k \\ z^\iota_k \end{pmatrix} + \begin{pmatrix} 1-a^\iota \\ -\delta t/Q^\iota \end{pmatrix} I^\iota_k \\
\theta^\iota_k \equiv V^\iota_k &= \alpha^\iota z^\iota_k + \beta^\iota - I^\iota_k R^{\iota,1} - I^{\iota,2}_k R^{\iota,2}
\end{aligned}
$$

where $a^\iota := e^{-\delta t/(R^{2,\iota} C^\iota)}$, δt is the time discretization step, and Q^ι is the cell capacity. Here, we assume that for all $k \in [0:\ell]$ the cell is neither fully charged or discharged (by e.g.,



(a) (b)

Fig. 2. (a) shows the equivalent circuit model of a lithium-ion battery cell in discharging mode (c.f. [33, Figure 2],[32, Figure 1]). (b) is taken from [32, Figure 3] and shows the nonlinear dependence of the open circuit voltage on the state of charge and its affine approximation.

requiring that $0 < z_0 - \sum_{k=0}^{\ell-1} \delta t I_k^\iota / Q^\iota < 1$ for all k and any candidate initial conditions and input currents) and so, the evolution of its voltage is accurately represented by the above difference equation. The initial condition comprising of the SoC z_0^ι and the current $I_0^{\iota,2}$ through $R^{\iota,2}$ is random with an unknown probability distribution. We also consider additive measurement noise with an unknown distribution, namely, we measure

$$\theta_k^\iota = \alpha^\iota z_k^\iota + \beta^\iota - I_k^\iota R^{\iota,1} - I_k^{\iota,2} R^{\iota,2} + v_k.$$

To track the evolution of each random element through a linear system of the form (3), we consider for each battery a nominal state trajectory $\chi_k^{\iota,\star} = (I_k^{\iota,2,\star}, z_k^{\iota,\star})$ initiated from the center of the support of its initial-state distribution. Setting $\xi_k^\iota = \chi_k^\iota - \chi_k^{\iota,\star}$ and $\zeta_k^\iota = \theta_k(\chi_k^\iota) - \theta_k(\chi_k^{\iota,\star})$,

$$\begin{aligned} \xi_{k+1}^\iota &= A_k^\iota \xi_k^\iota \\ \zeta_k^\iota &= H_k^\iota \xi_k^\iota + v_k, \end{aligned}$$

where $A_k^\iota := \operatorname{diag}(a, 1)$ and $H_k^\iota := (\alpha^\iota, -R^{\iota,2})$. Denoting $\boldsymbol{\xi} := (\xi^1, \ldots, \xi^{n_2})$ and $\boldsymbol{\zeta} := (\zeta^1, \ldots, \zeta^{n_2})$, we obtain a system of the form (3) for the dynamic random variable $\boldsymbol{\xi}$. Despite the fact that the state distribution $\boldsymbol{\xi}_k$ of the batteries across time is unknown, we assume having access to output data from N independent realizations of their dynamics over the horizon $[0 : \ell]$. Using these samples we exploit the results of the paper to build an ambiguity ball $\mathcal{P}^N$ of radius ε_N in the 2-Wasserstein distance (i.e., with $p = 2$), that contains the batteries' state distribution $P_{\boldsymbol{\xi}_\ell}$ at time ℓ with prescribed probability $1-\beta$. In particular, we take the samples from each realization $i \in [1 : N]$ and use an observer to estimate its state $\widehat{\boldsymbol{\xi}}_\ell^i$ at time ℓ. The ambiguity set is centered at the estimator-based empirical distribution $\widehat{P}_{\boldsymbol{\xi}_\ell}^N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\boldsymbol{\xi}}_\ell^i}$ and its radius can be determined using Theorem 4.7 and Proposition 4.6.

C. Decision problem as a distributionally robust optimization (DRO) problem

To solve the decision problem regarding whether or not to connect the batteries for economic dispatch, we formulate a distributionally robust optimization problem for the cost (29) using the ambiguity set $\mathcal{P}^N$. To do this, we derive an explicit expression of how the cost function $\mathcal{C}$ depends on the stochastic argument $\boldsymbol{\xi}_\ell$. Notice first that the power injected by each battery at time ℓ is

$$\begin{aligned} S^\iota = I_\ell^\iota V_\ell^\iota &= I_\ell^\iota\big(\alpha^\iota z_\ell^\iota + \beta^\iota - I_\ell^\iota R^{\iota,1} - I_\ell^{\iota,2} R^{\iota,2}\big) \\ &= \langle (-I_\ell^\iota R^{\iota,2}, \alpha^\iota I_\ell^\iota), \chi_\ell^\iota \rangle + \beta^\iota I_\ell^\iota - (I_\ell^\iota)^2 R^{\iota,1} \\ &= \langle \widehat{\alpha}^\iota, \xi_\ell^\iota \rangle + \widehat{\beta}^\iota \equiv (\widehat{\alpha}^\iota)^\top \xi_\ell^\iota + \widehat{\beta}^\iota, \end{aligned}$$

with $\widehat{\alpha}^\iota := (-I_\ell^\iota R^{\iota,2}, \alpha^\iota I_\ell^\iota)$ and

$$\begin{aligned} \widehat{\beta}^\iota &:= \langle \widehat{\alpha}^\iota, \chi_\ell^{\iota,\star} \rangle + I_\ell^\iota \beta^\iota - (I_\ell^\iota)^2 R^{\iota,1} \\ &= I_\ell^\iota I_\ell^{\iota,2,\star} R^{\iota,2} - \alpha^\iota I_\ell^\iota z_\ell^{\iota,\star} + I_\ell^\iota \beta^\iota - (I_\ell^\iota)^2 R^{\iota,1}. \end{aligned}$$

