Capture, Propagate, and Control Distributional Uncertainty

Liviu Aolaritei, Nicolas Lanzetti, and Florian Dörfler

Abstract— We study stochastic dynamical systems in settings where only partial statistical information about the noise is available, e.g., in the form of a limited number of noise realizations. Such systems are particularly challenging to analyze and control, primarily due to an absence of a distributional uncertainty model which: (1) is expressive enough to capture practically relevant scenarios; (2) can be easily propagated through system maps; (3) is closed under propagation; and (4) allows for computationally tractable control actions. In this paper, we propose to model distributional uncertainty via Optimal Transport ambiguity sets and show that such modeling choice satisfies all of the above requirements. We then specialize our results to stochastic LTI systems, and start by showing that the distributional uncertainty can be efficiently captured, with high probability, within an Optimal Transport ambiguity set on the space of noise trajectories. Then, we show that such ambiguity sets propagate exactly through the system dynamics, giving rise to stochastic tubes that contain, with high probability, all trajectories of the stochastic system. Finally, we show that the control task is very interpretable, unveiling an interesting decomposition between the roles of the feedforward and the feedback control terms. Our results are actionable and successfully applied in stochastic reachability analysis and in trajectory planning under distributional uncertainty.

I. INTRODUCTION

In the era of data science, it is increasingly common to encounter stochastic (dynamical) systems for which only partial statistical information on the noise is available (e.g., samples). We are therefore confronted with so-called *distributional uncertainty*, whereby not only is the system affected by noise but also the underlying noise probability distribution is unknown and only partially observable.

In Operation Research and Machine Learning, Wasserstein ambiguity sets have emerged as a prominent model for distributional uncertainty. These are balls in the probability space defined in terms of the Wasserstein distance [1], a distance between probability distributions, and centered at a reference distribution $\widehat{\mathbb{P}}$. Examples of applications are distributionally robust optimization [2], [3], regression and classification [4]–[6], adversarial training [7], etc.

More recently, Wasserstein ambiguity sets, and more generally optimal transport, penetrated the control community, with application in uncertainty quantification in dynamical systems [8], [9], model predictive control [10]–[12], distribution steering [13], optimal control [14], multi-agent stochastic optimization [15], [16], linear quadratic differential games [17], probability/multi-agent control [18], and filtering [19]–[22],

to name a few. In this paper, we demonstrate that *Optimal Transport (OT) ambiguity sets*, which encompass Wasserstein ambiguity sets, are also easy to propagate. This makes them very natural to model distributional uncertainty in the context of dynamical systems, enjoying the following three desirable features:

- *Expressivity*. They are rich enough to capture relevant stochasticity of real-world systems; see Section II.
- *Propagation*. They are easily and exactly propagated through linear maps and they are *closed* under the propagation (i.e., the result of the propagation is itself an OT ambiguity set); see Section III.
- *Computation*. They allow for computationally tractable decision-making (e.g., computing worst-case risk over the uncertainty set); see Sections II and V.

More specifically, our contributions are threefold. First, we study the propagation of OT ambiguity sets through linear maps and show that the propagation of an OT ambiguity set is itself an OT ambiguity set or can be tightly upper bounded by an OT ambiguity set. Second, we instantiate our results in the context of stochastic linear time-invariant systems and unveil a decomposition between the roles of the feedforward and feedback control terms. Finally, we deploy our results for stochastic reachability analysis and trajectory planning under distributional uncertainty. Among others, we demonstrate that OT ambiguity sets enable the data-driven robust design of reachability sets and feedforward input trajectories. This way, decisions perform well for the true (but unknown) probability distribution governing the noise even if the decision-maker only disposes of a few samples.

This theory is successfully exploited in the concomitant paper [12] to formulate a Wasserstein Tube MPC capable of optimally trading between safety and performance.

A. Mathematical Preliminaries and Notation

Throughout the paper, $\mathcal{P}(\mathbb{R}^d)$ denotes the space of probability distributions over $\mathbb{R}^d$. Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$, we denote by $\mathbb{P} \otimes \mathbb{Q}$ their product distribution and by $\mathbb{P}^{\otimes t}$ the t-fold product distribution $\mathbb{P} \otimes \ldots \otimes \mathbb{P}$ with t terms. The delta probability distribution at $x \in \mathbb{R}^d$ is denoted by δ_x. We focus on two classes of transformations of probability distributions: *pushforward* via a linear transformation and the *convolution* with a delta distribution. We start with the pushforward:

Definition 1. Let $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and $A \in \mathbb{R}^{m \times d}$. The pushforward of $\mathbb{P}$ via the linear map $x \mapsto Ax$ is denoted by $A_{\#}\mathbb{P}$, and is defined by $(A_{\#}\mathbb{P})(\mathcal{B}) \coloneqq \mathbb{P}(A^{-1}(\mathcal{B}))$, for all Borel sets $\mathcal{B} \subset \mathbb{R}^m$.

Equivalently, Definition 1 says that if $x \sim \mathbb{P}$, then $A_\#\mathbb{P}$ is the probability distribution of the random variable $y = Ax$.

Example 1. Let $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\widehat{x}^{(i)}}$ be an empirical distribution. Then, $A_\#\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{A\widehat{x}^{(i)}}$ is empirical as well, supported on the propagated samples.

Moreover, given $x \sim \mathbb{P}$ on $\mathbb{R}^d$ and $y \in \mathbb{R}^d$, $x + y$ is distributed according to the convolution $\delta_y * \mathbb{P}$ defined below.

Definition 2. Let $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and $y \in \mathbb{R}^d$. Then, the convolution of $\mathbb{P}$ and δ_y is denoted by $\delta_y * \mathbb{P}$, and is defined by $(\delta_y * \mathbb{P})(\mathcal{A}) = \mathbb{P}(\mathcal{A} - y)$, for all Borel sets $\mathcal{A} \subset \mathbb{R}^d$.

