

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

Boskos, Dimitris; Cortes, Jorge; Martinez, Sonia

**DOI**

[10.1109/TAC.2023.3273815](https://doi.org/10.1109/TAC.2023.3273815)

**Publication date**

2024

**Document Version**

Final published version

**Published in**

IEEE Transactions on Automatic Control

**Citation (APA)**

Boskos, D., Cortes, J., & Martinez, S. (2024). High-Confidence Data-Driven Ambiguity Sets for Time-Varying Linear Systems. *IEEE Transactions on Automatic Control*, 69(2), 797-812.  
<https://doi.org/10.1109/TAC.2023.3273815>

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

***Green Open Access added to TU Delft Institutional Repository***

***'You share, we take care!' - Taverne project***

***<https://www.openaccess.nl/en/you-share-we-take-care>***

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

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

Dimitris Boskos<sup>1</sup>, Member, IEEE, Jorge Cortés<sup>2</sup>, Fellow, IEEE, and Sonia Martínez<sup>3</sup>, Fellow, IEEE

**Abstract**—This article 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, estimation, linear system observers, stochastic systems.

## 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.

Manuscript received 27 November 2021; revised 16 November 2022; accepted 22 April 2023. Date of publication 8 May 2023; date of current version 30 January 2024. This work was supported by DARPA Lagrange program through award N66001-18-2-4027 and Grant ONR N00014-18-S-B001. Recommended by Associate Editor D. Antunes. A preliminary version of this work appeared as [DOI: 10.23919/ACC45564.2020.9147334] in the American Control Conference. (Corresponding author: Dimitris Boskos.)

Dimitris Boskos is with the Delft Center for Systems and Control, Delft University of Technology, 2628 Delft, Netherlands (e-mail: d.boskos@tudelft.nl).

Jorge Cortés and Sonia Martínez are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA 94607 USA (e-mail: cortes@ucsd.edu; soniamd@eng.ucsd.edu).

Digital Object Identifier 10.1109/TAC.2023.3273815

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 article, 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, such as expected-cost minimization and chance-constrained optimization, are the cornerstones of stochastic programming [41]. Distributionally robust versions of stochastic optimization problems [2], [5], [40] 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 [44]. 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 active current research area [16], [17], [18], [26], [45], [46].

Toward the exploitation of Wasserstein ambiguity sets for DRO problems, the work [34] introduces tractable reformulations with finite-sample guarantees, further exploited in [12], [25] 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], [20], [38]. This article [30]

exploits Wasserstein balls to robustify data-driven online optimization algorithms, and [39] 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 [48] and their data-driven extensions [49], 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 [23], [27] and economic dispatch [33], [37], [47].

Time-varying aspects of Wasserstein ambiguity sets are considered in our previous work: in [28] for dynamic traffic models, in [29] 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], 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, such as 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.

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 presentation self-contained and are omitted due to space constraints.<sup>1</sup>

## II. PRELIMINARIES

Here we present concepts from probability theory that are used throughout this article. Let  $\mathcal{B}(\mathbb{R}^d)$  denote the Borel  $\sigma$ -algebra on  $\mathbb{R}^d$ , and  $\mathcal{P}(\mathbb{R}^d)$  the probability measures on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ . For any  $p \geq 1$ ,  $\mathcal{P}_p(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} \|x\|^p d\mu < \infty\}$  is the set of probability measures in  $\mathcal{P}(\mathbb{R}^d)$  with finite  $p$ th moment. The Wasserstein distance between  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  is

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

where,  $\mathcal{H}(\mu, \nu)$  is the set of all couplings between  $\mu$  and  $\nu$ , i.e., probability measures on  $\mathbb{R}^d \times \mathbb{R}^d$  with marginals  $\mu$  and  $\nu$ , respectively. Intuitively,  $W_p(\mu, \nu)$  describes the minimum overall cost required to transfer the mass of distribution  $\mu$  from its respective locations to form distribution  $\nu$  (or vice versa), when moving unit mass from  $x$  to  $y$  has cost  $\|x - y\|^p$ . For any  $\mu \in \mathcal{P}(\mathbb{R}^d)$ , its support is the closed set  $\text{supp}(\mu) := \{x \in \mathbb{R}^d \mid \mu(U) > 0 \text{ for each neighborhood } U \text{ of } x\}$ . For a random variable  $X$  with distribution  $\mu$  we also denote  $\text{supp}(X) \equiv \text{supp}(\mu)$ . 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  $L_{\psi_p} := \{X \mid \mathbb{E}[\psi_p(|X|/t)] < \infty \text{ for some } t > 0\}$  on  $(\Omega, \mathcal{F})$ , the  $\psi_p$ -Orlicz norm (cf., [43, Section 2.7.1]) of  $X \in L_{\psi_p}$  is

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

When  $p = 1$  and  $p = 2$ , each random variable in  $L_{\psi_p}$  is subexponential and sub-Gaussian, respectively. We also denote by  $\|X\|_p \equiv (\mathbb{E}[|X|^p])^{1/p}$  the norm of a scalar random variable with finite  $p$ th 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  $p$ th norm of a vector in  $\mathbb{R}^n$  or a random variable in  $L^p$  should be clear from the context throughout this article. Given a set  $\{X_i\}_{i \in I}$  of random variables, we denote by  $\sigma(\{X_i\}_{i \in I})$  the  $\sigma$ -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  $\text{supp}(X) \subset K \subset \mathbb{R}^{n_1}$ . Then,  $\mathbb{E}[g(X, Y)] \geq \mathbb{E}[k(X)]$ .

<sup>1</sup>We adopt the following general notation. We denote by  $\|\cdot\|_p$  the  $p$ th 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  $\rho$  in  $\mathbb{R}^n$  with the  $p$ th norm,  $p \in [1, \infty]$ . 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 from the context (this avoids writing double transposes). The diameter of a set  $S \subset \mathbb{R}^n$  with the  $p$ th norm is defined as  $\text{diam}_p(S) := \sup\{\|x - y\|_p \mid x, y \in S\}$  and for  $z \in \mathbb{R}^n$ ,  $S + z := \{x + z \mid x \in S\}$ . We denote the induced Euclidean norm of a matrix  $A \in \mathbb{R}^{m \times n}$  by  $\|A\| := \max_{\|x\|=1} \|Ax\|$ . Given  $B \subset \Omega$ ,  $\mathbf{1}_B$  is the indicator function of  $B$  on  $\Omega$ , with  $\mathbf{1}_B(x) = 1$  for  $x \in B$  and  $\mathbf{1}_B(x) = 0$  for  $x \notin B$ .

### III. PROBLEM FORMULATION

Consider the stochastic optimization problem

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

where, the random variable  $\xi$  has an *unknown* distribution  $P_\xi$ . Given a finite amount of i.i.d. samples  $\xi^1, \dots, \xi^N$  from  $P_\xi$ , one can form the empirical distribution

$$P_\xi^N := \frac{1}{N} \sum_{i=1}^N \delta_{\xi^i} \quad (1)$$

to approximate  $P_\xi$ , where  $\delta_{\xi^i}$  denotes the Dirac measure at  $\xi^i$ , and solve the sample average approximation (SAA) problem

$$\inf_{x \in \mathcal{X}} \mathbb{E}_{P_\xi^N} [f(x, \xi)] = \inf_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N f(x, \xi^i).$$

Although the value of the SAA converges to the optimal value of the original problem as  $N \rightarrow \infty$  for any prescribed probability and under general assumptions (cf., [41, Chapter 5]), this is no longer guaranteed for a finite number of samples. To hedge this uncertainty, one can instead consider the DRO problem

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

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  $P_\xi$  with high confidence. We are interested in approaches that employ data, and in particular the empirical distribution  $P_\xi^N$ , 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  $\hat{P}_\xi^N$  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

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

with linear output

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

The initial condition  $\xi_0$  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  $\ell$  from multiple independent realizations of the process, denoted by  $\xi^i$ ,  $i \in [1 : N]$ . Data from independent realizations of identical systems are for instance collected in multiagent scenarios where the agents are subject to the same dynamics [50]. Populations of identical dynamical

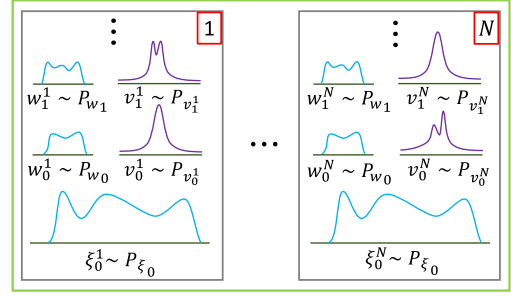


Fig. 1. Illustration of the probabilistic models for the random variables in the dynamics and observations according to Assumption 3.2.

systems are often also encountered in industrial plants, where multiple copies of the same machine are used to speed up the production process. Such machines are further often engaged in repetitive tasks, where the system is reset and repeats its control cycle from the beginning. 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  $[0 : \ell]$ . We also make certain hypotheses on the classes of the distributions  $P_{\xi_0}$  of the initial condition,  $P_{w_k}$  of the dynamics noise, and  $P_{v_k}$  of the measurement errors (cf., Fig. 1). These assumptions are made for individual realizations and allow us to consider nonidentical observation error distributions.

**Assumption 3.2 (Distribution classes):** Consider a finite sequence of realizations  $\xi^i$ ,  $i \in [1 : N]$  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  $\psi_p$ -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  $\psi_p$ -Orlicz/ $L_p$ -norm ratio):** By definition,  $\psi_p$ -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  $C_v/m_v := 0$ . We assume the specific values of the parameters  $m_v$ ,  $M_v$ , and  $C_v$  are known. These may be determined based on hardware specifications or further information about the measurement noise.  $\square$

Since the samples do not measure the full state, we aim to leverage the dynamics and estimate it from the assimilated output values. We make the next assumption to bound the state estimation errors over arbitrary evolution horizons.

**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  $\hat{\xi}_\ell^i(\zeta_0^i, \dots, \zeta_\ell^i)$  for the state of each realization and build an ambiguity set for the state distribution at time  $\ell$  with probabilistic guarantees. Using this construction as a building block, we then seek to derive ambiguity sets that contain the state distribution across a finite time horizon  $[\ell_1 : \ell_2] \subset [0 : \ell]$ . 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 start with DRO formulations at a single time instance because of their relative simplicity and the fact that they are still relevant for problems with time-varying uncertainties. For instance, in optimal power flow the operator may seek to optimally curtail PV generation during the peak hour, over which demand and generation can be conveniently modeled as constant and estimated by combining historical data with a dynamic model for their time evolution during the day.