Considering further affine costs $h^\iota(S) := \bar{\alpha}^\iota S + \bar{\beta}^\iota$ for the power provided by the batteries, the overall cost $\mathcal{C}$ becomes

$$\begin{aligned} \mathcal{C}(\boldsymbol{P}, \boldsymbol{\eta}) = g(\boldsymbol{P}) &+ (\boldsymbol{\eta} * \widetilde{\boldsymbol{\alpha}})^\top \boldsymbol{\xi}_\ell + \boldsymbol{\eta}^\top \widetilde{\boldsymbol{\beta}} \\ &+ c\big(\mathbf{1}^\top \boldsymbol{P} + (\boldsymbol{\eta} * \widehat{\boldsymbol{\alpha}})^\top \boldsymbol{\xi}_\ell + \boldsymbol{\eta}^\top \widehat{\boldsymbol{\beta}} - D\big)^2, \end{aligned} \tag{30}$$

where $*$ denotes the Khatri-Rao product (cf. Section II) and

$$\begin{aligned} g(\boldsymbol{P}) &:= \sum_{j=1}^{n_1} g^j(P^j), \ \widehat{\boldsymbol{\alpha}} := (\widehat{\alpha}^1, \ldots, \widehat{\alpha}^{n_2}), \\ \widehat{\boldsymbol{\beta}} &:= (\widehat{\beta}^1, \ldots, \widehat{\beta}^{n_2}), \ \widetilde{\boldsymbol{\alpha}} := (\bar{\alpha}^1 \widehat{\alpha}^1, \ldots, \bar{\alpha}^{n_2} \widehat{\alpha}^{n_2}), \\ \widetilde{\boldsymbol{\beta}} &:= (\bar{\alpha}^1 \widehat{\beta}^1 + \bar{\beta}^1, \ldots, \bar{\alpha}^{n_2} \widehat{\beta}^{n_2} + \bar{\beta}^{n_2}). \end{aligned}$$

Using the equivalent description (30) for $\mathcal{C}$ and recalling the upper and lower bounds $P_{\min}^j$ and $P_{\max}^j$ for the generator's power, we formulate the DRO power dispatch problem

$$\inf_{\boldsymbol{\eta}, \boldsymbol{P}} \Big\{ f_{\boldsymbol{\eta}}(\boldsymbol{P}) + \sup_{P_{\boldsymbol{\xi}_\ell} \in \mathcal{P}^N} \mathbb{E}_{P_{\boldsymbol{\xi}_\ell}}\big[h_{\boldsymbol{\eta}}(\boldsymbol{P}, \boldsymbol{\xi}_\ell)\big] \Big\}, \tag{31a}$$

$$\text{s.t. } P_{\min}^j \le P^j \le P_{\max}^j \quad \forall j \in [1 : n_1], \tag{31b}$$

with the ambiguity set $\mathcal{P}^N$ introduced above and

$$\begin{aligned} f_{\boldsymbol{\eta}}(\boldsymbol{P}) &:= g(\boldsymbol{P}) + c\boldsymbol{P}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{P} \\ &\quad + 2c(\boldsymbol{\eta}^\top \widehat{\boldsymbol{\beta}} - D)\mathbf{1}^\top \boldsymbol{P} + c(\boldsymbol{\eta}^\top \widehat{\boldsymbol{\beta}} - D)^2 + \boldsymbol{\eta}^\top \widetilde{\boldsymbol{\beta}} \\ h_{\boldsymbol{\eta}}(\boldsymbol{P}, \boldsymbol{\xi}_\ell) &:= c\boldsymbol{\xi}_\ell^\top (\boldsymbol{\eta} * \widehat{\boldsymbol{\alpha}})(\boldsymbol{\eta} * \widehat{\boldsymbol{\alpha}})^\top \boldsymbol{\xi}_\ell + \big(2c\big(\mathbf{1}^\top \boldsymbol{P} \\ &\quad + \boldsymbol{\eta}^\top \widehat{\boldsymbol{\beta}} - D)(\boldsymbol{\eta} * \widehat{\boldsymbol{\alpha}})^\top + (\boldsymbol{\eta} * \widetilde{\boldsymbol{\alpha}})^\top\big)\boldsymbol{\xi}_\ell, \end{aligned}$$

This formulation aims to minimize the worst-case expected cost with respect to the plausible distributions of $\boldsymbol{\xi}$ at time ℓ.

D. Tractable reformulation of the DRO problem

Our next goal is to obtain a tractable reformulation of the optimization problem (31). To this end, we first provide an equivalent description for the inner maximization in (31), which is carried out over a space of probability measures. Exploiting strong duality (see [22, Corollary 2(i)] or [5, Remark 1]) and recalling that our ambiguity set $\mathcal{P}^N$ is based on the 2-Wasserstein distance, we equivalently write the inner maximization problem $\sup_{P_{\boldsymbol{\xi}_\ell} \in \mathcal{P}^N} \mathbb{E}_{P_{\boldsymbol{\xi}_\ell}}\big[h_{\boldsymbol{\eta}}(\boldsymbol{P}, \boldsymbol{\xi}_\ell)\big]$ as

$$\inf_{\lambda \ge 0} \Big\{ \lambda \psi_N^2 + \frac{1}{N} \sum_{i=1}^N \sup_{\xi_\ell \in \Xi} \{ h_{\boldsymbol{\eta}}(\boldsymbol{P}, \boldsymbol{\xi}_\ell) - \lambda \|\boldsymbol{\xi}_\ell - \widehat{\boldsymbol{\xi}}_\ell^i\|^2 \} \Big\}, \tag{32}$$

where $\psi_N \equiv \psi_N(\beta)$ is the radius of the ambiguity ball, $\Xi \subset \mathbb{R}^{2n_2}$ is the support of the batteries' unknown state distribution, and the $\widehat{\boldsymbol{\xi}}_\ell^i$ are the estimated states of their

realizations. We slightly relax the problem, by allowing the ambiguity ball to contain all distributions with distance ψ_N from $\widehat{P}^N_{\boldsymbol{\xi}_\ell}$ that are supported on $\mathbb{R}^{2n_2}$ and not necessarily on Ξ. Thus, we first look to solve for each estimated state $\widehat{\boldsymbol{\xi}}^i_\ell$ the optimization problem

$$\sup_{\xi_\ell\in\mathbb{R}^{2n_2}} \{h_{\boldsymbol{\eta}}(\boldsymbol{P},\boldsymbol{\xi}_\ell) - \lambda\|\boldsymbol{\xi}_\ell - \widehat{\boldsymbol{\xi}}^i_\ell\|^2\},$$