Finally, we are interested in probabilistic constraints based on the *conditional value-at-risk (CVaR)*. Given $f : \mathbb{R}^d \to \mathbb{R}$ and a random variable $x \sim \mathbb{Q}$ on $\mathbb{R}^d$, the CVaR of $f(x)$ at probability level $1 - \gamma$ is defined as

$$\mathrm{CVaR}_{1-\gamma}^{\mathbb{Q}}(f(x)) = \inf_{\tau \in \mathbb{R}} \; \tau + \frac{1}{\gamma}\mathbb{E}_{\mathbb{Q}}\left[\max\{0, f(x) - \tau\}\right]. \quad (1)$$

II. Capture Distributional Uncertainty

We start by formalizing the notion of OT ambiguity sets and exposing their expressivity and their geometric, statistical, and computational features and guarantees.

A. Definition of Optimal Transport ambiguity sets

Consider a non-negative lower semi-continuous function $c : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ (henceforth, referred to as *transportation cost*) and two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$. Then, the *OT discrepancy* between $\mathbb{P}$ and $\mathbb{Q}$ is defined by

$$W^c(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x_1 - x_2)\mathrm{d}\gamma(x_1, x_2), \quad (2)$$

where $\Gamma(\mathbb{P}, \mathbb{Q})$ is the set of all probability distributions over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathbb{P}$ and $\mathbb{Q}$, often called *transport plans* or *couplings* [1]. The semantics are as follows: we seek the minimum cost to transport the probability distribution $\mathbb{P}$ onto the probability distribution $\mathbb{Q}$ when transporting a unit of mass from x_1 to x_2 costs $c(x_1 - x_2)$. Intuitively, $W^c(\mathbb{P}, \mathbb{Q})$ quantifies the discrepancy between $\mathbb{P}$ and $\mathbb{Q}$ and it naturally provides us with a definition of ambiguity in the space of probability distributions. In particular, the *OT ambiguity set* of radius ε centered at $\mathbb{P}$ is defined by

$$\mathbb{B}_{\varepsilon}^c(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^d) : W^c(\mathbb{P}, \mathbb{Q}) \leq \varepsilon\} \subset \mathcal{P}(\mathbb{R}^d). \quad (3)$$

In words, $\mathbb{B}_{\varepsilon}^c(\mathbb{P})$ includes all probability distributions onto which $\mathbb{P}$ can be transported with a budget of at most ε.

B. Properties of Optimal Transport ambiguity sets

OT ambiguity sets are attractive to capture distributional uncertainty for various reasons, which we detail next.

a) Expressivity: OT ambiguity sets are highly expressive: they contain both continuous and discrete distributions, distributions not concentrated on the support of $\mathbb{P}$, and even distributions whose mass asymptotically escapes to infinity:

Example 2. Let $c(x_1 - x_2) = |x_1 - x_2|^2$ on $\mathbb{R}$, $\varepsilon > 0$, and let $\mathbb{Q}$ be the Gaussian distribution with mean 0 and

variance ε. Then, $W^c(\delta_0, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[|x|^2\right] = \varepsilon$. Moreover, $W^c(\delta_0, \delta_{\sqrt{\varepsilon}}) = \varepsilon$, and $W^c(\delta_0, \varepsilon\frac{1}{n^2}\delta_n + (1 - \varepsilon\frac{1}{n^2})\delta_0) = \varepsilon$.

These properties cease to hold if the discrepancy between probability distributions is measured via the Kullback-Leibler (KL) divergence or Total Variation (TV) distance [23].

b) Geometric properties: OT ambiguity sets encapsulate the geometry that the transportation cost c induces on $\mathbb{R}^d$; e.g., if x_0, x_1 satisfy $c(x_0 - x_1) \leq \varepsilon$, then $\delta_{x_1} \in \mathbb{B}_{\varepsilon}^c(\delta_{x_0})$. Moreover, (2) and (3) readily show that ambiguity sets are well-behaved under monotone changes in c and ε:

Lemma 1. *Let $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$, c, c_1, c_2 be transportation costs over $\mathbb{R}^d$, and $\varepsilon, \varepsilon_1, \varepsilon_2 > 0$. Then,*
(i) *if $\varepsilon_1 \leq \varepsilon_2$ then $\mathbb{B}_{\varepsilon_1}^c(\mathbb{P}) \subseteq \mathbb{B}_{\varepsilon_2}^c(\mathbb{P})$;*
(ii) *if $c_1 \leq c_2$, then $\mathbb{B}_{\varepsilon}^{c_2}(\mathbb{P}) \subseteq \mathbb{B}_{\varepsilon}^{c_1}(\mathbb{P})$.*

In words, an increase in the transportation cost shrinks the ambiguity set, whereas an increase of the radius enlarges it. These simple observations arm practitioners with actionable knobs to control the level of distributional uncertainty.

c) Statistical properties: In most applications, probability distributions are not directly observable and must be estimated from data. Specifically, suppose one has access to n i.i.d. samples $\{\widehat{x}^{(i)}\}_{i=1}^n$ from $\mathbb{P}$, and constructs the *empirical probability distribution* $\widehat{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n}\delta_{\widehat{x}^{(i)}}$. A straightforward generalization of [24, Theorem 2] stipulates that if $c(x_1 - x_2) \leq \|x_1 - x_2\|^p$ for some $p \geq 1$ and the true distribution $\mathbb{P}$ is light-tailed, then $\mathbb{P} \in \mathbb{B}_{\varepsilon}^c(\widehat{\mathbb{P}})$ with high probability, provided that the radius ε is carefully chosen.

d) Computational tractability: For any $\mathbb{P}$-integrable upper semi-continuous function $\ell : \mathbb{R}^d \to \mathbb{R}$, the evaluation of the *worst-case risk* over OT ambiguity sets admits a powerful dual reformulation [25],

$$\sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon}^c(\mathbb{P})} \mathbb{E}_{\mathbb{Q}}\left[\ell(x)\right] = \inf_{\lambda \geq 0} \lambda\varepsilon + \mathbb{E}_{\mathbb{P}}\left[\sup_{\xi \in \mathbb{R}^d} \ell(\xi) - \lambda c(\xi - x)\right],$$

which collapses to computationally tractable finite-dimensional optimization problems for many cases of practical interest [3].

