Efficient Uncertainty Propagation with Guarantees in Wasserstein Distance

Eduardo Figueiredo^{*,1}, Steven Adams^{*,1}, Peyman Mohajerin Esfahani^{1,2}, Luca Laurenti^{1}

Abstract. In this paper, we consider the problem of propagating an uncertain distribution by a possibly non-linear function and quantifying the resulting uncertainty. We measure the uncertainty using the Wasserstein distance, and for a given input set of distributions close in the Wasserstein distance, we compute a set of distributions centered at a discrete distribution that is guaranteed to contain the pushforward of any distribution in the input set. Our approach is based on approximating a nominal distribution from the input set to a discrete support distribution for which the exact computation of the pushforward distribution is tractable, thus guaranteeing computational efficiency to our approach. Then, we rely on results from semi-discrete optimal transport and distributional robust optimization to show that for any $\epsilon > 0$ the error introduced by our approach can be made smaller than ϵ. Critically, in the context of dynamical systems, we show how our results allow one to efficiently approximate the distribution of a stochastic dynamical system with a discrete support distribution for a possibly infinite horizon while bounding the resulting approximation error. We empirically investigate the effectiveness of our framework on various benchmarks, including a 10-D non-linear system, showing the effectiveness of our approach in quantifying uncertainty in linear and non-linear stochastic systems.

1. Introduction

Modern cyber-physical systems are commonly affected by various sources of *uncertainty*. These include both the uncertainty caused by the intrinsic randomness in the system dynamics [38] and the uncertainty due to the use of statistical learning algorithms to estimate the unknown components/parameters of the system [28, 41]. Consequently, it is common that mathematical models are not only stochastic, but the distribution of the various random variables are themselves uncertain [39]. As a result, when these models are used in safety-critical applications, the resulting uncertainty cannot be neglected [13] and must be propagated through possibly non-linear functions. For instance, this is the case for stochastic dynamical systems, where the input distribution and the distribution of the noise affecting the system are commonly estimated from data and need to be propagated through the system dynamics for multiple (possibly infinite) time steps [7]. Unfortunately, how to propagate uncertain distributions through non-linear functions is still an open question. This leads to the main question in this paper: how can we efficiently propagate an uncertain distribution through a non-linear function with formal guarantees of correctness?

[*] Equal contribution. The authors are with (1) Delft University of Technology, and (2) the University of Toronto. Corresponding author's email: e.figueiredo@tudelft.nl. This research is partially supported by the NWO (grant OCENW.M.22.056) and the ERC Starting Grant TRUST-949796.

Propagating a distribution $\mathbb{P}$ through a function f is equivalent to computing the push-forward distribution of $\mathbb{P}$ by f denoted by $f\#\mathbb{P}$, which in the context of stochastic dynamical systems is equivalent to computing the Chapman-Kolmogorov Equation [33]. Unfortunately, in general, even when $\mathbb{P}$ is known, computing $f\#\mathbb{P}$ in closed form is not possible and requires approximations [27], such as moment matching [14] or discretization-based methods [25]. Unfortunately, these techniques either come with no correctness guarantees or are too computationally demanding due to the need to discretize the full state space and do not support any uncertainty in $\mathbb{P}$. When $\mathbb{P}$ is uncertain, the problem is exacerbated by the additional challenge of dealing with a possibly infinite set of distributions that must all be propagated through f. While this problem is receiving increasing interest [6, 16, 2], existing approaches are either limited to linear f or lack formality and scalability.

In this paper, given an uncertain distribution $\mathbb{P}$ and a non-linear function f, we present a framework to efficiently approximate $f\#\mathbb{P}$ via discrete distributions with formal quantification of the resulting uncertainty. To quantify the uncertainty, we rely on the Wasserstein distance [40]. This choice is motivated by the properties of the Wasserstein distance (i.e., it is a metric, it bounds the distance of the moments of the distributions, and convergence in the Wasserstein distance guarantees weak convergence) and its connection with optimal transport, which allows us to devise particularly efficient algorithms to solve our problem. Our approach is based on the fact that the Wasserstein distance between a continuous and a discrete distribution can be characterized as the solution of a semi-discrete optimal problem for which optimal solutions can be efficiently computed [34]. By using this connection and using techniques from distributional robust optimization and stochastic optimization [9, 10, 31, 18], we show that given a discrete distribution approximating $\mathbb{P}$, the Wasserstein distance between the pushforward of $\mathbb{P}$ by f and of its discrete approximation can be efficiently bounded, even when $\mathbb{P}$ is uncertain and f non-linear. The resulting bound can then be minimized by appropriately selecting the support of the approximating discrete distribution. This allows us to derive an efficient algorithmic framework that, given an uncertain distribution $\mathbb{P}$ and a non-linear function f and a given error threshold $\epsilon > 0$, returns a discrete distribution whose push-forward through f is guaranteed to be closer than ϵ to $f\#\mathbb{P}$.

We then show how our framework can be applied to formally approximate the state distribution of stochastic dynamical systems over time. We show that in contrast to existing results [6, 17], our approach can be successfully applied to linear and non-linear systems and for both finite and infinite time prediction horizons. In particular, under relatively mild assumptions on f, we prove the convergence of the approximation error of our approach over time to a fixed point. To further illustrate the usefulness of our framework, we perform an empirical evaluation on various benchmarks. In particular, we consider various linear and non-linear systems, including standard control benchmarks such as the Mountain Car [37] and Dubins Car [8], and a 10-D model of a neural network. The empirical analysis highlights how our framework can successfully approximate the push-forward distributions in both linear and non-linear cases and with relatively small discrete distributions, thus showcasing its potential to efficiently approximate complex distributions even in complex iterative prediction settings.