which is written

$$\begin{aligned}
&\sup_{\boldsymbol{\xi}_\ell\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}_\ell^\top\mathfrak{A}\boldsymbol{\xi}_\ell + \big(2c\big(\mathbf{1}^\top\boldsymbol{P}+\boldsymbol{\eta}^\top\widehat{\boldsymbol{\beta}}-D\big)(\boldsymbol{\eta}*\widehat{\boldsymbol{\alpha}})^\top \\
&\qquad + (\boldsymbol{\eta}*\widetilde{\boldsymbol{\alpha}})^\top\big)\boldsymbol{\xi}_\ell - \lambda(\boldsymbol{\xi}_\ell-\widehat{\boldsymbol{\xi}}^i_\ell)^\top(\boldsymbol{\xi}_\ell-\widehat{\boldsymbol{\xi}}^i_\ell)\big\} \\
&= -\lambda(\widehat{\boldsymbol{\xi}}^i_\ell)^\top\widehat{\boldsymbol{\xi}}^i_\ell + \sup_{\boldsymbol{\xi}_\ell\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}_\ell^\top(\mathfrak{A}-\lambda I_{2n_2})\boldsymbol{\xi}_\ell \\
&\qquad + \big(2c\big(\mathbf{1}^\top\boldsymbol{P}+\boldsymbol{\eta}^\top\widehat{\boldsymbol{\beta}}-D\big)(\boldsymbol{\eta}*\widehat{\boldsymbol{\alpha}})^\top \\
&\qquad + (\boldsymbol{\eta}*\widetilde{\boldsymbol{\alpha}})^\top + 2\lambda(\widehat{\boldsymbol{\xi}}^i_\ell)^\top\big)\boldsymbol{\xi}_\ell\big\} \\
&= -\lambda(\widehat{\boldsymbol{\xi}}^i_\ell)^\top\widehat{\boldsymbol{\xi}}^i_\ell + \sup_{\boldsymbol{\xi}_\ell\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}_\ell^\top(\mathfrak{A}-\lambda I_{2n_2})\boldsymbol{\xi}_\ell + (\boldsymbol{r}^i)^\top\boldsymbol{\xi}_\ell\big\}
\end{aligned}$$

where $\boldsymbol{r}^i \equiv \boldsymbol{r}^i_{\boldsymbol{\eta}}(\boldsymbol{P},\lambda) := 2c(\mathbf{1}^\top\boldsymbol{P}+\boldsymbol{\eta}^\top\widehat{\boldsymbol{\beta}}-D)(\boldsymbol{\eta}*\widehat{\boldsymbol{\alpha}})+\boldsymbol{\eta}*\widetilde{\boldsymbol{\alpha}}+2\lambda\widehat{\boldsymbol{\xi}}^i_\ell$ and $\mathfrak{A}\equiv\mathfrak{A}_{\boldsymbol{\eta}} := c(\boldsymbol{\eta}*\widehat{\boldsymbol{\alpha}})(\boldsymbol{\eta}*\widehat{\boldsymbol{\alpha}})^\top$ is a symmetric positive semi-definite matrix with diagonalization $\mathfrak{A}=\mathfrak{Q}^\top\mathfrak{D}\mathfrak{Q}$ where the eigenvalues decrease along the diagonal. Hence, we get

$$\begin{aligned}
&\sup_{\boldsymbol{\xi}_\ell\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}_\ell^\top(\mathfrak{A}-\lambda I_{2n_2})\boldsymbol{\xi}_\ell + (\boldsymbol{r}^i)^\top\boldsymbol{\xi}_\ell\big\} \\
&\quad= \sup_{\boldsymbol{\xi}_\ell\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}_\ell^\top(\mathfrak{Q}^\top\mathfrak{D}\mathfrak{Q}-\mathfrak{Q}^\top\lambda I_{2n_2}\mathfrak{Q})\boldsymbol{\xi}_\ell + (\boldsymbol{r}^i)^\top\boldsymbol{\xi}_\ell\big\} \\
&\quad= \sup_{\boldsymbol{\xi}\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}^\top(\mathfrak{D}-\lambda I_{2n_2})\boldsymbol{\xi} + (\widehat{\boldsymbol{r}}^i)^\top\boldsymbol{\xi}\big\}
\end{aligned}$$

with $\widehat{\boldsymbol{r}}^i := \mathfrak{Q}\boldsymbol{r}^i$ and denoting $\lambda_{\max}(\mathfrak{A})$ the maximum eigenvalue of $\mathfrak{A}$ we have

$$\begin{aligned}
&\sup_{\boldsymbol{\xi}\in\mathbb{R}^{2n_2}} \big\{\boldsymbol{\xi}^\top(\mathfrak{D}-\lambda I_{2n_2})\boldsymbol{\xi} + (\widehat{\boldsymbol{r}}^i)^\top\boldsymbol{\xi}\big\} \\
&\quad= \begin{cases} \infty & \text{if } 0\le\lambda<\lambda_{\max}(\mathfrak{A}) \\ \frac{1}{4}(\widehat{\boldsymbol{r}}^i)^\top(\lambda I_{2n_2}-\mathfrak{D})^{-1}\widehat{\boldsymbol{r}}^i & \text{if } \lambda>\lambda_{\max}(\mathfrak{A}). \end{cases}
\end{aligned} \tag{33}$$

To obtain this we exploited that $Q(\boldsymbol{\xi}) := \boldsymbol{\xi}^\top(\mathfrak{D}-\lambda I_{2n_2})\boldsymbol{\xi} + (\widehat{\boldsymbol{r}}^i)^\top\boldsymbol{\xi}$ is maximized when

$$\begin{aligned}
\nabla Q(\boldsymbol{\xi}_\star)=0 &\iff 2(\mathfrak{D}-\lambda I_{2n_2})\boldsymbol{\xi}_\star + \widehat{\boldsymbol{r}}^i = 0 \\
&\iff \boldsymbol{\xi}_\star = \frac{1}{2}(\lambda I_{2n_2}-\mathfrak{D})^{-1}\widehat{\boldsymbol{r}}^i,
\end{aligned}$$

which gives the optimal value $Q(\boldsymbol{\xi}_\star) = \frac{1}{4}(\widehat{\boldsymbol{r}}^i)^\top(\lambda I_{2n_2}-\mathfrak{D})^{-1}\widehat{\boldsymbol{r}}^i$. Note that we do not need to specify the value of the expression in (33) for $\lambda=\lambda_{\max}$. In particular, since the function we minimize in (32) is convex in λ, the inner part of the DRO problem is equivalently written

$$\begin{aligned}
\inf_{\lambda>\lambda_{\max}(\mathfrak{A})} \Bigg\{&\lambda\bigg(\psi_N^2 - \frac{1}{N}\sum_{i=1}^N(\widehat{\boldsymbol{\xi}}^i_\ell)^\top\widehat{\boldsymbol{\xi}}^i_\ell\bigg) \\
&+ \frac{1}{4N}\sum_{i=1}^N \widehat{\boldsymbol{r}}^i_{\boldsymbol{\eta}}(\boldsymbol{P},\lambda)^\top(\lambda I_{2n_2}-\mathfrak{D})^{-1}\widehat{\boldsymbol{r}}^i_{\boldsymbol{\eta}}(\boldsymbol{P},\lambda)\Bigg\}.
\end{aligned}$$