III. Propagate Distributional Uncertainty

In this section, we study how OT ambiguity sets propagate via linear transformations. Before doing so, we show that naive approaches (in particular, propagation of the center only, or propagation based on Lipschitz bounds) fail to effectively capture the propagation of distributional uncertainty.

A. Naive Approaches and their Shortcomings

Given the OT ambiguity set $\mathbb{B}_{\varepsilon}^c(\mathbb{P})$, one might be tempted to approximate the result of the propagation $A_\#\mathbb{B}_{\varepsilon}^c(\mathbb{P})$ by $\mathbb{B}_{\varepsilon}^c(A_\#\mathbb{P})$. This approach suffers from fundamental limitations already in very simple settings, easily resulting in crude overestimation or in catastrophic underestimation of the ambiguity set, as shown in the next example.

Example 3. Let $c(x_1 - x_2) = |x_1 - x_2|$ on $\mathbb{R}$, and $\varepsilon > 0$.
- Let $A = 0$. Then, $A_\#\mathbb{B}_{\varepsilon}^c(\mathbb{P})$ only contains δ_0, whereas $\mathbb{B}_{\varepsilon}^c(A_\#\mathbb{P}) = \mathbb{B}_{\varepsilon}^c(\delta_0)$ contains all distributions whose first

moment is at most ε. Therefore, $\mathbb{B}_\varepsilon^c(A_\#\mathbb{P})$ *overestimates* the true distributional uncertainty.

- Let $A = 2x$, and $\mathbb{P} = \delta_0$. Then, $\mathbb{B}_\varepsilon^c(A_\#\mathbb{P}) = \mathbb{B}_\varepsilon^c(\delta_0)$. In Theorem 2 we show that $A_\#\mathbb{B}_\varepsilon^c(\mathbb{P}) = \mathbb{B}_{2\varepsilon}^c(\delta_0)$. Thus, $\mathbb{B}_\varepsilon^c(A_\#\mathbb{P})$ *underestimates* the true uncertainty.

Moreover, one might be tempted to bound the propagated distributional uncertainty with the Lipschitz constant L of A, i.e., to "upper bound" $A_\#\mathbb{B}_\varepsilon^c(\mathbb{P})$ with $\mathbb{B}_{L\varepsilon}^c(\mathbb{P})$. However, this approach suffers from two major limitations. First, the transportation cost $c(x_1 - x_2)$ might not be $\|x_1 - x_2\|$, which makes the Lipschitz bound not directly applicable. Indeed, already in Example 3, an increase by a factor of 2 in the radius does not alleviate the underestimation of the true ambiguity set: one needs to use $L^2 = 2^2$ to account for the transportation cost not being $\|x_1 - x_2\|$. Second, even if the transportation cost is $\|x_1 - x_2\|$, Lipschitz bounds might be overly conservative, as shown next.

Example 4. Let $c(x_1 - x_2) = \|x_1 - x_2\|$ on $\mathbb{R}^d$, $\varepsilon > 0$, and A be a diagonal matrix with diagonal entries $\{0, 0, \ldots, n\}$ (thus, $L = n$). Then, $A_\#\mathbb{B}_\varepsilon^c(\mathbb{P})$ eliminates all the distributional uncertainty in the first $d-1$ dimensions. However, $\mathbb{B}_{n\varepsilon}^c(\delta_0)$ contains, among others, all distributions of the form $\delta_0 \otimes \mathbb{Q}$ with $\delta_0 \in \mathcal{P}(\mathbb{R})$, and $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^{d-1})$ satisfying $\mathbb{E}_\mathbb{Q}[\|x\|] \leq n\varepsilon$.

These shortcomings of naive uncertainty propagation prompt us to study the propagation of OT ambiguity sets.

B. Propagation via Linear Transformations

We now investigate how OT ambiguity sets are propagated through linear transformations defined by the matrix $A \in \mathbb{R}^{m \times d}$. To do so, we require the following mild structural assumption on the transportation cost.

Assumption 1. The transportation cost c is orthomonotone: $c(x_1 + x_2) \geq c(x_1)$ for all $x_1, x_2 \in \mathbb{R}^d$ satisfying $x_1^\top x_2 = 0$.

Remark. All transportation costs of the form $c(x_1 - x_2) = \psi(\|x_1 - x_2\|)$, where $\|\cdot\|$ is the Euclidean norm and $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is monotone and lower semi-continuous, are orthomonotone: for $x_1, x_2 \in \mathbb{R}^d$ such that $x_1^\top x_2 = 0$

$$\psi(\|x_1 + x_2\|) = \psi\left(\sqrt{\|x_1\|^2 + \|x_2\|^2}\right) \geq \psi(\|x_1\|).$$

We can now state our main result on the propagation via linear transformations.

Theorem 2 (Linear transformations). *Let* $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ $c : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ *satisfy Assumption 1. Consider a linear transformation defined by a matrix* $A \in \mathbb{R}^{m \times d}$. *Then,*

$$A_\#\mathbb{B}_\varepsilon^c(\mathbb{P}) \subseteq \mathbb{B}_\varepsilon^{c \circ A^\dagger}(A_\#\mathbb{P}), \tag{4}$$

with $A^\dagger$ *the Moore–Penrose pseudoinverse of A. Moreover, if the matrix A is full row-rank, then*

$$A_\#\mathbb{B}_\varepsilon^c(\mathbb{P}) = \mathbb{B}_\varepsilon^{c \circ A^\dagger}(A_\#\mathbb{P}). \tag{5}$$

with $A^\dagger = A^\top(AA^\top)^{-1}$.

The proof of this and all subsequent results is deferred to the online extended version [26]. In words, Theorem 2

asserts that the result of the propagation $A_\#\mathbb{B}_\varepsilon^c(\mathbb{P})$ is itself an OT ambiguity set, with the same radius ε, propagated center $A_\#\mathbb{P}$, and an A-induced transportation cost $c \circ A^\dagger$.