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. In the last part of Section IV we extend the results to construct ambiguity sets over finite time horizons. 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  $\ell$ , when samples are collected from (3) according to Assumption 3.1. If we had access to  $N$  independent full-state samples  $\xi_\ell^1, \dots, \xi_\ell^N$  from the distribution of  $\xi$  at  $\ell$ , 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., [34, Theorem 3.5], to specify an ambiguity ball radius  $\varepsilon_N(\beta)$  so that the true distribution of  $\xi_\ell$  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  $\xi$  at time  $\ell$ . The dynamics of the observer, initialized at zero, is given by

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

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

$$\hat{P}_{\xi_k}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\hat{\xi}_k^i}. \quad (5)$$

Denoting by  $e_k := \xi_k - \hat{\xi}_k$  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  $\xi_0$  is the initial condition of (3a). In particular

$$e_k = \Psi_k \xi_0 + \sum_{\kappa=1}^k \left( \Psi_{k,k-\kappa+1} G_{k-\kappa} w_{k-\kappa} + \Psi_{k,k-\kappa+1} K_{k-\kappa} v_{k-\kappa} \right) \quad (6)$$

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  $\ell$ , we set its center at the estimator-based empirical distribution  $\hat{P}_{\xi_\ell}^N$  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  $1 - \beta$ .

**Remark 4.1 (Generic lack of asymptotic consistency):** Even if a distributionally robust framework is not employed, replacing the empirical distribution by the estimator-based empirical distribution in (5) does no longer guarantee consistency, in the sense that the estimator-based empirical distribution *does not necessarily converge (weakly) to the true distribution*. Hence, there is no indication that the solution to the associated estimator-based SAA problem, i.e.,

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

with  $\xi_\ell^i$  replaced by  $\hat{\xi}_\ell^i$ , 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., when one has access to full-state samples and the measurement noise is known).  $\square$

Note that the random variable  $\xi_k^i$  of a system realization at time  $k$  is a function  $\xi_k^i(\xi_0^i, \mathbf{w}_k^i)$  of the random initial condition  $\xi_0^i$  and the dynamics noise  $\mathbf{w}_k^i \equiv (w_0^i, \dots, w_{k-1}^i)$ . Analogously, the estimated state  $\hat{\xi}_k^i$  of each observer realization is a stochastic variable  $\hat{\xi}_k^i(\xi_0^i, \mathbf{w}_k^i, \mathbf{v}_k^i)$  with additional randomness induced by the output noise  $\mathbf{v}_k^i \equiv (v_0^i, \dots, v_{k-1}^i)$ . Using the compact notation  $\xi_0 \equiv (\xi_0^1, \dots, \xi_0^N)$ ,  $\mathbf{w}_k \equiv (\mathbf{w}_k^1, \dots, \mathbf{w}_k^N)$ , and  $\mathbf{v}_k \equiv (\mathbf{v}_k^1, \dots, \mathbf{v}_k^N)$  for the corresponding initial conditions, dynamics noise, and output noise of all realizations, respectively, we can denote the empirical and the estimator-based empirical distribution at time  $\ell$  as  $P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell)$  and  $\hat{P}_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell, \mathbf{v}_\ell)$ . We also denote by  $P_{\xi_\ell}$  the true distribution of the data at discrete time  $\ell$ , where from (3a)

$$\xi_\ell = \Phi_\ell \xi_0 + \sum_{k=1}^{\ell} \Phi_{\ell-k+1} G_{\ell-k} \mathbf{w}_{\ell-k} \quad (7)$$

where,  $\Phi_\ell := \Phi_{\ell,0}$  and  $\Phi_{\ell,\ell} := I$  (and with  $\Phi_{k+s,k}$  defined in Assumption 3.4). Then, it follows from **H1** and **H2** in Assumption 3.2 that the random states  $\xi_\ell^i$  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

$$\mathbb{P}(W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}) \leq \psi_N(\beta)) \geq 1 - \beta. \quad (8)$$

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

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

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

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

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*  $\hat{\varepsilon}_N(\beta_{\text{ns}})$  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 two sections, we present the precise individual bounds for these terms and then combine them to obtain the overall ambiguity radius in Section IV-C.

### 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 4.2 (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  $\mu$ . Then for any  $p \geq 1$ ,  $N \geq 1$ , and confidence  $1 - \beta$  with  $\beta \in (0, 1)$ , we have  $\mathbb{P}(W_p(\mu^N, \mu) \leq \varepsilon_N(\beta, \rho)) \geq 1 - \beta$ , where

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

$\mu^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_i}$ ,  $\rho := \frac{1}{2} \text{diam}_\infty(\text{supp}(\mu))$ ,  $h(x) := \frac{x^2}{(\ln(2+1/x))^2}$ ,  $x > 0$ , 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 [18].

**Remark 4.3 (Determination of constants in nominal ambiguity radius):** 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  $C := \frac{C_*^d}{2\sqrt{d}^d}$  and  $c := \frac{1}{2^d \sqrt{d}^d}$ , where  $C_* := \sqrt{d} 2^{(d-2)/(2p)} \left( \frac{1}{1-2^{p-d/2}} + \frac{1}{1-2^{-p}} \right)^{1/p}$ . Recent work [4], [6], [19] 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 [24], [36].  $\square$

### B. Noise Ambiguity Radius

In this section, we quantify the noise ambiguity radius  $\hat{\varepsilon}_N(\beta_{\text{ns}})$  for any prescribed confidence  $1 - \beta_{\text{ns}}$ . For the results of this section, the initial condition and the internal noise are interpreted as deterministic quantities. To clarify this distinction, we use the alternative notation  $P_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega})$  and  $\hat{P}_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega}, \mathbf{v}_\ell)$  for the corresponding empirical and estimator-based empirical distribution, where  $\mathbf{z} = (z^1, \dots, z^N)$ ,  $z^1 \equiv \xi_0^1, \dots, z^N \equiv \xi_0^N$ , and  $\boldsymbol{\omega} = (\boldsymbol{\omega}^1, \dots, \boldsymbol{\omega}^N)$ ,  $\boldsymbol{\omega}^1 \equiv \mathbf{w}_\ell^1, \dots, \boldsymbol{\omega}^N \equiv \mathbf{w}_\ell^N$ . We next uniformly bound the distance between the empirical and the estimator-based empirical distribution with prescribed confidence for all values of the initial condition and the internal noise from a set that contains the support of their joint distribution.

**Lemma 4.4 (Distance between empirical and estimator-based empirical distribution):** Let  $(\mathbf{z}, \boldsymbol{\omega}) \in B_\infty^{Nd}(\rho_{\xi_0}) \times B_\infty^{N\ell q}(\rho_w)$

and consider the discrete distribution  $P_{\xi_\ell}^N \equiv P_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega})$  and the empirical distribution  $\hat{P}_{\xi_\ell}^N \equiv \hat{P}_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega}, \mathbf{v}_\ell)$ , where  $\mathbf{v}_\ell$  is the measurement noise of the realizations. Then

$$W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}^N) \leq 2^{\frac{p-1}{p}} \mathfrak{M}_w + 2^{\frac{p-1}{p}} \left( \frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p \right)^{\frac{1}{p}} \quad (12a)$$

where

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

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

The next result gives bounds for the norms of the random variables  $\mathfrak{E}^i$  in Lemma 4.4.

**Lemma 4.5 (Orlicz- and  $L^p$ -norm bounds for  $\mathfrak{E}^i$ ):** The random variables  $\mathfrak{E}^i$  in (12c) satisfy

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

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

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

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

The proofs of both abovementioned results are given in Appendix A. By bounding the term  $(\frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p)^{\frac{1}{p}}$  in (12a) we obtain the main result of this section that bounds the Wasserstein distance between the empirical and the estimator-based empirical distribution.

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

$$\hat{\varepsilon}_N(\beta_{\text{ns}}) := 2^{\frac{p-1}{p}} \left( \mathfrak{M}_w + \mathfrak{M}_v + \mathfrak{M}_v \alpha_p^{-1} \left( \frac{\mathfrak{R}^2}{c'N} \ln \frac{2}{\beta_{\text{ns}}} \right) \right) \quad (14)$$

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

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

$$\mathfrak{R} := \mathfrak{C}_v / m_v + 1 / \ln 2 \quad (16)$$

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

$$\mathbb{P} \left( W_p(\hat{P}_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega}, \mathbf{v}_\ell), P_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega})) \leq \hat{\varepsilon}_N(\beta_{\text{ns}}) \right) \geq 1 - \beta_{\text{ns}}. \quad (17)$$

The proof is given in Appendix A.