Taking further into account that

$$(\lambda I_{2n_2}-\mathfrak{D})^{-1} = \operatorname{diag}\Big(\frac{1}{\lambda-\lambda_{\max}(\mathfrak{A})},\dots,\frac{1}{\lambda-\lambda_{\min}(\mathfrak{A})}\Big),$$

as well as the constraints (31b) on the decision variable $\boldsymbol{P}$, the overall DRO problem is reformulated as

$$\begin{aligned}
\min_{\boldsymbol{\eta}}\inf_{\boldsymbol{P},\lambda} \Bigg\{&f_{\boldsymbol{\eta}}(\boldsymbol{P}) + \lambda\bigg(\psi_N^2 - \frac{1}{N}\sum_{i=1}^N(\widehat{\boldsymbol{\xi}}^i_\ell)^\top\widehat{\boldsymbol{\xi}}^i_\ell\bigg) \\
&+ \frac{1}{4N}\sum_{i=1}^N \widehat{\boldsymbol{r}}^i_{\boldsymbol{\eta}}(\boldsymbol{P},\lambda)^\top \\
&\times \operatorname{diag}\Big(\frac{1}{\lambda-\lambda_{\max}(\mathfrak{A})},\dots,\frac{1}{\lambda-\lambda_{\min}(\mathfrak{A})}\Big)\widehat{\boldsymbol{r}}^i_{\boldsymbol{\eta}}(\boldsymbol{P},\lambda)\Bigg\}
\end{aligned} \tag{34a}$$

$$\text{subject to } P^j_{\min}\le P^j\le P^j_{\max}\quad\forall j\in[1:n_1]$$
$$\lambda>\lambda_{\max}(\mathfrak{A}). \tag{34b}$$

E. Simulation results

For the simulations we consider $n_1=4$ generators and $n_2=3$ batteries with the same characteristics. We assume that the distributions of each initial SoC z^ι_0 and current $I^{\iota,2}_0$ are known to be supported on the intervals $[0.45,0.9]$ and $[1.5,1.7]$, respectively. The true SoC distribution for batteries 2 and 3 at time zero is $P_{z^2_0}=P_{z^3_0}=\mathcal{U}[0.45,0.65]$ ($\mathcal{U}$ denotes uniform distribution). On the other hand, the provider of battery 1 has access to the distinct batteries 1A and 1B and selects randomly one among them with probabilities 0.9 and 0.1, respectively. The SoC distribution of battery 1A at time zero is $P_{z^{1A}_0}=\mathcal{U}[0.45,0.65]$, whereas that of battery 1B is $P_{z^{1B}_0}=\mathcal{U}[0.84,0.86]$. Thus, we get the bimodal distribution $P_{z^1_0}=0.9\mathcal{U}[0.45,0.65]+0.1\mathcal{U}[0.84,0.86]$, which is responsible for non-negligible empirical distribution variations, since for small numbers of samples, it can fairly frequently occur that the relative percentage of samples from 1B deviates significantly from its expected one. On the other hand, we assume that the true initial currents $I^{\iota,2}_0$ of all batteries are fixed to 1.6308, namely, $P_{I^{1,2}_0}=P_{I^{2,2}_0}=P_{I^{3,2}_0}=\delta_{1.6308}$. For the measurements, we consider the Gaussian mixture noise model $P_{v_k}=0.5\mathcal{N}(0.01,0.01^2)+0.5\mathcal{N}(-0.01,0.01^2)$ with $\mathcal{N}(\mu,\sigma^2)$ denoting the normal distribution with mean μ and variance σ^2.

To compute the ambiguity radius for the reformulated DRO problem (34), we specify its nominal and noise components $\varepsilon_N(\beta_{\text{nom}},\rho_{\boldsymbol{\xi}_\ell})$ and $\widehat{\varepsilon}_N(\beta_{\text{ns}})$, where due to Proposition 4.1, $\rho_{\boldsymbol{\xi}_\ell}$ can be selected as half the diameter of any set containing the support of $P_{\boldsymbol{\xi}_\ell}$ in the infinity norm. It follows directly from the specific dynamics of the batteries that $\rho_{\boldsymbol{\xi}_\ell}$ does not exceed half the diameter of the initial conditions' distribution support, which is isometric to $[0.45,0.9]^3\times[1.5,1.7]^3\subset\mathbb{R}^6$. Hence, using Proposition 19 in the online version [10] with $p=2$, $d=6$, and $\rho_{\boldsymbol{\xi}_\ell}=0.225$, we obtain

$$\varepsilon_N(\beta_{\text{nom}},\rho_{\boldsymbol{\xi}_\ell}) = 4.02N^{-\frac{1}{6}} + 1.31(\ln\beta^{-1}_{\text{nom}})^{\frac{1}{4}}N^{-\frac{1}{4}}.$$

To determine the noise radius, we first compute lower and upper bounds m_v and M_v for the L_2 norm of the Gaussian mixture noise v_k and an upper bound C_v for its ψ_2 norm. Denoting by $\mathbb{E}_P$ the integral with respect to the distribution P, we have for $P_{v_k}=0.5\mathcal{N}(\mu_1,\sigma_1^2)+0.5\mathcal{N}(\mu_2,\sigma_2^2)$ that $\|v_k\|_2^2 = \mathbb{E}_{\frac{1}{2}(P_1+P_2)}[v_k^2] = \frac{1}{2}(\mu_1^2+\sigma_1^2+\mu_2^2+\sigma_2^2)$, where $P_1=\mathcal{N}(\mu_1,\sigma_1^2)$, $P_2=\mathcal{N}(\mu_2,\sigma_2^2)$ and we used the fact that

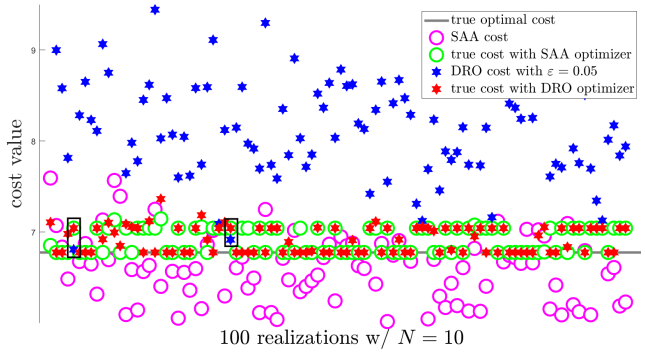


Fig. 3. Results from 100 realizations of the power dispatch problem with $N = 10$ independent samples used for each realization. We compute the optimizers of the SAA and DRO problems, plot their corresponding optimal values (termed "SAA cost" and "DRO cost"), and also evaluate their performance with respect to the true distribution ("true cost with SAA optimizer" and "true cost with DRO optimizer"). With the exception of two realizations (whose DRO cost and true cost with the DRO optimizer are framed inside black boxes), the DRO cost is above the true cost of the DRO optimizer, namely, this happens with high probability. From the plot, it is also clear that the SAA solution tends to over-promise since its value is most frequently below the true cost of the SAA optimizer.