Remark. Theorem 2 continues to hold if the OT ambiguity set $\mathbb{B}_\varepsilon^c(\mathbb{P})$ is defined over a subset $\mathcal{X} \subset \mathbb{R}^d$, with $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, $c : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, and $A : \mathcal{X} \to \mathcal{Y} := A\mathcal{X}$. In that case, the propagated ambiguity set $\mathbb{B}_\varepsilon^{c \circ A^\dagger}(A_\#\mathbb{P})$ is restricted to all distributions supported on $A\mathcal{X}$.

The following example shows that the equality (5) does generally *not* hold for non-surjective linear maps.

Example 5. Consider $A := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ with pseudoinverse $A^\dagger = A$, the quadratic transportation cost $c(x_1 - x_2) = \|x_1 - x_2\|^2$, the probability distribution $\mathbb{P} = \delta_{(0,0)} \in \mathcal{P}(\mathbb{R}^2)$, and an arbitrary radius $\varepsilon > 0$. Let $\mathbb{Q} = \delta_{(0,1)} \in \mathcal{P}(\mathbb{R}^2)$. Since $(0,1) \notin \text{Range}(A)$, $\mathbb{Q}$ does not belong to $A_\#\mathbb{B}_\varepsilon^c(\mathbb{P})$. However, $W^{c \circ A^\dagger}(\mathbb{Q}, \mathbb{P}) = \left\|A^\dagger \begin{bmatrix} 0 \\ 0 \end{bmatrix} - A^\dagger \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\|^2 = 0$. Thus, $\mathbb{Q} \in \mathbb{B}_\varepsilon^{c \circ A^\dagger}(A_\#\mathbb{P})$, and $A_\#\mathbb{B}_\varepsilon^c(\mathbb{P}) \subsetneq \mathbb{B}_\varepsilon^{c \circ A^\dagger}(A_\#\mathbb{P})$.

IV. STOCHASTIC LINEAR CONTROL SYSTEMS

In this section, we focus on the stochastic linear time-invariant control system

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + Dw_t \\ u_t &= Kx_t + v_t \end{aligned} \tag{6}$$

and show that the theory of Section III can be efficiently exploited to capture, propagate, and even control distributional uncertainty. We assume that the system matrices $A \in \mathbb{R}^{d \times d}$, $B \in \mathbb{R}^{d \times m}$, $D \in \mathbb{R}^{d \times r}$ are known, the initial condition $x_0 \in \mathbb{R}^d$ is known and deterministic, and the noise sequence $\{w_t\}_{t \in \mathbb{N}} \subset \mathbb{R}^r$ is i.i.d. according to an *unknown* light-tailed distribution $\mathbb{P}$ that belongs to the OT ambiguity set $\mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}})$, with reference distribution $\widehat{\mathbb{P}}$ and (translation-invariant and orthomonotone) transportation cost $c(\cdot) = \|\cdot\|_2^2$.

Remark. Modelling the noise using the OT ambiguity set $\mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}})$ allows us to capture many important scenarios. First, we generalize the works which assume the noise to belong to a specific class of distributions (e.g., Gaussian). Second, it allows us to robustly capture the system uncertainty when only partial statistical information about the noise is available (e.g., samples, moments, etc.). In many such cases, $\mathbb{P} \in \mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}})$ can be guaranteed with high probability (see the statistical properties in Section II-B).

A. Capture Distributional Uncertainty in Linear Systems

We start by defining, for any $t \in \mathbb{N}$, the vectors $\mathbf{v}_{[t-1]} = \begin{bmatrix} v_{t-1}^\top & \cdots & v_0^\top \end{bmatrix}^\top$ and $\mathbf{w}_{[t-1]} = \begin{bmatrix} w_{t-1}^\top & \cdots & w_0^\top \end{bmatrix}^\top$, and rewriting the system dynamics (6) in the form

$$\begin{aligned} x_t &= (A + BK)^t x_0 + \mathbf{B}_{t-1}\mathbf{v}_{[t-1]} + \mathbf{D}_{t-1}\mathbf{w}_{[t-1]}, \\ \mathbf{B}_{t-1} &= \begin{bmatrix} B & (A+BK)B & \ldots & (A+BK)^{t-1}B \end{bmatrix}, \\ \mathbf{D}_{t-1} &= \begin{bmatrix} D & (A+BK)D & \ldots & (A+BK)^{t-1}D \end{bmatrix}. \end{aligned} \tag{7}$$

Eq. (7) unveils that the distributional uncertainty of the state x_t can be characterized through the pushforward via the matrix $\mathbf{D}_{t-1}$ of the OT ambiguity set associated to $\mathbf{w}_{[t-1]}$.

The following lemma explains how to construct the OT ambiguity set of the noise trajectory $\mathbf{w}_{[t-1]}$, starting from the ambiguity set of w_0, i.e., $\mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}})$. Since $\mathbf{w}_{[t-1]}$ is composed of t i.i.d. random variables distributed according to $\mathbb{P}$, its distribution is the t-fold product distribution $\mathbb{P}^{\otimes t}$.