### 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, which is the central result of this article.

**Theorem 4.7 (Ambiguity set under noisy dynamics and 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  $\hat{\varepsilon}_N(\beta_{\text{ns}})$  are given by (11) and (14), respectively, with

$$\rho_{\xi_\ell} := \sqrt{d} \|\Phi_\ell\| \rho_{\xi_0} + \sqrt{q} \sum_{k=1}^{\ell} \|\Phi_{\ell, \ell-k+1} G_{\ell-k}\| \rho_w. \quad (18)$$

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

$$\begin{aligned} \{W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}) \leq \psi_N(\beta)\} &\supset \{W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}) \leq \hat{\varepsilon}_N(\beta_{\text{ns}})\} \\ &\cap \{W_p(P_{\xi_\ell}^N, P_{\xi_\ell}) \leq \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})\}. \end{aligned}$$

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

$$\begin{aligned} \mathbb{E} \left[ \mathbf{1}_{\{W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}) - \hat{\varepsilon}_N(\beta_{\text{ns}}) \leq 0\}} \right. \\ \left. \times \mathbf{1}_{\{W_p(P_{\xi_\ell}^N, P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \leq 0\}} \right] \geq 1 - \beta. \quad (19) \end{aligned}$$

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

$$\begin{aligned} g(\xi_0, \mathbf{w}_\ell, \mathbf{v}_\ell) &:= \mathbf{1}_{\{W_p(P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell), P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \leq 0\}} \\ &\times \mathbf{1}_{\{W_p(\hat{P}_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell, \mathbf{v}_\ell), P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell)) - \hat{\varepsilon}_N(\beta_{\text{ns}}) \leq 0\}}. \end{aligned}$$

Due to (17),  $\mathbb{E}[\mathbf{1}_{\{W_p(\hat{P}_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega}, \mathbf{v}_\ell), P_{\xi_\ell}^N(\mathbf{z}, \boldsymbol{\omega})) - \hat{\varepsilon}_N(\beta_{\text{ns}}) \leq 0\}}] \geq 1 - \beta_{\text{ns}}$  for any  $x = (\mathbf{z}, \boldsymbol{\omega}) \in K$  and thus  $\mathbb{E}[g(x, Y)] \geq \mathbf{1}_{\{W_p(P_{\xi_\ell}^N(x), P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \leq 0\}} \times (1 - \beta_{\text{ns}}) =: k(x)$ , for all  $x \in K$ . Hence, since  $X \equiv (\xi_0, \mathbf{w}_\ell)$  and  $Y \equiv \mathbf{v}_\ell$  are independent by H2, we deduce from Lemma 2.1 that

$$\begin{aligned} \mathbb{E}[g(X, Y)] &\geq \mathbb{E} \left[ \mathbf{1}_{\{W_p(P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell), P_{\xi_\ell}) - \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell}) \leq 0\}} (1 - \beta_{\text{ns}}) \right] \\ &= (1 - \beta_{\text{ns}}) \mathbb{P}(W_p(P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell), P_{\xi_\ell}) \leq \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})). \end{aligned}$$

From (7) and H3 in Assumption 3.2, it follows that  $P_{\xi_\ell}$  is supported on the compact set  $B_{\infty}^d(\rho_{\xi_\ell})$  with  $\text{diam}_{\infty}(B_{\infty}^d(\rho_{\xi_\ell})) = 2\rho_{\xi_\ell}$  and  $\rho_{\xi_\ell}$  given in (18). In addition, due to H1 and H2 in Assumption 3.2 the random states  $\xi_\ell^i$  in the empirical distribution  $P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell) = \frac{1}{N} \sum_{i=1}^N \delta_{\xi_\ell^i}$  are i.i.d.. Thus, we get from Proposition 4.2 that  $\mathbb{P}(W_p(P_{\xi_\ell}^N(\xi_0, \mathbf{w}_\ell), P_{\xi_\ell}) \leq \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})) \geq 1 - \beta_{\text{nom}}$ , which implies  $\mathbb{E}[g(X, Y)] \geq (1 - \beta_{\text{ns}})(1 - \beta_{\text{nom}}) = 1 - \beta$ . Finally, (19) follows from this and the definition of  $g$ .

With this result at hand, we deduce from the expressions (11) and (17) 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  $\psi_N$  generalizes the nominal ambiguity radius  $\varepsilon_N$  in the DRO literature (even when dynamic random variables are considered [9]) and reduces to  $\varepsilon_N$  in the noise-free case where  $\hat{\varepsilon}_N = 0$ .

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  $\ell$ , 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 (14) 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.  $\square$

**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

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

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.  $\square$

**Remark 4.10 (Potential conservativeness of the overall radius):** It is worth noting that the ambiguity radius  $\psi_N$  can be rather conservative. The main reasons for this are the following.

- 1) Conservativeness of the concentration of measure results used for the derivation of the nominal radius.
- 2) Potential lack of homogeneity of the distribution's support since some components of the state distribution may have a much smaller support than others.
- 3) Potential independence between components of the state distribution, which is not exploited.
- 4) Conservative upper bounds for the estimation error.

Although there is room to sharpen all these aspects, they require additional contributions that are beyond the scope of this article.

A practical approach to select the ambiguity radius in a data-driven manner for a fixed (single) optimization problem would consist of two steps. The first step will use a technique, such as cross-validation to tune the radius based on the states of the estimator-based empirical distribution [34, Section 7]. The second step will enlarge this radius using i.i.d.

data of the discrepancy  $e_k^j$ ,  $j = 1, \dots, M$  between the true and the estimated state (provided that these are available), where  $M = mN$  for some  $m \in \mathbb{N}$ . This enlargement is analogous to  $\frac{1}{m}(\frac{1}{N} \sum_{j=(m-1)N+1}^{mN} \|e_k^j\|^p)^{\frac{1}{p}}$ , which is an estimator for the average Wasserstein distance between the true and the estimator-based empirical distribution.  $\square$

**Remark 4.11 (Chance-constrained problems):** The ambiguity sets of this article can also be used in chance-constrained problems where the samples are no longer drawn from the true distribution because of the type of nonidealities presented here. Then, one can directly use the formulations of e.g., [25], by replacing the empirical distribution there with the estimator-based empirical distribution and tuning the ambiguity radius according to Theorem 4.7 or the practical approach in Remark 4.10.  $\square$

#### 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  $\ell$ . In this case we assume again that samples are collected over the time window  $[0 : \ell]$  but we seek to build an ambiguity set about the state distribution along  $[\ell_1 : \ell_2]$ , with  $0 \leq \ell_1 \leq \ell_2 \leq \ell$ . 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  $\xi_\ell := (\xi_{\ell_1}, \dots, \xi_{\ell_2}) \in \mathbb{R}^{\tilde{\ell}d}$  with  $\ell := (\ell_1, \dots, \ell_2)$  and  $\tilde{\ell} = \ell_2 - \ell_1 + 1$ , comprising of all states over the interval of interest and using the concentration of measure result of Proposition 4.2 for  $\tilde{\ell}d$ -dimensional random variables. This has the drawback that the ambiguity radius decays slowly with the number of trajectories due to the high dimension of  $\xi_\ell$ . The other description derives an ambiguity set about the state distribution  $P_{\xi_{\ell_1}}$  at time  $\ell_1$  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 4.12 (Ambiguity set over a bounded time horizon):** Consider output data collected from  $N$  realizations of system (3) over the interval  $[0 : \ell]$  and let Assumption 3.2 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

$$\hat{P}_{\xi_\ell}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\xi_\ell^i}$$

over the horizon  $[\ell_1 : \ell_2] \subset [0 : \ell]$ , where  $\hat{\xi}_\ell^i := (\hat{\xi}_{\ell_1}^i, \dots, \hat{\xi}_{\ell_2}^i) \in \mathbb{R}^{\tilde{\ell}d}$  and each  $\hat{\xi}_\ell^i$  is given by the observer (4). Then

$$\mathbb{P}(W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}) \leq \psi_N(\beta)) \geq 1 - \beta \quad (20)$$

holds, where  $\xi_\ell := (\xi_{\ell_1}, \dots, \xi_{\ell_2})$  and  $\psi_N(\beta)$  is given in (10). The nominal component  $\varepsilon_N(\beta_{\text{nom}}) \equiv \varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})$  is given

by (11) (with  $d$  in the expression substituted by  $\tilde{\ell}d$ )

$$\rho_{\xi_\ell} := \max_{\ell \in [\ell_1 : \ell_2]} \left\{ \sqrt{d} \|\Phi_\ell\| \rho_{\xi_0} + \sqrt{q} \sum_{k=1}^{\ell} \|\Phi_{\ell, \ell-k+1} G_{\ell-k}\| \rho_w \right\} \quad (21)$$

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

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

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

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

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.,

$$\mathbb{P} \left( P_{\xi_\ell} \in \mathcal{B}_{\psi_{N,\ell}}(\tilde{P}_{\xi_\ell}^N) \forall \ell \in [\ell_1 : \ell_2] \right) \geq 1 - \beta \quad (22)$$

where,  $\mathcal{B}_{\psi_{N,\ell}}(\tilde{P}_{\xi_\ell}^N) := \{P \in \mathcal{P}_p(\mathbb{R}^d) \mid W_p(P, \tilde{P}_{\xi_\ell}^N) \leq \psi_{N,\ell}\}$  and  $\tilde{P}_{\xi_\ell}^N$  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

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

To obtain the result in this case and retain sufficiently sharp bounds for the ambiguity radius, we partially strengthen our assumptions **H2** and **H3** about the internal noise distribution. We assume that the sequence  $w_k$  is independent and that we either know a (uniform) bound on its  $p$ th moment or that we know its precise distribution. The technical approach is substantially different from that of Theorem 4.12 and relies on the notion of the convolution  $\mu \star \nu$  of two distributions  $\mu$  and  $\nu$  on  $\mathbb{R}^d$ , which is the image of their product measure  $\mu \otimes \nu$  on  $\mathbb{R}^d \times \mathbb{R}^d$  under the mapping  $(x, y) \mapsto x + y$ ; equivalently,  $\mu \star \nu(B) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbf{1}_B(x + y) \mu \otimes \nu(dx, dy)$  for any  $B \in \mathcal{B}(\mathbb{R}^d)$  (cf., [7, Pages 207, 208]).

**Theorem 4.13 (Pointwise ambiguity sets over a bounded time horizon):** Let the assumptions of Theorem 4.12 hold, assume that the internal noise sequence  $w_\ell$  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)  $P_{w_\ell}$  is not known and  $\mathbb{E}[\|w_\ell\|^p] \leq q_w$ .
- ii)  $P_{w_\ell}$  is known.

Then, for any confidence  $1 - \beta$ , and  $\beta_{\text{nom}}, \beta_{\text{ns}} \in (0, 1)$  satisfying (9), (22) holds, with  $\tilde{P}_{\xi_{\ell_1}}^N := \tilde{P}_{\xi_{\ell_1}}^N$  and  $\psi_{N,\ell_1}$  as given

by Theorem 4.7 (for  $\ell \equiv \ell_1$ ), and  $\tilde{P}_{\xi_\ell}^N$ ,  $\psi_{N,\ell}$ ,  $\ell \in [\ell_1 + 1 : \ell_2]$  defined as follows for the respective abovementioned two cases.

- i) The ambiguity set center is  $\tilde{P}_{\xi_\ell}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{\xi}_\ell^i}$  with  $\tilde{\xi}_\ell^i := \Phi_{\ell, \ell_1} \hat{\xi}_{\ell_1}^i$  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  $\tilde{P}_{\xi_\ell}^N := ((A_{\ell-1})_{\#} \tilde{P}_{\xi_{\ell-1}}^N) \star P_{w_{\ell-1}}$  and the radius is  $\psi_{N,\ell} := \|A_{\ell-1}\| \dots \|A_{\ell_1}\| \psi_{N,\ell_1}$ .