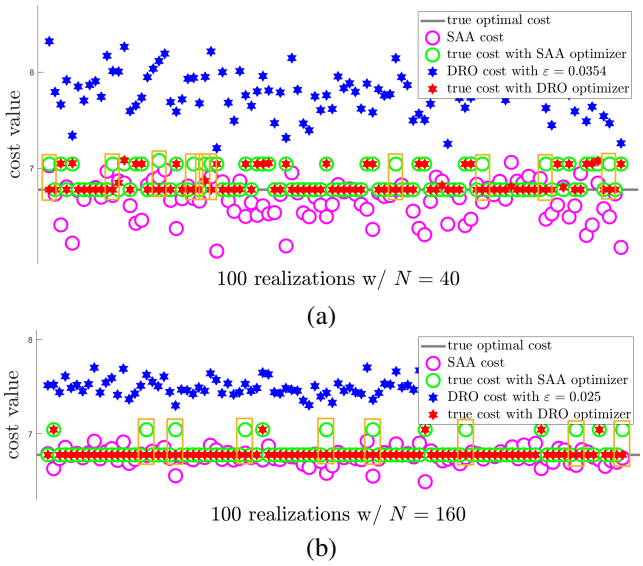


Fig. 4. Analogous results to those of Figure 3, from 100 realizations with (a) $N = 40$ and (b) $N = 160$ independent samples, and the ambiguity radius tuned so that the same confidence level is preserved. In both cases, the DRO cost is above the true cost of the DRO optimizer with high probability (in fact, always). Furthermore, the cost of the DRO optimizer (red star) is strictly better than the true cost of the SAA one (green circle) for a considerable number of realizations (highlighted in the illustrated boxes).

$\mathbb{E}_{P_i}\left[v_k^2\right] = \mu_i^2 + \mathbb{E}_{P_i}\left[(v_k - \mu_i)^2\right] = \mu_i^2 + \sigma_i^2$. Hence, in our case, where $\mu_i = \sigma_i = 0.01$, we can pick $m_v = M_v = 0.01\sqrt{2}$. Further, using Proposition 21 from the online version [10], we can select $C_v = 0.01(\sqrt{8/3} + \sqrt{\ln 2})$. To perform the state estimation from the output samples we used a Kalman filter. Its initial condition covariance matrix corresponds to independent Gaussian distributions for each SoC z_0^ι and current $I_0^{\iota,2}$ with a standard deviation of the order of their assumed support. We also select the same covariance as in the components of the Gaussian mixture noise to model the measurement noise of the Kalman filter. Using the dynamics of the filter and the values of m_v, M_v, and C_v above, we obtain from (12b), (13a)-(13c), and (17) the constants $\mathfrak{M}_w = 0.325$, $\mathfrak{M}_v = 0.008$, and $\mathfrak{R} = 2.72$ for the expression of the noise radius. In particular, we have from Proposition 4.6 that $\widehat{\varepsilon}_N(\beta_{\text{ns}}) = 0.47 + 0.0113\sqrt{74.98/N \ln(2/\beta_{\text{ns}})}$ and the overall radius is

$$\psi_N(\beta) = 0.47 + 4.02N^{-\frac{1}{6}} + 1.31(\ln \beta_{\text{nom}}^{-1})^{\frac{1}{4}} N^{-\frac{1}{4}} + 0.0973(\ln(2\beta_{\text{ns}}^{-1}))^{\frac{1}{2}} N^{-\frac{1}{2}}. \quad (35)$$

We assume that the energy cost of the generators is lower than that of the batteries and select the quadratic power generation cost $g(\boldsymbol{P}) = 0.25\sum_{j=1}^{4}(P^j - 0.1)^2$ and the same lower/upper power thresholds $P^j_{\min} = 0.2/P^j_{\max} = 0.5$ for all generators. For the batteries, we pick the same resistances $R^{\iota,1} = 0.34$ and $R^{\iota,2} = 0.17$, and we take $a^\iota = 0.945$ and $I_k^\iota = 8$ for all times. We nevertheless use different linear costs $h^\iota(S) = \bar{\alpha}^\iota S$ for their injected powers, with $\bar{\alpha}^1 = 1$ and $\bar{\alpha}^2 = \bar{\alpha}^3 = 1.3$, since battery 1 is less reliable due to the large SoC fluctuation among its two modes.

We solve 100 independent realizations of the overall economic dispatch problem. For each of them, we generate independent samples from the batteries' initial condition distributions and solve the associated sample average approximation (SAA) and DRO problems for $N = 10$, $N = 40$, and $N = 160$ samples, respectively, using CVX [23]. It is worth noting that the radius ψ_N given by (35) is rather conservative. The main reasons for this are 1) conservativeness of the concentration of measure results used for the derivation of the nominal radius, 2) lack of homogeneity of the distribution's support (the a priori known support of the $I_0^{\iota,2}$ components is much smaller than that of the z_0^ι ones), 3) independence of the batteries' individual distributions, which we have not exploited, and 4) conservative upper bounds for the estimation error. Although there is room to sharpen all these aspects, it requires multiple additional contributions and lies beyond the scope of the paper. Nevertheless, the formula (35) gives a qualitative intuition about the decay rates for the ambiguity radius. In particular, it indicates that under the same confidence level and for small sample sizes, an ambiguity radius proportional to $N^{-\frac{1}{4}}$ is a reasonable choice. Based on this, we selected the ambiguity radii 0.05, 0.0354, and 0.025 for $N = 10$, $N = 40$, and $N = 160$. The associated simulation results are shown in Figures 3, 4(a), and 4(b), respectively. We plot there the optimal values of the SAA and DRO problems (termed "SAA cost" and "DRO cost") and provide the expected performance of their respective decisions with respect to the true distribution ("true cost with SAA optimizer" a.k.a. out-of-sample performance and "true cost with DRO optimizer"). We observe that in all three cases, the DRO value is above the true cost of the DRO optimizer for nearly all realizations (and for all when N is 40 or 160), which verifies the finite sample guarantees of DRO formulations [18, Theorem 3.5]. In addition, when solving the problem for 40 or 160 samples, we witness a clear out-of-sample superiority of the DRO decision compared to the one of the non-robust SAA, because it considerably improves the true cost for a significant number of realizations (cf. Figure 4).