Lemma 3. *Let* $\mathbb{P} \in \mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}})$ *with probability* $1-\delta$, *for some* $\delta \geq 0$. *Then,*

$$\mathbb{P}^{\otimes t} \in \mathbb{B}_{t\varepsilon}^{\|\cdot\|_2^2}(\widehat{\mathbb{P}}^{\otimes t}), \quad \text{with probability } 1-\delta.$$

The true power of Lemma 3 is revealed in data-driven scenarios in the settings of control tasks over a specified prediction horizon $t \in \mathbb{N}$ (e.g., model predictive control). In such cases, we have access to only n i.i.d. noise samples $\{\widehat{w}^{(i)}\}_{i=1}^n$ from $\mathbb{P}$, and we construct the empirical distribution $\widehat{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n \delta_{\widehat{w}^{(i)}}$. By [24, Theorem 2], $\mathbb{P} \in \mathbb{B}_\varepsilon^c(\widehat{\mathbb{P}})$ with high probability, provided that the radius ε is in the order of $n^{-1/\max\{2,r\}}$. Then, Lemma 3 guarantees that the distribution $\mathbb{P}^{\otimes t}$ of the noise trajectory $\mathbf{w}_{[t-1]}$ belongs, with high-probability, to $\mathbb{B}_{\varepsilon_1}^{\|\cdot\|_2^2}(\widehat{\mathbb{P}}^{\otimes t})$, with radius ε_1 in the order of $tn^{-1/\max\{2,r\}}$. This is preferable over the alternative strategy of working directly with n noise trajectories $\{(\widehat{w}_{t-1}^{(i)}, \ldots, \widehat{w}_0^{(i)})\}_{i=1}^n$, and constructing an OT ambiguity set $\mathbb{B}_{\varepsilon_1}^{\|\cdot\|_2^2}(\widehat{\mathbb{Q}})$ around the empirical distribution $\widehat{\mathbb{Q}} = \frac{1}{n}\sum_{i=1}^n \delta_{(\widehat{w}_{t-1}^{(i)}, \ldots, \widehat{w}_0^{(i)})}$ based solely on [24, Theorem 2]. In that case, the radius should be in the order of $n^{-1/\max\{2,tr\}}$. In practical cases, where the dimension of the noise r is low, the *linear* dependence on the horizon in $tn^{-1/tr}$ ensures that the ambiguity radius shrinks much faster with the number of samples n, as opposed to the exponential dependence in $n^{-1/\max\{2,tr\}}$.

B. Propagate and Control Distributional Uncertainty

We can now study the propagation of the uncertainty from the noise $\mathbf{w}_{[t-1]}$ to the state x_t. Importantly, the resulting OT ambiguity set capturing the distributional uncertainty of x_t unveils the role of the two components in the control input $u_t = Kx_t + v_t$: the feedforward term v_t controls the *center*, while the feedback gain matrix K controls the *shape and size* of this OT ambiguity set. This is explained in the following proposition and the subsequent discussion.

Proposition 4. *Consider the linear control system* (6), *with i.i.d. noise* $\{w_t\}_{t\in\mathbb{N}}$. *Moreover, let* $\mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}})$ *capture the distributional uncertainty of* w_t, $\forall t \in \mathbb{N}$. *Then, the distributional uncertainty of* x_t *is captured by*

$$\mathbb{B}_{t\varepsilon}^{\|\cdot\|_2^2 \circ \mathbf{D}_{t-1}^\dagger}\left(\delta_{(A+BK)^t x_0 + \mathbf{B}_{t-1}\mathbf{v}_{[t-1]}} * (\mathbf{D}_{t-1\#}\widehat{\mathbb{P}}^{\otimes t})\right). \quad (8)$$

Recall from Theorem 2 that the propagation of the distributional uncertainty via $\mathbf{D}_{t-1}$ is *exact* whenever the matrix $\mathbf{D}_{t-1}$ is full row-rank. This trivially holds when D is the identity matrix, and more generally can be guaranteed by an appropriate choice of the feedback gain matrix K. In the following, we inspect the three components of the OT ambiguity set (8) to shed light on the roles of the feedforward control trajectory $\mathbf{v}_{[t-1]}$ and of the feedback gain matrix K.

(1) **Ambiguity radius** $t\varepsilon$. This quantity grows linearly in the horizon t, and can be shrunk only by shrinking ε. This, in turn, requires having access to a higher number of noise samples $\{\widehat{w}^{(i)}\}_{i=1}^n$ (recall that ε decreases as $n^{-1/\max\{2,r\}}$).

(2) **Center** $\delta_{(A+BK)^t x_0 + \mathbf{B}_{t-1}\mathbf{v}_{[t-1]}} * (\mathbf{D}_{t-1\#}\widehat{\mathbb{P}}^{\otimes t})$. The center distribution is influenced by both K and $\mathbf{v}_{[t-1]}$. In particular, K determines the shape of the center distribution (through the pushforward via the matrix $\mathbf{D}_{t-1}$); $\mathbf{v}_{[t-1]}$, instead, translates the support of the center distribution in $\mathbb{R}^d$. This observation becomes more clear in the data-driven scenario: if $\widehat{\mathbb{P}}$ is the empirical distribution supported on the points $\{\widehat{w}^{(i)}\}_{i=1}^n$, then the distribution $\widehat{\mathbb{P}}^{\otimes t}$ is the empirical distribution supported on the noise trajectories $\widehat{\mathbf{w}}_{[t-1]}^{(\mathbf{i})} := \left[(w^{(i_1)})^\top \cdots (w^{(i_t)})^\top\right]^\top$, for $\mathbf{i} = [i_1, \ldots, i_t]$, $i_j \in [n]$, $j \in [t]$. In that case, the center distribution becomes the empirical distribution supported on the points

$$(A+BK)^t x_0 + \mathbf{B}_{t-1}\mathbf{v}_{[t-1]} + \mathbf{D}_{t-1}\widehat{\mathbf{w}}_{[t-1]}^{(\mathbf{i})}.$$

Thus, $\mathbf{D}_{t-1}$ (and consequently, K) maps the noise trajectory to the point $(A + BK)^t x_0 + \mathbf{D}_{t-1}\widehat{\mathbf{w}}_{[t-1]}^{(\mathbf{i})} \in \mathbb{R}^d$, and $\mathbf{B}_{t-1}\mathbf{v}_{[t-1]}$ controls this point in $\mathbb{R}^d$.