The proof is given in Appendix B. Note that when the internal noise distribution is known, all individual ambiguity balls of Theorem 4.13 shrink at the exact same decay rate with the number of sampled trajectories, which overcomes the slow decay rate of the ambiguity radius of Theorem 4.12 for larger time horizons.

## 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. The focus here is on the noise component because it can be controlled under appropriate observability assumptions, whereas the nominal radius grows unbounded with time for unstable dynamics. 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  $K_*, K^*, G^* > 0$  and  $\Psi_s^* > 0$ ,  $s \in \mathbb{N}_0$ , so that  $\|G_k\| \leq G^*$ ,  $K_* \leq \|K_k\| \leq K^*$ , and  $\|\Psi_{k+s,k}\| \leq \Psi_s^*$  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  $s \geq s_0$ .

The proof is given in Appendix C. 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 (14) 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 (16), are uniformly upper bounded as

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

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

The proof is given in the online version [10].

**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

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

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

with  $\ell_0$  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, respectively. 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.  $\square$

## 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  $\delta t$  of the continuous-time domain. To this end, each generator and storage unit supplies the network with positive power  $P^j$  and  $S^i$ , respectively, at time  $\ell$ . We assume we can control the power of the generators, which additionally needs to be within the upper and lower thresholds  $P_{\min}^j$  and  $P_{\max}^j$ , respectively. Each battery is modeled as an uncertain dynamic element with an unknown initial state distribution and we can decide whether it is connected ( $\eta^i = 1$ ) or not ( $\eta^i = 0$ ) to the network at time  $\ell$ . 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

$$\mathcal{C}(\mathbf{P}, \boldsymbol{\eta}) := \sum_{j=1}^{n_1} g^j(P^j) + \sum_{i=1}^{n_2} \eta^i h^i(S^i) + c \left( \sum_{j=1}^{n_1} P^j + \sum_{i=1}^{n_2} \eta^i S^i - D \right)^2 \quad (23)$$

where,  $\mathbf{P} := (P^1, \dots, P^{n_1})$ ,  $\boldsymbol{\eta} := (\eta^1, \dots, \eta^{n_2})$ , and  $g^j, h^i$  are cost functions for the power provided by generator  $j$  and storage

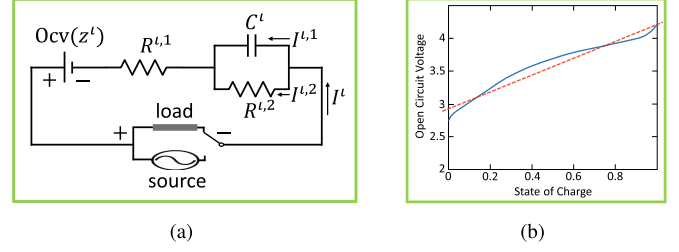


Fig. 2. (a) shows the equivalent circuit model of a lithium-ion battery cell in discharging mode (cf., [32, Fig. 2], [31, Fig. 1]). (b) is taken from [31, Fig. 3] and shows the nonlinear dependence of the open circuit voltage on the SoC and its affine approximation.

unit  $l$ , 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 (23). Due to the uncertainty about the batteries' state and their injected powers  $S^i$ , the minimization of (23) 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  $I^l$  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 Fig. 2(a) (see, e.g., [31], [32]), where  $z^l$  is the state of charge (SoC) of the cell and  $\text{Ocv}(z^l)$  is its corresponding open-circuit voltage, which we approximate by the affine function  $\alpha^l z^l + \beta^l$  in Fig. 2(b). The associated discrete-time cell model is

$$\chi_{k+1}^l \equiv \begin{pmatrix} I_{k+1}^{l,2} \\ z_{k+1}^l \end{pmatrix} = \begin{pmatrix} a^l & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} I_k^{l,2} \\ z_k^l \end{pmatrix} + \begin{pmatrix} 1 - a^l \\ -\delta t / Q^l \end{pmatrix} I_k^l$$

$$\theta_k^l \equiv V_k^l = \alpha^l z_k^l + \beta^l - I_k^l R^{l,1} - I_k^{l,2} R^{l,2}$$

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

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

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  $\chi_k^{l,*} = (I_k^{l,2,*}, z_k^{l,*})$  initiated from the center of the support of its initial-state distribution. Setting  $\xi_k^l = \chi_k^l - \chi_k^{l,*}$  and  $\zeta_k^l = \theta_k^l(\chi_k^l) - \theta_k^l(\chi_k^{l,*})$

$$\xi_{k+1}^l = A_k^l \xi_k^l$$

$$\zeta_k^l = H_k^l \xi_k^l + v_k$$

where,  $A_k^t := \text{diag}(a, 1)$  and  $H_k^t := (\alpha^t, -R^{t,2})$ . Denoting  $\xi := (\xi^1, \dots, \xi^{n_2})$  and  $\zeta := (\zeta^1, \dots, \zeta^{n_2})$ , we obtain a system of the form (3) for the dynamic random variable  $\xi$ . Despite the fact that the state distribution  $\xi_k$  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 this article to build an ambiguity ball  $\mathcal{P}^N$  of radius  $\varepsilon_N$  in the 2-Wasserstein distance (i.e., with  $p = 2$ ), that contains the batteries' state distribution  $P_{\xi_\ell}$  at time  $\ell$  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  $\hat{\xi}_\ell^i$  at time  $\ell$ . The ambiguity set is centered at the estimator-based empirical distribution  $\hat{P}_{\xi_\ell}^N = \frac{1}{N} \sum_{i=1}^N \delta_{\hat{\xi}_\ell^i}$  and its radius can be determined using Theorem 4.7 and Proposition 4.6.

### C. Decision Problem as DRO Problem

To solve the decision problem regarding whether or not to connect the batteries for economic dispatch, we formulate a DRO problem for the cost (23) 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  $\xi_\ell$ . Notice first that the power injected by each battery at time  $\ell$  is

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

where  $\langle \cdot, \cdot \rangle$  denotes inner product,  $\hat{\alpha}^t := (-I_\ell^t R^{t,2}, \alpha^t I_\ell^t)$  and

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

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

$$\begin{aligned} \mathcal{C}(\mathbf{P}, \boldsymbol{\eta}) &= g(\mathbf{P}) + (\boldsymbol{\eta} * \tilde{\alpha})^\top \xi_\ell + \boldsymbol{\eta}^\top \tilde{\beta} \\ &\quad + c \left( \mathbf{1}^\top \mathbf{P} + (\boldsymbol{\eta} * \hat{\alpha})^\top \xi_\ell + \boldsymbol{\eta}^\top \hat{\beta} - D \right)^2 \end{aligned} \quad (24)$$

where,  $*$  denotes the Khatri-Rao product (i.e., for  $\mathbf{a} \equiv (a^1, \dots, a^d) \in \mathbb{R}^d$  and  $\mathbf{b} \equiv (b^1, \dots, b^d) \in \mathbb{R}^{dn}$  with  $b^i \in \mathbb{R}^n$ ,  $\mathbf{a} * \mathbf{b} := (a^1 b^1, \dots, a^d b^d)$ ) and

$$\begin{aligned} g(\mathbf{P}) &:= \sum_{j=1}^{n_1} g^j(P^j), \quad \hat{\alpha} := (\hat{\alpha}^1, \dots, \hat{\alpha}^{n_2}) \\ \hat{\beta} &:= (\hat{\beta}^1, \dots, \hat{\beta}^{n_2}), \quad \tilde{\alpha} := (\bar{\alpha}^1 \hat{\alpha}^1, \dots, \bar{\alpha}^{n_2} \hat{\alpha}^{n_2}) \\ \tilde{\beta} &:= (\bar{\alpha}^1 \hat{\beta}^1 + \bar{\beta}^1, \dots, \bar{\alpha}^{n_2} \hat{\beta}^{n_2} + \bar{\beta}^{n_2}). \end{aligned}$$

Using the equivalent description (24) 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

$$\inf_{\boldsymbol{\eta}, \mathbf{P}} \left\{ f_{\boldsymbol{\eta}}(\mathbf{P}) + \sup_{P_{\xi_\ell} \in \mathcal{P}^N} \mathbb{E}_{P_{\xi_\ell}} [h_{\boldsymbol{\eta}}(\mathbf{P}, \xi_\ell)] \right\} \quad (25a)$$

$$\text{s.t. } P_{\min}^j \leq P^j \leq P_{\max}^j \quad \forall j \in [1 : n_1] \quad (25b)$$

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

$$\begin{aligned} f_{\boldsymbol{\eta}}(\mathbf{P}) &:= g(\mathbf{P}) + c \mathbf{P}^\top \mathbf{1} \mathbf{P} \\ &\quad + 2c(\boldsymbol{\eta}^\top \hat{\beta} - D) \mathbf{1}^\top \mathbf{P} + c(\boldsymbol{\eta}^\top \hat{\beta} - D)^2 + \boldsymbol{\eta}^\top \tilde{\beta} \\ h_{\boldsymbol{\eta}}(\mathbf{P}, \xi_\ell) &:= c \xi_\ell^\top (\boldsymbol{\eta} * \hat{\alpha}) (\boldsymbol{\eta} * \hat{\alpha})^\top \xi_\ell + \left( 2c \left( \mathbf{1}^\top \mathbf{P} \right. \right. \\ &\quad \left. \left. + \boldsymbol{\eta}^\top \hat{\beta} - D \right) (\boldsymbol{\eta} * \hat{\alpha})^\top + (\boldsymbol{\eta} * \tilde{\alpha})^\top \right) \xi_\ell. \end{aligned}$$

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

### D. Tractable Reformulation of the DRO Problem

Our next goal is to obtain a tractable reformulation of the optimization problem (25). To this end, we first provide an equivalent description for the inner maximization in (25), which is carried out over a space of probability measures. Exploiting strong duality (see [21, 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  $\sup_{P_{\xi_\ell} \in \mathcal{P}^N} \mathbb{E}_{P_{\xi_\ell}} [h_{\boldsymbol{\eta}}(\mathbf{P}, \xi_\ell)]$  as

$$\inf_{\lambda \geq 0} \left\{ \lambda \psi_N^2 + \frac{1}{N} \sum_{i=1}^N \sup_{\xi_\ell \in \Xi} \{ h_{\boldsymbol{\eta}}(\mathbf{P}, \xi_\ell) - \lambda \|\xi_\ell - \hat{\xi}_\ell^i\|^2 \} \right\} \quad (26)$$