F. Discussion

The SAA solution tends to consistently promise a better outcome compared to what the true distribution reveals for the same decision (e.g., magenta circle being usually under the green circle in all figures). This rarely happens for the DRO solution, and when it does, it is only by a small margin. This makes the DRO approach preferable over the SAA one in the context of power systems operations where honoring committments at a much higher cost than anticipated might result in significant losses, and not fulfilling committments may lead to penalties from the system operator.

VII. Conclusions

We have constructed high-confidence ambiguity sets for dynamic random variables using partial-state measurements from independent realizations of their evolution. In our model, both the dynamics and measurements are subject to disturbances with unknown probability distributions. The ambiguity sets are built using an observer to estimate the full state of each realization and leveraging concentration of measure inequalities. For systems that are either time-invariant and detectable, or uniformly observable, we have established uniform boundedness of the ambiguity radius. To aid the associated probabilistic guarantees, we also provided auxiliary concentration of measure results. Future research will include the consideration of robust state estimation criteria to mitigate the noise effect on the ambiguity radius, the extension of the results to nonlinear dynamics, and the construction of ambiguity sets with information about the moments.

VIII. Appendix

Here we give proofs of various results of the paper.

Proof of Lemma 2.1: Since X and Y are independent, their joint distribution $P_{(X,Y)}$ is the product measure $P_X \otimes P_Y$ of the individual distributions P_X and P_Y. Thus, from the Fubini theorem [1, Theorem 2.6.5] and integrability of g, k we get

$$\begin{aligned}\mathbb{E}[g(X,Y)] &= \int_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} g(x,y)dP_{(X,Y)} \\ &= \int_{\mathbb{R}^{n_1}}\int_{\mathbb{R}^{n_2}} g(x,y)dP_Y dP_X = \int_{\mathbb{R}^{n_1}} \mathbb{E}[g(x,Y)]dP_X \\ &= \int_K \mathbb{E}[g(x,Y)]dP_X \geq \int_K k(x)dP_X \\ &= \int_{\mathbb{R}^{n_1}} k(x)dP_X = \mathbb{E}[k(X)],\end{aligned}$$

which concludes the proof. ∎

Proof of Lemma 4.3: Using [9, Lemma A.2] to bound the Wasserstein distance of two discrete distributions, we get $W_p(\widehat{P}^N_{\xi_\ell}, P^N_{\xi_\ell}) \leq \left(\frac{1}{N}\sum_{i=1}^N \|\widehat{\xi}^i_\ell - \xi^i_\ell\|^p\right)^{\frac{1}{p}} = \left(\frac{1}{N}\sum_{i=1}^N \|e^i_\ell\|^p\right)^{\frac{1}{p}}$. From (6), we have

$$\begin{aligned}\|e^i_\ell\| &= \Big\|\Psi_\ell z^i + \sum_{k=1}^{\ell}\big(\Psi_{\ell,\ell-k+1}G_{\ell-k}\omega^i_{\ell-k} \\ &\qquad + \Psi_{\ell,\ell-k+1}K_{\ell-k}v^i_{\ell-k}\big)\Big\| \\ &\leq \|\Psi_\ell\|\|z^i\| + \sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}G_{\ell-k}\|\|\omega^i_{\ell-k}\| \\ &\quad + \sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|\|v^i_{\ell-k}\|_1 =: \mathfrak{M}(z^i,\boldsymbol{\omega}^i) + \mathfrak{E}(\boldsymbol{v}^i),\end{aligned}$$

with $\mathfrak{E}(\boldsymbol{v}^i) \equiv \mathfrak{E}^i$ given in the statement. Since $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ for $a,b\geq 0$ and $p\geq 1$,

$$W_p(\widehat{P}^N_{\xi_\ell}, P^N_{\xi_\ell}) \leq \Big(\frac{1}{N}2^{p-1}\sum_{i=1}^N(\mathfrak{M}(z^i,\boldsymbol{\omega}^i)^p + (\mathfrak{E}^i)^p)\Big)^{\frac{1}{p}}.$$

Next, using $(a+b)^{\frac{1}{p}} \leq a^{\frac{1}{p}} + b^{\frac{1}{p}}$ for $a,b\geq 0$ and $p\geq 1$, we have

$$\begin{aligned}W_p(\widehat{P}^N_{\xi_\ell}, P^N_{\xi_\ell}) &\leq \Big(\frac{1}{N}2^{p-1}\sum_{i=1}^N \mathfrak{M}(z^i,\boldsymbol{\omega}^i)^p\Big)^{\frac{1}{p}} \\ &\quad + \Big(\frac{1}{N}2^{p-1}\sum_{i=1}^N(\mathfrak{E}^i)^p\Big)^{\frac{1}{p}}. \qquad (36)\end{aligned}$$

Finally, since $(\boldsymbol{z},\boldsymbol{\omega}) \in B^{Nd}_\infty(\rho_{\xi_0}) \times B^{N\ell q}_\infty(\rho_w)$, we get $\mathfrak{M}(z^i,\boldsymbol{\omega}^i)^p \leq \|\Psi_\ell\|\sqrt{d}\|z^i\|_\infty + \sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}G_{\ell-k}\|\sqrt{q}\|\omega^i_{\ell-k-1}\|_\infty \leq \mathfrak{M}_w$. This combined with (36) yields (12a). ∎

Proof of Lemma 4.4: From **H4** in Assumption 3.2, we obtain for each summand in (12c)

$$\begin{aligned}&\big\|\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|\|v^i_{\ell-k}\|_1\big\|_{\psi_p} \\ &\quad\leq \|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|\big(\big\|v^i_{\ell-k,1}\big\|_{\psi_p} + \cdots + \big\|v^i_{\ell-k,r}\big\|_{\psi_p}\big) \\ &\quad\leq C_v r\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|.\end{aligned}$$

Hence, we deduce that

$$\begin{aligned}\|\mathfrak{E}^i\|_{\psi_p} &\leq \sum_{k=1}^{\ell}\big\|\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|\|v^i_{\ell-k}\|_1\big\|_{\psi_p} \\ &\leq C_v r\sum_{k=1}^{\ell}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|.\end{aligned}$$

For the L^p bounds, note that $\|\mathfrak{E}^i\|_p = \big\|\sum_{k\in[1:\ell],l\in[1:r]}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\||v^i_{\ell-k,l}|\big\|_p$. Thus, from the inequality $\big\|\sum_i c_iX_i\big\|_p \leq \sum_i c_i\|X_i\|_p$, which holds for any nonnegative c_i and X_i in L^p,

$$\|\mathfrak{E}^i\|_p \leq \sum_{k\in[1:\ell],l\in[1:r]}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|\|v^i_{\ell-k,l}\|_p,$$

which, by the upper bound in **H4** of Assumption 3.2, implies (13a). For the other bound, we exploit linearity of the expectation and the inequality $\big(\sum_i c_i\big)^p \geq \sum_i c_i^p$, which holds for any nonnegative c_i, to get

$$\begin{aligned}&\big(\mathbb{E}\big[(\mathfrak{E}^i)^p\big]\big)^{\frac{1}{p}} \\ &\quad\geq \Bigg(\sum_{k\in[1:\ell],l\in[1:r]}\|\Psi_{\ell,\ell-k+1}K_{\ell-k}\|^p\mathbb{E}\big[|v^i_{\ell-k,l}|\big]^p\Bigg)^{\frac{1}{p}}.\end{aligned}$$

Thus, from the lower bound in **H4** of Assumption 3.2 we also obtain (13c). ∎

We next prove Proposition 4.5, along the lines of the proof of [44, Theorem 3.1.1], which considers the special

case of sub-Gaussian distributions. We rely on the following concentration inequality [44, Corollary 2.8.3].