(3) **Transportation cost** $\|\cdot\|_2^2 \circ \mathbf{D}_{t-1}^\dagger$. This function influences both the shape and size of the OT ambiguity set. For ease of exposition, we assume that $\mathbf{D}_{t-1}$ is full row-rank. Then, $\mathbf{D}_{t-1}^\dagger = \mathbf{D}_{t-1}^\top(\mathbf{D}_{t-1}\mathbf{D}_{t-1}^\top)^{-1}$. Moreover, if $U\Sigma V^\top$ is the SVD of $\mathbf{D}_{t-1}$, then $V\Sigma^\dagger U^\top$ is the SVD of $\mathbf{D}_{t-1}^\dagger$. In particular, the singular values of $\mathbf{D}_{t-1}^\dagger$ are obtained by taking the inverse of the singular values of $\mathbf{D}_{t-1}$. Consequently, if $\{\sigma_i\}_{i=1}^d$ are the singular values of $\mathbf{D}_{t-1}$ and $\{u_i\}_{i=1}^d$ are the orthonormal columns of U, the transportation cost becomes

$$\|\mathbf{D}_{t-1}^\dagger(x_1 - x_2)\|_2^2 = \sum_{i=1}^d \frac{1}{\sigma_i^2}\left|u_i^\top(x_1 - x_2)\right|^2.$$

In words, the cost of moving probability mass from the center distribution in the direction u_i costs $\|x_1 - x_2\|^2/\sigma_i^2$ (indeed, $u_i^\top(x_1 - x_2)$ is the orthogonal projection of $x_1 - x_2$ onto u_1). The feedback gain matrix K controls the amount of mass moved in this direction through the singular value σ_i of the matrix $\mathbf{D}_{t-1}$. Specifically, the higher the value of σ_i, the more probability mass is moved in the direction u_i. Similarly, the lower the value of σ_i, the less probability mass is moved in the direction u_i. This way, we can precisely control the displacement of probability mass from the center distribution and so the shape and size of the OT ambiguity set (8). Alternatively, if only the size of (8) is of interest, the maximum singular value of $\mathbf{D}_{t-1}$ yields the upper bound

$$\mathbb{B}_{t\varepsilon}^{\|\cdot\|_2^2 \circ \mathbf{D}_{t-1}^\dagger}(\widetilde{\mathbb{P}}) \subseteq \mathbb{B}_{t\varepsilon}^{\sigma_{\min}(\mathbf{D}_{t-1}^\dagger)^2\|\cdot\|_2^2}(\widetilde{\mathbb{P}}) = \mathbb{B}_{t\varepsilon\sigma_{\max}(\mathbf{D}_{t-1})^2}^{\|\cdot\|_2^2}(\widetilde{\mathbb{P}})$$

with $\widetilde{\mathbb{P}} := \delta_{(A+BK)^t x_0 + \mathbf{B}_{t-1}\mathbf{v}_{[t-1]}} * (\mathbf{D}_{t-1\#}\widehat{\mathbb{P}}^{\otimes t})$.

Summarizing, the careful inspection of the OT ambiguity set (8) brings to light the separation of the control tasks carried out by the two components of u_t. On the one hand, the feedforward term v_t controls the position in $\mathbb{R}^d$ of the support of the center distribution (and, with it, the position of the OT ambiguity set). On the other hand, the feedback

gain matrix K controls the shape and size of the center distribution, as well as the shape and size of the displacement of probability mass from the center distribution (through the transportation cost).

V. APPLICATIONS

In this section, we apply our theoretic results to stochastic reachability analysis and trajectory planning. In our setting, the decision-maker has only access to finitely many samples (i.e., noise trajectories) and seeks a deterministic set capturing the state of the system (in reachability analysis) or the cheapest control input to reach a given target (in trajectory planning) which perform well under the true distribution.

A. Preliminaries

Before diving into details, we fix the notation and present a preliminary result in distributionally robust optimization. Henceforth, we assume that the matrix $\mathbf{D}_{t-1}$ is full row-rank and that we have access to n noise sample trajectories

$$\widehat{\mathbf{w}}_{[t-1]}^{(i)} := \left[(\widehat{w}_{t-1}^{(i)})^\top \quad \dots \quad (\widehat{w}_0^{(i)})^\top \right]^\top,$$

for $i \in \{1, \dots, n\}$, which yield the empirical distribution $\widehat{\mathbb{P}}_t = \frac{1}{n} \sum_{i=1}^n \delta_{\widehat{\mathbf{w}}_{[t-1]}^{(i)}}$. Moreover, we assume that the distributional uncertainty of $\mathbf{w}_{[t-1]}$ is captured by the OT ambiguity set $\mathbb{B}_\varepsilon^{\|\cdot\|_2^2}(\widehat{\mathbb{P}}_t)$. This encompasses (and relaxes) the setting of the previous section, where statistical considerations forced ε and $\widehat{\mathbb{P}}_t$ to be of the form $t\tilde{\varepsilon}$ and $\tilde{\mathbb{P}}^{\otimes t}$, for some $\tilde{\varepsilon} > 0$ and some empirical distribution $\tilde{\mathbb{P}}$ on $\mathbb{R}^r$, respectively. Then, similarly to Proposition 4, the distributional uncertainty of x_t is captured by

$$\mathbb{S}_t(\mathbf{v}_{[t-1]}) := \mathbb{B}_\varepsilon^{\|\cdot\|_2^2 \circ \mathbf{D}_{t-1}^\dagger} \left(\delta_{(A+BK)^t x_0 + \mathbf{B}_{t-1} \mathbf{v}_{[t-1]}} * (\mathbf{D}_{t-1} \# \widehat{\mathbb{P}}_t) \right).$$

The center of the ambiguity set (8) is supported on the n controlled state samples

$$\widehat{x}_t^{(i)} = (A+BK)^t x_0 + \mathbf{B}_{t-1} \mathbf{v}_{[t-1]} + \mathbf{D}_{t-1} \widehat{\mathbf{w}}_{[t-1]}^{(i)}.$$