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  $\hat{\xi}_\ell^i$  are the estimated states of their realizations. We slightly relax the problem, by allowing the ambiguity ball to contain all distributions with distance  $\psi_N$  from  $\hat{P}_{\xi_\ell}^N$  that are supported on  $\mathbb{R}^{2n_2}$  and not necessarily on  $\Xi$ .<sup>2</sup> Thus, we first look to solve for each estimated state  $\hat{\xi}_\ell^i$  the optimization problem

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

which is written

$$\begin{aligned} &\sup_{\xi_\ell \in \mathbb{R}^{2n_2}} \left\{ \xi_\ell^\top \mathfrak{A} \xi_\ell + \left( 2c \left( \mathbf{1}^\top \mathbf{P} + \boldsymbol{\eta}^\top \hat{\beta} - D \right) (\boldsymbol{\eta} * \hat{\alpha})^\top \right. \right. \\ &\quad \left. \left. + (\boldsymbol{\eta} * \tilde{\alpha})^\top \right) \xi_\ell - \lambda (\xi_\ell - \hat{\xi}_\ell^i)^\top (\xi_\ell - \hat{\xi}_\ell^i) \right\} \\ &= -\lambda (\hat{\xi}_\ell^i)^\top \hat{\xi}_\ell^i + \sup_{\xi_\ell \in \mathbb{R}^{2n_2}} \left\{ \xi_\ell^\top (\mathfrak{A} - \lambda I_{2n_2}) \xi_\ell + (\mathbf{r}^i)^\top \xi_\ell \right\} \end{aligned}$$

where,  $\mathbf{r}^i \equiv \mathbf{r}_\eta^i(\mathbf{P}, \lambda) := 2c(\mathbf{1}^\top \mathbf{P} + \boldsymbol{\eta}^\top \hat{\beta} - D)(\boldsymbol{\eta} * \hat{\alpha}) + \boldsymbol{\eta} * \tilde{\alpha} + 2\lambda \hat{\xi}_\ell^i$  and  $\mathfrak{A} \equiv \mathfrak{A}_\eta := c(\boldsymbol{\eta} * \hat{\alpha})(\boldsymbol{\eta} * \hat{\alpha})^\top$  is a symmetric positive semidefinite matrix with diagonalization  $\mathfrak{A} = \mathbf{\Omega}^\top \mathbf{\mathcal{D}} \mathbf{\Omega}$  where the eigenvalues decrease along the diagonal. Hence, we get  $\sup_{\xi_\ell \in \mathbb{R}^{2n_2}} \{ \xi_\ell^\top (\mathfrak{A} - \lambda I_{2n_2}) \xi_\ell + (\mathbf{r}^i)^\top \xi_\ell \} = \sup_{\xi \in \mathbb{R}^{2n_2}} \{ \xi^\top (\mathfrak{D} - \lambda I_{2n_2}) \xi + (\hat{\mathbf{r}}^i)^\top \xi \}$  with  $\hat{\mathbf{r}}^i := \mathbf{\Omega} \mathbf{r}^i$  and

<sup>2</sup>The exact same reformulations can be obtained for any quadratic cost function  $\xi_\ell^\top Q \xi_\ell + b(\mathbf{P})^\top \xi_\ell + c$ , with  $Q$  positive semidefinite and  $b(\mathbf{P})$  affine in  $\mathbf{P}$ . Analogous reformulations for dynamic programming problems with quadratic costs can be found in [49].

denoting  $\lambda_{\max}(\mathfrak{A})$  the maximum eigenvalue of  $\mathfrak{A}$  we have

$$\sup_{\xi \in \mathbb{R}^{2n_2}} \{ \xi^\top (\mathfrak{D} - \lambda I_{2n_2}) \xi + (\hat{\mathbf{r}}^i)^\top \xi \} \\ = \begin{cases} \infty & \text{if } 0 \leq \lambda < \lambda_{\max}(\mathfrak{A}) \\ \frac{1}{4}(\hat{\mathbf{r}}^i)^\top (\lambda I_{2n_2} - \mathfrak{D})^{-1} \hat{\mathbf{r}}^i & \text{if } \lambda > \lambda_{\max}(\mathfrak{A}). \end{cases} \quad (27)$$

To obtain this we exploited that  $Q(\xi) := \xi^\top (\mathfrak{D} - \lambda I_{2n_2}) \xi + (\hat{\mathbf{r}}^i)^\top \xi$  is maximized when  $\nabla Q(\xi_\star) = 0 \iff \xi_\star = \frac{1}{2}(\lambda I_{2n_2} - \mathfrak{D})^{-1} \hat{\mathbf{r}}^i$  which gives the optimal value  $Q(\xi_\star) = \frac{1}{4}(\hat{\mathbf{r}}^i)^\top (\lambda I_{2n_2} - \mathfrak{D})^{-1} \hat{\mathbf{r}}^i$ . Note that we do not need to specify the value of the expression in (27) for  $\lambda = \lambda_{\max}$ . In particular, since the function we minimize in (26) is convex in  $\lambda$ , the inner part of the DRO problem is equivalently written

$$\inf_{\lambda > \lambda_{\max}(\mathfrak{A})} \left\{ \lambda \left( \psi_N^2 - \frac{1}{N} \sum_{i=1}^N (\hat{\xi}_\ell^i)^\top \hat{\xi}_\ell^i \right) + \frac{1}{4N} \sum_{i=1}^N \hat{\mathbf{r}}_\eta^i(P, \lambda)^\top (\lambda I_{2n_2} - \mathfrak{D})^{-1} \hat{\mathbf{r}}_\eta^i(P, \lambda) \right\}.$$

Taking further into account that  $(\lambda I_{2n_2} - \mathfrak{D})^{-1} = \text{diag} \left( \frac{1}{\lambda - \lambda_{\max}(\mathfrak{A})}, \dots, \frac{1}{\lambda - \lambda_{\min}(\mathfrak{A})} \right)$  as well as the constraints (25b) on the decision variable  $\mathbf{P}$ , the overall DRO problem is reformulated as

$$\min_{\eta} \inf_{\mathbf{P}, \lambda} \left\{ f_\eta(\mathbf{P}) + \lambda \left( \psi_N^2 - \frac{1}{N} \sum_{i=1}^N (\hat{\xi}_\ell^i)^\top \hat{\xi}_\ell^i \right) + \frac{1}{4N} \sum_{i=1}^N \hat{\mathbf{r}}_\eta^i(P, \lambda)^\top \times \text{diag} \left( \frac{1}{\lambda - \lambda_{\max}(\mathfrak{A})}, \dots, \frac{1}{\lambda - \lambda_{\min}(\mathfrak{A})} \right) \hat{\mathbf{r}}_\eta^i(P, \lambda) \right\} \quad (28a)$$

$$\text{subject to } P_{\min}^j \leq P^j \leq P_{\max}^j \quad \forall j \in [1 : n_1] \\ \lambda > \lambda_{\max}(\mathfrak{A}). \quad (28b)$$

## 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  $z_0^i$  and current  $I_0^{\iota,2}$  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  $P_{z_0^2} = P_{z_0^3} = \mathcal{U}[0.45, 0.65]$  ( $\mathcal{U}$  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  $P_{z_0^{1A}} = \mathcal{U}[0.45, 0.65]$ , whereas that of battery 1B is  $P_{z_0^{1B}} = \mathcal{U}[0.84, 0.86]$ . Thus, we get the bimodal distribution  $P_{z_0^1} = 0.9\mathcal{U}[0.45, 0.65] + 0.1\mathcal{U}[0.84, 0.86]$ , which is responsible for nonnegligible 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  $I_0^{\iota,2}$  of all batteries are fixed to 1.6308, namely,  $P_{I_0^{1,2}} = P_{I_0^{2,2}} = P_{I_0^{3,2}} = \delta_{1.6308}$ . 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  $\mu$  and variance  $\sigma^2$ . To compute the ambiguity radius for the reformulated DRO problem (28), we specify its nominal and noise components  $\varepsilon_N(\beta_{\text{nom}}, \rho_{\xi_\ell})$  and  $\hat{\varepsilon}_N(\beta_{\text{ns}})$ , where due to Proposition 4.2,  $\rho_{\xi_\ell}$  can be selected as half the diameter of any set containing the support of  $P_{\xi_\ell}$  in the infinity norm. It follows directly from the specific dynamics of the batteries that  $\rho_{\xi_\ell}$  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  $\rho_{\xi_\ell} = 0.225$ , we obtain

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

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  $\psi_2$  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  $\mathbb{E}_{P_i}[v_k^2] = \mu_i^2 + \mathbb{E}_{P_i}[(v_k - \mu_i)^2] = \mu_i^2 + \sigma_i^2$ . 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  $z_0^i$  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  $C_v$  previously, we obtain from (12b), (13a)–(13c), and (16) 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  $\hat{\varepsilon}_N(\beta_{\text{ns}}) = 0.47 + 0.0113\sqrt{74.98/N \ln(2/\beta_{\text{ns}})}$  and the overall radius is

$$\psi_N(\beta) = 0.47 + 4.02 N^{-\frac{1}{6}} + 1.31(\ln \beta_{\text{nom}}^{-1})^{\frac{1}{4}} N^{-\frac{1}{4}} + 0.0973(\ln(2\beta_{\text{ns}}^{-1}))^{\frac{1}{2}} N^{-\frac{1}{2}}. \quad (29)$$

We assume that the energy cost of the generators is lower than that of the batteries and select the quadratic power generation cost  $g(\mathbf{P}) = 0.25 \sum_{j=1}^4 (P^j - 0.1)^2$  and the same lower/upper power thresholds  $P_{\min}^j = 0.2/P_{\max}^j = 0.5$  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.