Proposition 8.1: ***(Bernstein inequality).*** Let $X_1, \dots, X_N$ be scalar, mean-zero, sub-exponential, independent random variables. Then, for every $t \geq 0$ we have

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} X_i\Big| \geq t\Big) \leq 2\exp\Big(-c' \min\Big\{\frac{t^2}{R^2}, \frac{t}{R}\Big\}N\Big),$$

where $c' = 1/10$ and $R := \max_{i\in[1:N]} \|X_i\|_{\psi_1}$.

The precise constant c' above is not specified in [44] but we provide an independent proof of this result in the online version [10, Section 8.2].

Proof of Proposition 4.5: Note that each random variable $X_i^p - 1$ is mean zero by assumption. Additionally, we have that $\|X_i^p - 1\|_{\psi_1} \leq \|X_i^p\|_{\psi_1} + \|1\|_{\psi_1} = \|X_i\|_{\psi_p} + 1/\ln 2 \leq R$, where we took into account that

$$\mathbb{E}[\psi_1(X_i^p/t^p)] = \mathbb{E}[\psi_p(X_i/t)] \Rightarrow \|X_i^p\|_{\psi_1} = \|X_i\|_{\psi_p},$$

and the following fact, shown after the proof.

▷ *Fact I.* For any constant random variable $X = \mu \in \mathbb{R}$, it holds $\|X\|_{\psi_p} = |\mu|/(\ln 2)^{\frac{1}{p}}$. ◁

Thus, we get from Proposition 8.1 that

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} X_i^p - 1\Big| \geq t\Big) \leq 2\exp\Big(-\frac{c'N}{R^2}\min\{t^2, t\}\Big), \tag{37}$$

where we used the fact that $R > 1$. We will further leverage the following facts shown after the proof of the proposition.

▷ *Fact II.* For all $p \geq 1$ and $z \geq 0$ it holds that $|z-1| \geq \delta \Rightarrow |z^p - 1| \geq \max\{\delta, \delta^p\}$. ◁

▷ *Fact III.* For any $\delta \geq 0$, if $u = \max\{\delta, \delta^p\}$, then $\min\{u, u^2\} = \alpha_p(\delta)$, with α_p as given by (15). ◁

By exploiting Fact II, we get

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\Big(\frac{1}{N}\sum_{i=1}^{N} X_i^p\Big)^{\frac{1}{p}} - 1\Big| \geq t\Big) \\
&\quad\leq \mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} X_i^p - 1\Big| \geq \max\{t, t^p\}\Big) \\
&\quad\leq 2\exp\Big(-\frac{c'N}{R^2}\min\{\max\{t,t^p\}^2, \max\{t,t^p\}\}\Big).
\end{aligned}$$

Thus, since $\mathbb{P}(|Y| \geq t) \geq \mathbb{P}(Y \geq t)$ for any random variable Y, we obtain (14) from Fact III and conclude the proof. ■

Proof of Fact I: From the ψ_p norm definition, $\|X\|_{\psi_p} = \inf\big\{t > 0 \,|\, \mathbb{E}\big[e^{(|X|/t)^p}\big] \leq 2\big\} = \inf\big\{t > 0 \,|\, t \geq |\mu|/(\ln 2)^{\frac{1}{p}}\big\} = |\mu|/(\ln 2)^{\frac{1}{p}}$, which establishes the result. ■

Proof of Fact II: Assume first that $z < 1$. Then, we have that $|z^p - 1| = 1 - z^p > 1 - z \geq \delta \geq \delta^p$. Next, let $z \geq 1$. Then, we get $|z^p - 1| = z^p - 1 \geq z - 1 \geq \delta$. In addition, when $\delta^p \geq \delta$, namely, when $\delta \geq 1$, we have that $z^p - (z-1)^p \geq 1$, and hence, $|z^p - 1| = z^p - 1 \geq (z-1)^p \geq \delta^p$. ■

Proof of Fact III: We consider two cases. Case (i): $0 \leq \delta \leq 1 \Rightarrow \delta \geq \delta^p \Rightarrow u = \max\{\delta, \delta^p\} = \delta$. Then $\min\{u, u^2\} = \min\{\delta, \delta^2\} = \delta^2$. Case (ii): $\delta > 1 \Rightarrow \delta \leq \delta^p \Rightarrow u = \max\{\delta, \delta^p\} = \delta^p$. Then $\min\{u, u^2\} = \min\{\delta^p, \delta^{2p}\} = \delta^p$. Thus, we get that $\min\{u, u^2\} = \alpha_p(\delta)$ for all $\delta \geq 0$. ■

REFERENCES

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

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

[3] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," *Mathematical Programming*, vol. 171, no. 1-2, pp. 217–282, 2018.

[4] 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.

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

[6] 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.

[7] V. Bogachev, *Measure theory*. Springer, 2007, vol. 1.

[8] 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, July 2020, pp. 4491–4496.

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

[10] D. Boskos, J. Cortés, and S. Martínez, "High-confidence data-driven ambiguity sets for time-varying linear systems," *https://arxiv.org/abs/2102.01142*, 2021.

[11] F. Boso, D. Boskos, J. Cortés, S. Martínez, and D. M. Tartakovsky, "Dynamics of data-driven ambiguity sets for hyperbolic conservation laws with uncertain inputs," *SIAM Journal on Scientific Computing*, vol. 43, no. 3, pp. A2102–A2129, 2021.

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

[13] A. Cherukuri and J. Cortés, "Distributed coordination of DERs with storage for dynamic economic dispatch," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 835–842, 2018.

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

[15] J. Coulson, J. Lygeros, and F. Dörfler, "Data-enabled predictive control: In the shallows of the DeePC," in *European Control Conference*, 2019, pp. 307–312.