Consider now the polyhedral constraint set

$$\mathcal{X} := \left\{ x \in \mathbb{R}^d : \max_{j \in [J]} a_j^\top x + b_j \leq 0, \, J \in \mathbb{N} \right\}, \quad (9)$$

where $a_j \in \mathbb{R}^d$, $b_j \in \mathbb{R}$, and $[J] := \{1, \dots, J\}$. For some $\gamma \in (0, 1)$, we impose the distributionally robust CVaR constraint (with CVaR defined in (1))

$$\sup_{\mathbb{Q} \in \mathbb{S}_t(\mathbf{v}_{[t-1]})} \text{CVaR}_{1-\gamma}^{\mathbb{Q}} \left(\max_{j \in [J]} a_j^\top x_t + b_j \right) \leq 0. \quad (10)$$

By [3, Proposition 2.12 and Theorem 2.16], (10) can be reformulated as a set of convex deterministic constraints:

Proposition 5. *Constraint* (10) *is equivalent to the following set of convex constraints, denoted by* $\Gamma_t(\mathbf{v}_{[t-1]}, A, b)$:

$$\forall i \in [N], \forall j \in [J+1]:$$

$$\begin{cases} \tau \in \mathbb{R}, \lambda \in \mathbb{R}_+, s_i \in \mathbb{R} \\ \lambda \varepsilon N + \sum_{i=1}^N s_i \leq 0 \\ \alpha_j^\top \widehat{x}_t^{(i)} + \beta_j(\tau) + \frac{1}{4\lambda} \alpha_j^\top \left((\mathbf{D}_{t-1}^\dagger)^\top \mathbf{D}_{t-1}^\dagger \right)^{-1} \alpha_j \leq s_i \end{cases}$$
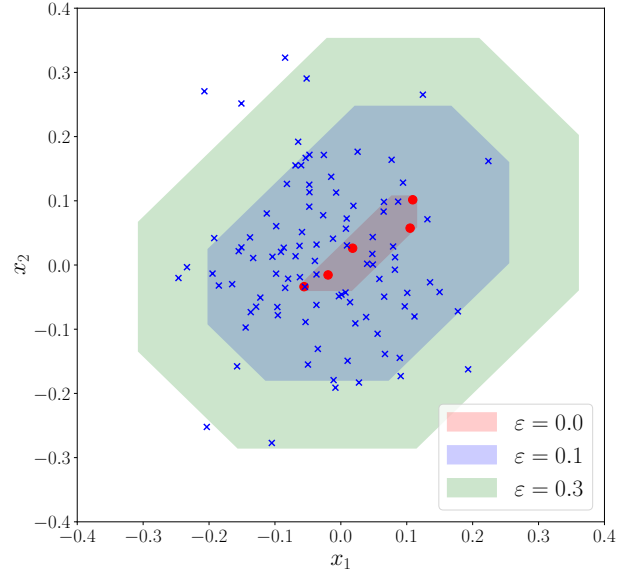

Fig. 1: Set containing the state of the system at time t for various radii ε. In red, the training samples (i.e., the samples available to the decision maker); in blue, the testing samples (i.e., samples used to test the decision of the decision-maker). The radius ε arbitrates performance and robustness: smaller radii lead to smaller sets, which however perform poorly on unseen data points; larger radii result in bigger sets, which well capture unseen samples.

with $\alpha_j := a_j/\gamma$ and $\beta_j(\tau) := (b_j + \gamma\tau - \tau)/\gamma$, for $j \in [J]$, as well as $\alpha_{J+1} := 0$ and $\beta_{J+1}(\tau) := \tau$.

B. Distributionally Robust Reachability Analysis

In the setting of stochastic reachability analysis, for a given feedforward input $\mathbf{v}_{[t-1]}$, we look for the *smallest* (deterministic) set which contains, with high confidence, the state x_t of a stochastic linear time-invariant system at some future time t. Specifically, for predefined $a_j \in \mathbb{R}^n$, $\forall j \in [J]$, we parametrize the set as a polyhedron (9) and seek to solve

$$\begin{aligned} \max \quad & \sum_{j \in [J]} b_j \\ \text{s.t.} \quad & \sup_{\mathbb{Q} \in \mathbb{S}_t(\mathbf{v}_{[t-1]})} \text{CVaR}_{1-\gamma}^{\mathbb{Q}} \left(\max_{j \in [J]} a_j^\top x_t + b_j \right) \leq 0. \end{aligned}$$

Proposition 5 directly gives the convex reformulation

$$\begin{aligned} \max \quad & \sum_{j \in [J]} b_j \\ \text{s.t.} \quad & b \in \Gamma_t(\mathbf{v}_{[t-1]}, A, b). \end{aligned}$$

We evaluate our methodology on the two-dimensional linear system $A = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$, $B = I$, and $D = 0.1I$, with K being the LQR controller (designed with $Q = R = I$) and $t = 10$. We suppose that the decision-maker has access to 5 noise sample trajectories (the red points in Fig. 1) and that $v_t = 0$ (so, $\mathbf{v}_{[t-1]} = 0$). We select $\gamma = 0.05$ and choose $J = 8$ hyperplanes with $a_j = \begin{bmatrix} i & j \end{bmatrix}^\top$ with $i, j \in \{0, \pm 1\}$ (without the trivial case $a_j = 0$). We repeat our experiments for three values of ε. Our results are in Fig. 1. For low ε, the optimal set tightly includes the state resulting from the 5 samples trajectories but performs very poorly on unseen samples (blue crossed Fig. 1). A larger ε, instead, leads to an increasingly larger set, which performs well on test samples, so that ε arbitrates between performance and robustness.
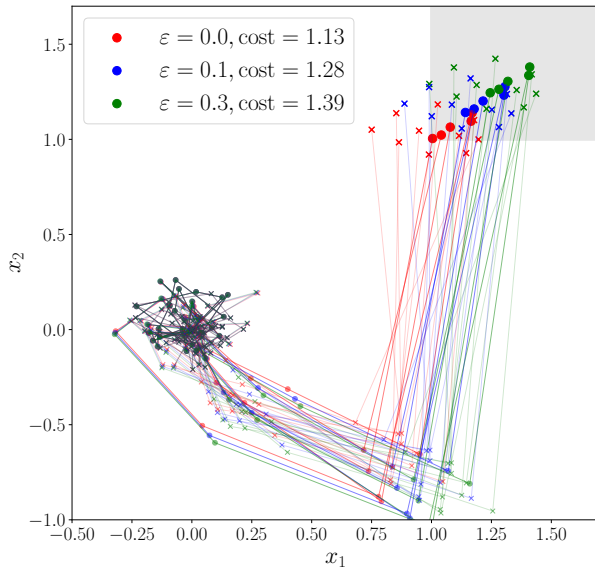
Fig. 2: Distributionally robust steering of a system from the origin to a target set (in grey) for various radii ε. Filled circles (with solid lines) are the training samples and crosses (with dotted lines) are testing samples. Small radii lead to trajectories approaching the boundary of the target, which however might not reach the target under different noise realizations (in red); larger radii, instead, push the trajectory to the interior of the target (in green).