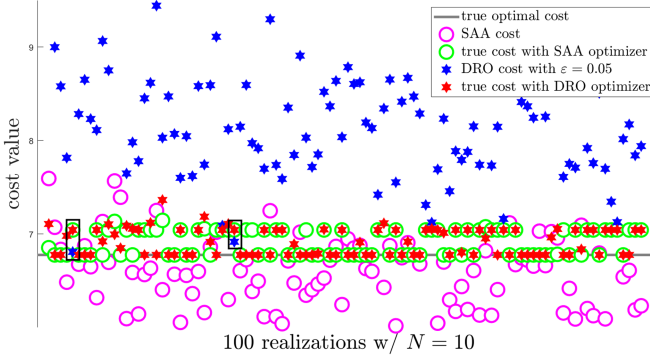


Fig. 3. Results from 100 realizations of the power dispatch problem with  $N = 10$  independent samples used for each realization. We compute the optimizers of the SAA and DRO problems, plot their corresponding optimal values (termed “SAA cost” and “DRO cost”), and also evaluate their performance with 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 overpromise since its value is most frequently below the true cost of the SAA optimizer.

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 SAA and DRO problems for  $N = 10$ ,  $N = 40$ , and  $N = 160$  samples, respectively, using CVX [22]. Despite the potential conservativeness of the ambiguity radius  $\psi_N$  (cf. Remark 4.10), formula (29) gives a qualitative intuition about its decay rate. 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 Figs. 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 [34, 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 optimizer compared to the one obtained by the nonrobust SAA, because it considerably improves the true cost for a significant number of realizations (cf., Fig. 4).

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 commitments at a much higher cost than anticipated might result in significant losses, and not fulfilling commitments may lead to penalties from the system operator.

Authorized licensed use limited to: TU Delft Library. Downloaded on February 13, 2024 at 09:39:49 UTC from IEEE Xplore. Restrictions apply.

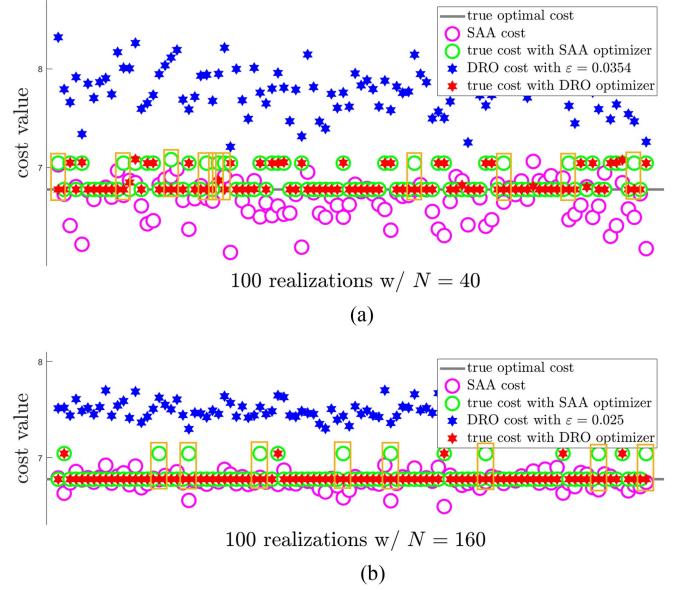


Fig. 4. Analogous results to those of Fig. 3, from 100 realizations with (a)  $N = 40$  and (b)  $N = 160$  independent samples, and the ambiguity radius tuned so that the same confidence level is preserved. In both cases, the DRO cost is above the true cost of the DRO optimizer with high probability (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).

## VII. CONCLUSION

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. 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.

## APPENDIX

Here we give proofs of various results of this article. The proofs are organized per section.

### A. Proofs From Section IV-B

*Proof of Lemma 4.4:* Using [9, Lemma A.2] to bound the Wasserstein distance of two discrete distributions, we get  $W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}^N) \leq (\frac{1}{N} \sum_{i=1}^N \|\hat{\xi}_\ell^i - \xi_\ell^i\|^p)^{\frac{1}{p}} = (\frac{1}{N} \sum_{i=1}^N \|e_\ell^i\|^p)^{\frac{1}{p}}$ . From (6), we have

$$\|e_\ell^i\| = \left\| \Psi_\ell z^i + \sum_{k=1}^{\ell} \left( \Psi_{\ell, \ell-k+1} G_{\ell-k} \omega_{\ell-k}^i + \Psi_{\ell, \ell-k+1} K_{\ell-k} v_{\ell-k}^i \right) \right\|$$

$$\leq \|\Psi_\ell\| \|z^i\| + \sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} G_{\ell-k}\| \|\omega_{\ell-k}^i\| \\ + \sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\| \|v_{\ell-k}^i\|_1 =: \mathfrak{M}(z^i, \omega^i) + \mathfrak{E}(v^i)$$

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

$$W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}^N) \leq \left( \frac{1}{N} 2^{p-1} \sum_{i=1}^N (\mathfrak{M}(z^i, \omega^i)^p + (\mathfrak{E}^i)^p) \right)^{\frac{1}{p}}.$$

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

$$W_p(\hat{P}_{\xi_\ell}^N, P_{\xi_\ell}^N) \leq \left( \frac{1}{N} 2^{p-1} \sum_{i=1}^N \mathfrak{M}(z^i, \omega^i)^p \right)^{\frac{1}{p}} \\ + \left( \frac{1}{N} 2^{p-1} \sum_{i=1}^N (\mathfrak{E}^i)^p \right)^{\frac{1}{p}}. \quad (30)$$

Finally, since  $(z, \omega) \in B_\infty^{Nd}(\rho_{\xi_0}) \times B_\infty^{N\ell q}(\rho_w)$ , we get  $\mathfrak{M}(z^i, \omega^i)^p \leq \|\Psi_\ell\| \sqrt{d} \|z^i\|_\infty + \sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} G_{\ell-k}\| \sqrt{q} \|\omega_{\ell-k}^i\|_\infty \leq \mathfrak{M}_w$ . This combined with (30) yields (12a).

*Proof of Lemma 4.5:* From **H4** in Assumption 3.2, we obtain for each summand in (12c)

$$\|\Psi_{\ell,\ell-k+1} K_{\ell-k}\| \|v_{\ell-k}^i\|_1 \|\psi_p \\ \leq \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\| \left( \|v_{\ell-k,1}^i\|_{\psi_p} + \dots + \|v_{\ell-k,r}^i\|_{\psi_p} \right) \\ \leq C_v r \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\|.$$

Hence, we deduce that

$$\|\mathfrak{E}^i\|_{\psi_p} \leq \sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\| \|v_{\ell-k}^i\|_1 \|\psi_p \\ \leq C_v r \sum_{k=1}^{\ell} \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\|.$$

For the  $L^p$  bounds, note that  $\|\mathfrak{E}^i\|_p = \|\sum_{k \in [1:\ell], l \in [1:r]} \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\| \|v_{\ell-k,l}^i\|_1\|_p$ . Thus, from the inequality  $\|\sum_i c_i X_i\|_p \leq \sum_i c_i \|X_i\|_p$ , which holds for any nonnegative  $c_i$  and  $X_i$  in  $L^p$

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

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  $(\sum_i c_i)^p \geq \sum_i c_i^p$ , which holds for any nonnegative  $c_i$ , to get

$$(\mathbb{E}[(\mathfrak{E}^i)^p])^{\frac{1}{p}} \\ \geq \left( \sum_{k \in [1:\ell], l \in [1:r]} \|\Psi_{\ell,\ell-k+1} K_{\ell-k}\|^p \mathbb{E}[\|v_{\ell-k,l}^i\|^p] \right)^{\frac{1}{p}}.$$

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

To prove Proposition 4.6, we exploit the following concentration of measure result around the mean of nonnegative independent random variables.

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

$$\mathbb{P}\left(\left|\frac{1}{N} \sum_{i=1}^N X_i^p\right|^{\frac{1}{p}} - 1 \geq t\right) \leq 2 \exp\left(-\frac{c' N}{R^2} \alpha_p(t)\right) \quad (31)$$

for every  $t \geq 0$ , with  $c' = 1/10$ ,  $R := \max_{i \in [1:N]} \|X_i\|_{\psi_p} + 1/\ln 2$ , and  $\alpha_p$  as given by (15).

We prove Proposition A.1 along the lines of the proof of [43, Theorem 3.1.1], which considers the special case of sub-Gaussian distributions. We rely on the following concentration inequality [43, Corollary 2.8.3].

*Proposition A.2 (Bernstein inequality):* Let  $X_1, \dots, X_N$  be scalar, mean-zero, subexponential, independent random variables. Then, for every  $t \geq 0$  we have

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

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

The precise constant  $c'$  in this result is not specified in [43] but we provide an independent proof of this result in the online version [10, Section 8.2].

*Proof of Proposition A.1:* 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 A.2 that

$$\mathbb{P}\left(\left|\frac{1}{N} \sum_{i=1}^N X_i^p - 1\right| \geq t\right) \leq 2 \exp\left(-\frac{c' N}{R^2} \min\{t^2, t\}\right) \quad (32)$$

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  $\alpha_p$  as given by (15). ◁

By exploiting Fact II, we get

$$\mathbb{P}\left(\left|\left(\frac{1}{N} \sum_{i=1}^N X_i^p\right)^{\frac{1}{p}} - 1\right| \geq t\right) \\ \leq \mathbb{P}\left(\left|\frac{1}{N} \sum_{i=1}^N X_i^p - 1\right| \geq \max\{t, t^p\}\right)$$

$$\leq 2 \exp \left( -\frac{c'N}{R^2} \min\{\max\{t, t^p\}^2, \max\{t, t^p\}\} \right).$$

Thus, since  $\mathbb{P}(|Y| \geq t) \geq \mathbb{P}(Y \geq t)$  for any random variable  $Y$ , we obtain (31) from Fact III and conclude the proof.