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

[17] S. Dereich, M. Scheutzow, and R. Schottstedt, "Constructive quantization: Approximation by empirical measures," *Annales de l'Institut Henri Poincar, Probabilits et Statistiques*, vol. 49, no. 4, p. 11831203, 2013.

[18] P. M. 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.

[19] 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. 707738, 2015.

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

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

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

[23] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[24] Y. Guo, K. Baker, E. DallAnese, Z. Hu, and T. H. Summers, "Data-based distributionally robust stochastic optimal power flow–Part I: Methodologies," *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1483–1492, 2018.

[25] 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.

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

[27] B. Kloeckner, "Empirical measures: regularity is a counter-curse to dimensionality," *arXiv preprint arXiv:1802.04038*, 2019.

[28] B. Li, J. Mathieu, and R. Jiang, "Distributionally robust chance constrained optimal power flow assuming log-concave distributions," in *Power Systems Computation Conference*, 2018, pp. 1–7.

[29] 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.

[30] ——, "Online learning of parameterized uncertain dynamical environments with finite-sample guarantees," *IEEE Control Systems Letters*, 2020.

[31] D. Li and S. Martínez, "Online data assimilation in distributionally robust optimization," in *IEEE Int. Conf. on Decision and Control*, Miami, FL, USA, December 2018, pp. 1961–1966.

[32] M. Li, "Li-ion dynamics and state of charge estimation," *Renewable Energy*, vol. 100, pp. 44–52, 2017.

[33] B. Y. Liaw, G. Nagasubramanian, R. G. Jungst, and D. H. Doughty, "Modeling of lithium ion cells–A simple equivalent-circuit model approach," *Solid State Ionics*, vol. 175, no. 1-4, pp. 835–839, 2004.

[34] J. Liu, Y. Chen, C. Duan, J. Lin, and J. Lyu, "Distributionally robust optimal reactive power dispatch with Wasserstein distance in active distribution network," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 3, pp. 426–436, 2020.

[35] S. Liu, "Matrix results on the Khatri-Rao and Tracy-Singh products," *Linear Algebra and its Applications*, vol. 289, no. 1, p. 267277, 1999.

[36] J. B. Moor and B. D. O. Anderson, "Coping with singular transition matrices in estimation and control stability theory," *International Journal of Control*, vol. 31, no. 3, pp. 571–586, 1980.

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

[38] 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.

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

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

[41] A. Shapiro, "Distributionally robust stochastic programming," *SIAM Journal on Optimization*, vol. 27, no. 4, p. 22582275, 2017.

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

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

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

[45] C. Villani, *Topics in optimal transportation*, ser. Graduate Studies in Mathematics. American Mathematical Society, 2003, no. 58.

[46] 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.

[47] J. Weed and Q. Berthe, "Estimation of smooth densities in Wasserstein distance," *arXiv preprint arXiv:1902.01778*, 2019.

[48] F. Xin, B.-M. Hodge, L. Fangxing, D. Ershun, and K. Chongqing, "Adjustable and distributionally robust chance-constrained economic dispatch considering wind power uncertainty," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 3, pp. 658–664, 2019.

[49] I. Yang, "A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance," *IEEE Control Systems Letters*, vol. 1, no. 1, pp. 164–169, 2017.

[50] ——, "Wasserstein distributionally robust stochastic control: A data-driven approach," *arXiv preprint arXiv:1812.09808*, 2018.

[51] S. Zeng, "Sample-based population observers," *Automatica*, vol. 101, pp. 166–174, 2019.

Dimitris Boskos (M' 15) was born in Athens, Greece in 1981. He has received the Diploma in Mechanical Engineering from the National Technical University of Athens (NTUA), Greece, in 2005, the M.Sc. in Applied mathematics from the NTUA in 2008 and the Ph.D. in Applied mathematics from the NTUA in 2014. Between August 2014 and August 2018, he has been a Postdoctoral Researcher with the Department of Automatic Control, School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden. Between August 2018 and August 2020, he is a Postdoctoral Researcher with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA. Since October 2020, he is an Assistant Professor with the Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands. His research interests include distributionally robust optimization, distributed control of multi-agent systems, formal verification, and nonlinear observer design.

Jorge Cortés (M'02, SM'06, F'14) received the Licenciatura degree in mathematics from Universidad de Zaragoza, Zaragoza, Spain, in 1997, and the Ph.D. degree in engineering mathematics from Universidad Carlos III de Madrid, Madrid, Spain, in 2001. He held postdoctoral positions with the University of Twente, Twente, The Netherlands, and the University of Illinois at Urbana-Champaign, Urbana, IL, USA. He was an Assistant Professor with the Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA, USA, from 2004 to 2007. He is currently a Professor in the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA. He is the author of Geometric, Control and Numerical Aspects of Nonholonomic Systems (Springer-Verlag, 2002) and co-author (together with F. Bullo and S. Martínez) of Distributed Control of Robotic Networks (Princeton University Press, 2009). At the IEEE Control Systems Society, he has been a Distinguished Lecturer (2010-2014), and is currently its Director of Operations and an elected member (2018-2020) of its Board of Governors. His current research interests include distributed control and optimization, network science, resource-aware control, nonsmooth analysis, reasoning and decision making under uncertainty, network neuroscience, and multi-agent coordination in robotic, power, and transportation networks.

Sonia Martínez (M'02-SM'07-F'18) is a Professor of Mechanical and Aerospace Engineering at the University of California, San Diego, CA, USA. She received the Ph.D. degree in Engineering Mathematics from the Universidad Carlos III de Madrid, Spain, in May 2002. Following a year as a Visiting Assistant Professor of Applied Mathematics at the Technical University of Catalonia, Spain, she obtained a Postdoctoral Fulbright Fellowship and held appointments at the Coordinated Science Laboratory of the University of Illinois, Urbana-Champaign during 2004, and at the Center for Control, Dynamical systems and Computation (CCDC) of the University of California, Santa Barbara during 2005. In a broad sense, her main research interests include the control of network systems, multi-agent systems, nonlinear control theory, and robotics. For her work on the control of underactuated mechanical systems she received the Best Student Paper award at the 2002 IEEE Conference on Decision and Control. She was the recipient of a NSF CAREER Award in 2007. For the paper "Motion coordination with Distributed Information," co-authored with Jorge Cortés and Francesco Bullo, she received the 2008 Control Systems Magazine Outstanding Paper Award. She has served on the editorial board of the European Journal of Control (2011-2013) and the Journal of Geometric Mechanics (2009-present), and currently serves as a Senior Editor of the IEEE Transactions on Control of Network Systems.