C. Distributionally Robust Trajectory Planning

Our second example concerns (distributionally robust) trajectory planning. Given a deterministic initial condition, we look for the cheapest feedforward input v_t steering the system to a given target, expressed in the form of a polyhedral set (cf. (9)) Accordingly, the trajectory planning problem reads

$$
\begin{aligned}
\min \quad & \sum_{t=0}^{t-1} \|v_t\|_2^2 \\
\text{s.t.} \quad & \sup_{\mathbb{Q}\in\mathbb{S}_t(\mathbf{v}_{[t-1]})} \text{CVaR}_{1-\gamma}^{\mathbb{Q}}\left(\max_{j\in[J]} a_j^\top x_t + b_j\right) \le 0.
\end{aligned}
$$

Proposition 5 readily gives the convex reformulation

$$
\begin{aligned}
\min \quad & \|\mathbf{v}_{[t-1]}\|_2^2 \\
\text{s.t.} \quad & \mathbf{v}_{[t-1]} \in \Gamma_t(\mathbf{v}_{[t-1]}, A, b).
\end{aligned}
$$

We apply our methodology to the setting described in Section V-B and choose the set $[1,2]\times[1,2]$ as the target (grey in Fig. 2). As shown in Fig. 2, the feedforward input resulting from $\varepsilon = 0$ (red in Fig. 2) performs well on the 5 sample trajectories, steering them to the boundary of the target set, but yields poor performance on unseen samples. For larger ε (blue and green in Fig. 2), instead, the system trajectories are successfully steered to the target set, even for unseen noise realizations, at the price of a slight increase in cost.

References

[1] Cédric Villani. *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg, 2009.

[2] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[3] Soroosh Shafieezadeh-Abadeh, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv preprint arXiv:2303.03900*, 2023.

[4] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[5] Nam Ho-Nguyen and Stephen J. Wright. Adversarial classification via distributional robustness with Wasserstein ambiguity. *arXiv:2005.13815*, 2020.

[6] Liviu Aolaritei, Soroosh Shafieezadeh-Abadeh, and Florian Dörfler. The performance of Wasserstein distributionally robust M-estimators in high dimensions. *arXiv preprint arXiv:2206.13269*, 2022.

[7] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pages 6808–6817. PMLR, 2019.

[8] Liviu Aolaritei, Nicolas Lanzetti, Hongruyu Chen, and Florian Dörfler. Distributional uncertainty propagation via optimal transport. *arXiv preprint arXiv:2205.00343*, 2023.

[9] Dimitris Boskos, Jorge Cortés, and Sonia Martínez. Data-driven ambiguity sets with probabilistic guarantees for dynamic processes. *IEEE Transactions on Automatic Control*, 66(7):2991–3006, 2020.

[10] Christoph Mark and Steven Liu. Stochastic MPC with distributionally robust chance constraints. *IFAC-PapersOnLine*, 53(2):7136–7141, 2020.

[11] Jeremy Coulson, John Lygeros, and Florian Dörfler. Distributionally robust chance constrained data-enabled predictive control. *IEEE Transactions on Automatic Control*, 67(7):3289–3304, 2021.

[12] Liviu Aolaritei, Marta Fochesato, John Lygeros, and Florian Dörfler. Wasserstein tube mpc with exact uncertainty propagation. *arXiv preprint arXiv:2304.12093*, 2023.

[13] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal transport in systems and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:89–113, 2021.

[14] Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870, 2020.

[15] Ashish Cherukuri, Alireza Zolanvari, Goran Banjac, and Ashish R Hota. Data-driven distributionally robust optimization over a network via distributed semi-infinite programming. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4771–4775. IEEE, 2022.

[16] Ashish Cherukuri and Jorge Cortés. Cooperative data-driven distributionally robust optimization. *IEEE Transactions on Automatic Control*, 65(10):4400–4407, 2019.

[17] Daniel Owusu Adu, Tamer Başar, and Bahman Gharesifard. Optimal transport for a class of linear quadratic differential games. *IEEE Transactions on Automatic Control*, 67(11):6287–6294, 2022.

[18] Antonio Terpin, Nicolas Lanzetti, and Florian Dörfler. Dynamic programming in probability spaces via optimal transport. *arXiv preprint arXiv:2302.13550*, 2023.

[19] Rahul Singh, Isabel Haasler, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen. Inference with aggregate data: An optimal transport approach. *arXiv preprint arXiv:2003.13933*, 2020.

[20] Amirhossein Taghvaei and Prashant G Mehta. Optimal transportation methods in nonlinear filtering. *IEEE Control Systems Magazine*, 41(4):34–49, 2021.

[21] Vishaal Krishnan and Sonia Martínez. A probabilistic framework for moving-horizon estimation: Stability and privacy guarantees. *IEEE Transactions on Automatic Control*, 66(4):1817–1824, 2020.

[22] Soroosh Shafieezadeh-Abadeh, Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems*, pages 8474–8483, 2018.

[23] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[24] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 8 2015.

[25] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 4 2019.

[26] Liviu Aolaritei, Nicolas Lanzetti, and Florian Dörfler. Capture, propagate, and control distributional uncertainty. *arXiv preprint arXiv:2304.02235*, 2023.