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

*Proof of Proposition 4.6:* 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 A.1 that

$$\mathbb{P}\left(\left(\frac{1}{N} \sum_{i=1}^N \left(\frac{\mathfrak{E}^i}{\|\mathfrak{E}^i\|_p}\right)^p\right)^{\frac{1}{p}} - 1 \geq t\right) \leq 2 \exp\left(-\frac{c'N}{R^2} \alpha_p(t)\right)$$

where,  $R = \max_{i \in [1:N]} \|\mathfrak{E}^i / \|\mathfrak{E}^i\|_p\|_{\psi_p} + 1/\ln 2$ . From (13b), (13c), and (16), we deduce  $\mathfrak{R} \geq R$ , and thus

$$\mathbb{P}\left(\left(\frac{1}{N} \sum_{i=1}^N \left(\frac{\mathfrak{E}^i}{\|\mathfrak{E}^i\|_p}\right)^p\right)^{\frac{1}{p}} - 1 \geq t\right) \leq 2 \exp\left(-\frac{c'N}{\mathfrak{R}^2} \alpha_p(t)\right).$$

Now, it follows from (13a) that:

$$\mathfrak{M}_v \left( \frac{1}{N} \sum_{i=1}^N \left( \frac{\mathfrak{E}^i}{\|\mathfrak{E}^i\|_p} \right)^p \right)^{\frac{1}{p}} - \mathfrak{M}_v \geq \left( \frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p \right)^{\frac{1}{p}} - \mathfrak{M}_v.$$

Thus, we deduce that

$$\begin{aligned} & \mathbb{P}\left(\left(\frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p\right)^{\frac{1}{p}} \geq \mathfrak{M}_v + s\right) \\ & \leq 2 \exp\left(-\frac{c'N}{\mathfrak{R}^2} \alpha_p\left(\frac{s}{\mathfrak{M}_v}\right)\right). \end{aligned} \quad (33)$$

To establish (17), it suffices by Lemma 4.4 to show that

$$\mathbb{P}\left(2^{\frac{p-1}{p}} \mathfrak{M}_w + 2^{\frac{p-1}{p}} \left(\frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p\right)^{\frac{1}{p}} \leq \widehat{\varepsilon}_N(\beta_{\text{ns}})\right) \geq 1 - \beta_{\text{ns}}.$$

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

$$\begin{aligned} & \mathbb{P}\left(\left(\frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p\right)^{\frac{1}{p}} < \mathfrak{M}_v + \mathfrak{M}_v \alpha_p^{-1}\left(\frac{\mathfrak{R}^2}{c'N} \ln \frac{2}{\beta_{\text{ns}}}\right)\right) \\ & \geq 1 - \beta_{\text{ns}}. \end{aligned}$$

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

$$\mathbb{P}\left(\left(\frac{1}{N} \sum_{i=1}^N (\mathfrak{E}^i)^p\right)^{\frac{1}{p}} \geq \mathfrak{M}_v + \tau \mathfrak{M}_v\right) \leq \beta_{\text{ns}}$$

which follows by (33) with  $s = \tau \mathfrak{M}_v$ .

## B. Proofs From Section IV-D

Here we give the proof of Theorem 4.13. We use the next result, whose proof can be found in [10]. The result examines the Wasserstein distance between the distributions of two random variables when other random variables are added.

*Lemma A.3 (Wasserstein distance under convolution):* Given  $p \geq 1$  and distributions  $P_1, P_2, Q \in \mathcal{P}_p(\mathbb{R}^d)$ , it holds that  $W_p(P_1, P_2) \leq W_p(P_1 \star Q, P_2 \star Q)$ . Also, if it holds that  $(\int_{\mathbb{R}^d} \|x\|^p Q(dx))^{\frac{1}{p}} \leq q$ , then  $W_p(P_1, P_2 \star Q) \leq W_p(P_1, P_2) + q$ .

*Proof of Theorem 4.13:* The proof is carried out by induction on  $\ell \in [\ell_1 : \ell_2]$ . It suffices to establish that

$$\begin{aligned} W_p(P_{\xi_{\ell_1}}, \tilde{P}_{\xi_{\ell_1}}^N) & \leq \psi_{N, \ell_1} \implies \\ W_p(P_{\xi_{\ell'}}, \tilde{P}_{\xi_{\ell'}}^N) & \leq \psi_{N, \ell'} \quad \forall \ell' \in [\ell_1 : \ell]. \end{aligned} \quad (34)$$

Note that from Theorem 4.7,  $\mathbb{P}(P_{\xi_{\ell_1}} \in \mathcal{B}_{\psi_{N, \ell_1}}(\tilde{P}_{\xi_{\ell_1}}^N)) \geq 1 - \beta$ .

From (34), this also implies that  $\mathbb{P}(P_{\xi_{\ell'}} \in \mathcal{B}_{\psi_{N, \ell'}}(\tilde{P}_{\xi_{\ell'}}^N) \mid \ell' \in [\ell_1 : \ell]) \geq 1 - \beta$ , establishing validity of the result for  $\ell = \ell_2$ .

For  $\ell = \ell_1$ , the induction hypothesis (34) (IH) 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  $W_p(P_{\xi_{\ell}}, \tilde{P}_{\xi_{\ell}}^N) \leq \psi_{N, \ell}$  then also  $W_p(P_{\xi_{\ell+1}}, \tilde{P}_{\xi_{\ell+1}}^N) \leq \psi_{N, \ell+1}$  for both cases (i) and (ii). Consider first (i) and note that then the ambiguity set center at  $\ell + 1$  satisfies  $\tilde{P}_{\xi_{\ell+1}}^N = \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{\xi}_{\ell+1}^i} = \frac{1}{N} \sum_{i=1}^N \delta_{A_{\ell} \tilde{\xi}_{\ell}^i} = (A_{\ell})_{\#} \tilde{P}_{\xi_{\ell}}^N$ , where we have exploited that  $\tilde{\xi}_{\ell+1}^i \equiv \Phi_{k, \ell+1} \tilde{\xi}_{\ell}^i$  and the definition of  $\Phi_{k, \ell+1}$  (for  $k = \ell - 1, \ell$ ) to derive the second equality. Using also the fact that  $P_{\xi_{\ell+1}} = ((A_{\ell})_{\#} P_{\xi_{\ell}}) \star P_{w_{\ell}}$ , we get from the second result of Lemma A.3 that

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

Here, we also used the fact that  $W_p(f_{\#} P, f_{\#} Q) \leq L W_p(P, Q)$  for any globally Lipschitz function  $f : \mathbb{R}^d \rightarrow \mathbb{R}^r$  with Lipschitz constant  $L$  in the second to last inequality (see, e.g., [44, Proposition A.16]).

Next, we prove IH for (ii). From Lemma A.3 and the definition of the ambiguity set center and radius

$$\begin{aligned} W_p(P_{\xi_{\ell+1}}, \tilde{P}_{\xi_{\ell+1}}^N) \\ = W_p(((A_{\ell})_{\#} P_{\xi_{\ell}}) \star P_{w_{\ell}}, ((A_{\ell})_{\#} \tilde{P}_{\xi_{\ell}}^N) \star P_{w_{\ell}}) \end{aligned}$$

$$\begin{aligned} &\leq W_p \left( (A_\ell)_{\#} P_{\xi_\ell}, (A_\ell)_{\#} \tilde{P}_{\xi_\ell}^N \right) \leq \|A_\ell\| W_p \left( P_{\xi_\ell}, \tilde{P}_{\xi_\ell}^N \right) \\ &\leq \|A_\ell\| \psi_{N,\ell} = \|A_\ell\| \|A_{\ell-1}\| \cdots \|A_{\ell_1}\| \psi_{N,\ell_1} = \psi_{N,\ell+1} \end{aligned}$$

completing the proof.

### C. Proofs From Section V

*Proof of Proposition 5.1:* 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  $G^*$ . 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 [35, p. 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  $K^*$  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  $K_*$  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  $\Psi_s^*$  follow from the uniform bounds for all  $A_k$  and  $H_k$  and the derived bound  $K^*$  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., [42, Theorem 31]), namely  $\lim_{s \rightarrow \infty} \|F^s\| = 0$ . 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  $\Psi_{k+s,k} = F^s$ . Under Assumption 3.4(ii), let

$$\tilde{e}_{k+1} = F_k \tilde{e}_k \quad (35)$$

be the recursive noise-free version of the error equation (6). Then, from [35, p. 577], there exists a quadratic time-varying Lyapunov function  $V(k, \tilde{e}) := \tilde{e}^\top Q_k \tilde{e}$  with each  $Q_k$  positive definite,  $a_1, a_2 > 0$ ,  $a_3 \in (0, 1)$ , and  $m \in \mathbb{N}$  so that

$$a_1 \leq \lambda_{\min}(Q_k) \leq \lambda_{\max}(Q_k) \leq a_2 \quad (36a)$$

$$V(k+m, \tilde{e}_{k+m}) - V(k, \tilde{e}_k) \leq -a_3 V(k, \tilde{e}_k) \quad (36b)$$

for any  $k$  and any solution of (35) with state  $\tilde{e}_k$  at time  $k$ . Thus,  $\Psi_{k+m,k}^\top Q_{k+m} \Psi_{k+m,k} \preceq (1-a_3)Q_k$ , and hence, by induction  $\Psi_{k+\nu m,k}^\top Q_{k+\nu m} \Psi_{k+\nu m,k} \preceq (1-a_3)^\nu Q_k$ , since

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

Next, pick  $\tilde{e}$  with  $\|\tilde{e}\| = 1$  and  $\|\Psi_{k+\nu m,k} \tilde{e}\| = \|\Psi_{k+\nu m,k}\|$ . Taking into account that  $\tilde{e}^\top \Psi_{k+\nu m,k}^\top Q_{k+\nu m} \Psi_{k+\nu m,k} \tilde{e} \leq (1-a_3)^\nu \tilde{e}^\top Q_k \tilde{e}$ , we get  $\lambda_{\min}(Q_{k+\nu m}) \|\Psi_{k+\nu m,k} \tilde{e}\|^2 \leq (1-a_3)^\nu \lambda_{\max}(Q_k)$ . Using (36a)

$$\|\Psi_{k+\nu m,k}\| \leq (1-a_3)^{\frac{\nu}{2}} (a_2/a_1)^{\frac{1}{2}}. \quad (37)$$

Now, select  $\nu$  so that  $(1-a_3)^{\frac{\nu}{2}} (a_2/a_1)^{\frac{1}{2}} \leq 1/(2 \max_{s \in [1:m]} \Psi_s^*)$  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 (37), part (i), and the selection of  $\nu$  that

$$\begin{aligned} \|\Psi_{k+sm,k}\| &= \|\Psi_{k+s'_0+m',k+s'_0} \Psi_{k+s'_0,k}\| \\ &\leq \|\Psi_{k+s'_0+m',k+s'_0}\| \|\Psi_{k+\nu'm,k}\| \leq \Psi_{m'}^* \frac{1}{2 \max_{s \in [1:m]} \Psi_s^*} \leq \frac{1}{2} \end{aligned}$$

which establishes the result.

## REFERENCES

- [1] R. B. Ash, *Real Analysis and Probability*. Cambridge, MA, USA: Academic Press, 1972.
- [2] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2009.
- [3] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," *Math. Program.*, 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," *J. Appl. Probability*, vol. 56, no. 3, pp. 830–857, 2019.
- [5] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Math. Operations Res.*, 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 *Proc. Tut. Ope. Res. Emerg. Optim. Methods Model. Techn. Appl.*, 2021, pp. 227–254.
- [7] V. Bogachev, *Measure Theory*, vol. 1, Berlin, Germany: Springer, 2007.
- [8] D. Boskos, J. Cortés, and S. Martínez, "Data-driven ambiguity sets for linear systems under disturbances and noisy observations," in *Proc. Amer. Control Conf.*, 2020, pp. 4491–4496.
- [9] D. Boskos, J. Cortés, and S. Martínez, "Data-driven ambiguity sets with probabilistic guarantees for dynamic processes," *IEEE Trans. Autom. Control*, vol. 66, no. 7, pp. 2991–3006, Jul. 2021.
- [10] D. Boskos, J. Cortés, and S. Martínez, "High-confidence data-driven ambiguity sets for time-varying linear systems," 2021, *arXiv:2102.01142*.
- [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 J. Sci. Comput.*, vol. 43, no. 3, pp. A2102–A2129, 2021.
- [12] Z. Chen, D. Kuhn, and W. Wiesemann, "Data-driven chance constrained programs over Wasserstein balls," *Operations Res.*, 2022.
- [13] A. Cherukuri and J. Cortés, "Distributed coordination of DERs with storage for dynamic economic dispatch," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 835–842, Mar. 2018.
- [14] A. Cherukuri and J. Cortés, "Cooperative data-driven distributionally robust optimization," *IEEE Trans. Autom. Control*, vol. 65, no. 10, pp. 4400–4407, Oct. 2020.
- [15] J. Coulson, J. Lygeros, and F. Dörfler, "Data-enabled predictive control: In the shadows of the DeePC," in *Proc. Eur. Control Conf.*, 2019, pp. 307–312.
- [16] J. Dedecker and F. Merlevède, "Behavior of the empirical Wasserstein distance in  $R^d$  under moment conditions," *Electron. J. Probability*, vol. 24, no. 6, pp. 1–32, 2019.
- [17] S. Dereich, M. Scheutzow, and R. Schottstedt, "Constructive quantization: Approximation by empirical measures," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 49, no. 4, pp. 1183–1203, 2013.
- [18] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory Related Fields*, vol. 162, no. 3–4, pp. 707–738, 2015.
- [19] R. Gao, "Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality," *Operations Res.*, 2022.
- [20] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributionally robust optimization and variation regularization," *Operations Res.*, 2022.
- [21] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *Math. Operations Res.*, vol. 48, no. 2, pp. 603–655, May 2023.
- [22] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," version 2.1, Mar. 2014. [Online]. Available: <http://cvxr.com/cvx>.

- [23] Y. Guo, K. Baker, E. Dall-Anese, Z. Hu, and T. H. Summers, "Data-based distributionally robust stochastic optimal power flow—Part I: Methodologies," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1483–1492, Mar. 2019.
- [24] A. Hakobyan and I. Yang, "Wasserstein distributionally robust motion planning and control with safety constraints using conditional value-at-risk," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 490–496.
- [25] A. Hota, A. Cherukuri, and J. Lygeros, "Data-driven chance constrained optimization under Wasserstein ambiguity sets," in *Proc. Amer. Control Conf.*, 2019, pp. 1501–1506.
- [26] B. Kloeckner, "Empirical measures: Regularity is a counter-curse to dimensionality," *ESAIM: Probability Statist.*, vol. 24, pp. 408–434, 2020.
- [27] B. Li, J. Mathieu, and R. Jiang, "Distributionally robust chance constrained optimal power flow assuming log-concave distributions," in *Power Syst. Computation Conf.*, 2018, pp. 1–7.
- [28] D. Li, D. Fooladivanda, and S. Martínez, "Data-driven variable speed limit design for highways via distributionally robust optimization," in *Proc. Eur. Control Conf.*, 2019, pp. 1055–1061.
- [29] D. Li, D. Fooladivanda, and S. Martínez, "Online learning of parameterized uncertain dynamical environments with finite-sample guarantees," *IEEE Contr. Syst. Lett.*, vol. 5, no. 4, pp. 1351–1356, Oct. 2021.
- [30] D. Li and S. Martínez, "Online data assimilation in distributionally robust optimization," in *Proc. IEEE Int. Conf. Decis. Control*, 2018, pp. 1961–1966.
- [31] M. Li, "Li-ion dynamics and state of charge estimation," *Renewable Energy*, vol. 100, pp. 44–52, 2017.
- [32] 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.
- [33] J. Liu, Y. Chen, C. Duan, J. Lin, and J. Lyu, "Distributionally robust optimal reactive power dispatch with Wasserstein distance in active distribution network," *J. Modern Power Syst. Clean Energy*, vol. 8, no. 3, pp. 426–436, 2020.
- [34] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, no. 1–2, pp. 115–166, 2018.
- [35] J. B. Moor and B. D. O. Anderson, "Coping with singular transition matrices in estimation and control stability theory," *Int. J. Control*, vol. 31, no. 3, pp. 571–586, 1980.
- [36] B. P. G. V. Parys, D. Kuhn, P. J. Goulart, and M. Morari, "Distributionally robust control of constrained stochastic systems," *IEEE Trans. Autom. Control*, vol. 61, no. 2, pp. 430–442, Feb. 2016.
- [37] B. K. Poolla, A. R. Hota, S. Bolognani, D. S. Callaway, and A. Cherukuri, "Wasserstein distributionally robust look-ahead economic dispatch," *IEEE Trans. Power Syst.*, vol. 36, no. 3, May 2021.
- [38] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, "Regularization via mass transportation," *J. Mach. Learn. Res.*, vol. 20, no. 103, pp. 1–68, 2019.
- [39] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8474–8483.
- [40] A. Shapiro, "Distributionally robust stochastic programming," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2258–2275, 2017.
- [41] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, vol. 16, Philadelphia, PA, USA: SIAM, 2014.
- [42] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Berlin, Germany: Springer, 1998.
- [43] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47, Cambridge, MA, USA: Cambridge Univ. Press, 2018.
- [44] C. Villani, "Topics in optimal transportation, ser.," in *Graduate Studies in Mathematics*. Providence, RI, USA: Amer. Math. Soc., 2003.
- [45] J. Weed and F. Bach, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *Bernoulli*, vol. 25, no. 4 A, pp. 2620–2648, 2019.
- [46] J. Weed and Q. Berthet, "Estimation of smooth densities in Wasserstein distance," in *Proc. Conf. Learn. Theory*, 2019, pp. 3118–3119.
- [47] F. Xin, B.-M. Hodge, L. Fangxing, D. Ershun, and K. Chongqing, "Adjustable and distributionally robust chance-constrained economic dispatch considering wind power uncertainty," *J. Modern Power Syst. Clean Energy*, vol. 7, no. 3, pp. 658–664, May 2019.
- [48] I. Yang, "A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance," *IEEE Contr. Syst. Lett.*, vol. 1, no. 1, pp. 164–169, Jul. 2017.
- [49] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *IEEE Trans. Autom. Control*, vol. 66, no. 8, pp. 3863–3870, Aug. 2021.
- [50] S. Zeng, "Sample-based population observers," *Automatica*, vol. 101, pp. 108–118, 2021.



**Dimitris Boskos** (Member, IEEE) was born in Athens, Greece, in 1981. He received the diploma in mechanical engineering in 2005, and the M.Sc. and Ph.D. degrees in applied mathematics in 2008 and 2014, respectively, all from the National Technical University of Athens, Athens, Greece.

Between 2014 and 2018, he has been a Postdoctoral Researcher with the Department of Automatic Control, School of Electrical Engineering, Royal Institute of Technology, Stockholm, Sweden. Between 2018 and 2020, he has been a Postdoctoral Researcher with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA. Since 2020, he has been 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 multiagent systems, formal verification, and nonlinear observer design.



**Jorge Cortés** (Fellow, IEEE) 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. From 2004 to 2007, he was an Assistant Professor with the Department of

Applied Mathematics and Statistics, University of California, Santa Cruz, CA, USA. He is currently a Professor with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA. He has authored *Geometric Control and Numerical Aspects of Nonholonomic Systems* (Springer-Verlag, 2002) and coauthored (together with F. Bullo and S. Martínez) *Distributed Control of Robotic Networks* (Princeton University Press, 2009). His research interests include distributed control and optimization, network science, resource-aware control, nonsmooth analysis, reasoning, and decision making under uncertainty, network neuroscience, and multiagent coordination in robotic, power, and transportation networks.

Dr. Cortés has been a Distinguished Lecturer (2010–2014) with the IEEE Control Systems Society, an Elected Member (2018–2020) of its Board of Governors, and Director of Operations of its Executive Committee (2019–2022). He is a Fellow of the IEEE, SIAM, and IFAC.



**Sonia Martínez** (Fellow, IEEE) received the Ph.D. degree in engineering mathematics from the Universidad Carlos III de Madrid, Madrid, Spain, in 2002.

Following a year as a Visiting Assistant Professor of applied mathematics with the Technical University of Catalonia, Barcelona, Spain, she received a Postdoctoral Fulbright Fellowship and held appointments with the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign, Urbana, IL, USA,

during 2004, and at the Center for Control, Dynamical systems and Computation of the University of California, Santa Barbara, CA, USA, during 2005. She is a Professor of mechanical and aerospace engineering with the University of California, San Diego, CA, USA. Her research interests include the control of network systems, multiagent systems, nonlinear control theory, and robotics.

Dr. Martínez was a recipient of the Best Student Paper award at the 2002 IEEE Conference on Decision and Control for her work on the control of underactuated mechanical systems. She was also a recipient of the NSF CAREER Award in 2007. For the paper "Motion coordination with Distributed Information," coauthored 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*, the *Journal of Geometric Mechanics*, and has been a senior editor for the IEEE TRANSACTIONS ON NETWORKED SYSTEMS. She is currently a Senior Editor for *Automatica*, and the editor in Chief of the IEEE OPEN JOURNAL OF CONTROL SYSTEMS.