In summary, the main contributions of this work are listed below:

- **Uncertainty propagation**: upper-bounds on the uncertainty measured in terms of the Wasserstein distance of the pushforward of an uncertain probability distribution through a possibly non-linear function (Theorem 5.1), and a refined version under no ambiguity (Theorem 5.2);
- **Algorithm & convergence rate**: an efficient algorithmic procedure (Algorithm 1) to approximate the pushforward of an uncertain probability distribution by a discrete distribution, with guaranteed convergence in ρ-Wasserstein distance (Theorem 6.2);
- **Approximation error dynamics**: an application of our framework to stochastic dynamical systems for both finite and infinite prediction horizons (Theorem 7.1).

The paper is organized as follows. We formulate the problem in Section 4, present the formal uncertainty propagation error bounds in Section 5, and introduce an algorithmic procedure to propagate the uncertain distributions and compute these bounds in Section 6. Finally, in Section 8, we conduct an extensive empirical validation on several benchmarks, including complex non-linear dynamical systems, such as the Mountain and Dubins Car, and a 10-D non-linear system.

2. Related Works

Our work is connected with the distributionally robust optimization literature. In distributionally robust optimization, one is usually interested in computing the worst expected value of a certain transformation of a random vector w.r.t. a family of distributions $\mathcal{P}$, i.e., $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\xi\sim\mathbb{P}}[f(\xi)]$ [35, 15, 12, 18]. In particular, [31] provides techniques to characterize this worst-case expectation via convex optimization in the case where $\mathcal{P}$ is defined as a Wasserstein ambiguity set. Similarly, [18] prove that the distributionally robust problem is equivalent to a dual minimization program in $\mathbb{R}$-space for a large class of functions f and spaces $\mathcal{P}$, a result from which we took inspiration to demonstrate some of our results in this paper. In our work, however, we aim to find the worst *Wasserstein distance* between push-forwarded measures, given that they belong to a given set of probabilities close in Wasserstein distance. Furthermore, the Wasserstein distance is defined not as an expectation on the $\mathcal{P}$-space but as the infimum expectation of a specific cost function on the coupling space. Thus, a different framework must be devised to solve the problem.

Uncertainty propagation for various classes of functions has also been studied in the context of dynamical systems [27]. In [6], the authors provide a framework for the propagation of a set of distributions close in the Wasserstein distance in dynamical systems, where a distribution needs to be propagated through the system dynamics multiple times. These results have been applied in the context of stochastic model predictive control [5, 30]. However, in terms of numerical tractability, these techniques are specific to linear systems. Instead, in [16], the authors consider the uncertainty propagation problem in the context of random differential equations. The resulting bounds, however, involve different Wasserstein spaces, i.e., they propose a bound of type

$\mathbb{W}_{\rho_1}(f\#\mathbb{P}, f\#\mathbb{Q}) \leq C(f, \mathbb{P}, \mathbb{Q})\mathbb{W}_{\rho_2}(\mathbb{P}, \mathbb{Q})$ where[1] $\rho_1 < \rho_2$, thus not allowing for its use in settings where the uncertainty must be propagated multiple times and the information on the moments must be conserved[2]. Uncertainty propagation in stochastic dynamical systems has also been considered in [17], where the authors consider mixture approximations of the distribution of a dynamical system over time with bounds in total variation. However, the resulting bounds cannot be applied in the context of our paper where we approximate a continuous distribution with a discrete one, grow linearly with time independently of f, and, consequently, become uninformative for a not small prediction time horizon. A related work is also [2], which views neural networks as stochastic dynamical systems and presents an algorithmic framework to approximate a stochastic neural network with a mixture of Gaussian distributions with error bounds in Wasserstein distance. This approach is, however, specific to neural networks.

Another related line of work is that of stochastic abstraction-based methods, where a stochastic system is abstracted into a variant of a discrete Markov chain [1, 26] and that have also been recently extended to support distributional uncertainty on the system dynamics [22]. However, these works suffer from the state-space explosion problem due to the need to finely discretize the full support of the distributions. In contrast, our approach approximates a continuous distribution with a discrete one by selecting the support of the discrete distribution to minimize the distance from the continuous one. This allows us to reduce the size of the support of the resulting discrete distribution by only placing locations in the regions with high probability mass.

3. Preliminaries

Here, we provide the necessary preliminaries on the Wasserstein distance and the quantization of probability distributions.

3.1. Notation

For a vector $x \in \mathbb{R}^d$, we denote by $x^{(i)}$ its i-element. For a set $\mathcal{X} \subseteq \mathbb{R}^d$, the indicator function for $\mathcal{X}$ is denoted as $\mathbb{1}_{\mathcal{X}}(x) := 1$ if $x \in \mathcal{X}$; otherwise 0. For $\mathcal{X} \subseteq \mathbb{R}^d$, we denote a partition of $\mathcal{X}$ in N *regions* $\boldsymbol{\mathcal{R}} := \{\mathcal{R}_i\}_{i=1}^N$, i.e. $\mathcal{R}_i \subseteq \mathcal{X}$, $\bigcup_{i=1}^N \mathcal{R}_i = \mathcal{X}$, and $\forall i \neq j, \mathcal{R}_i \cap \mathcal{R}_j = \emptyset$. Given a Borel measurable space $\mathcal{X} \subseteq \mathbb{R}^d$, we denote by $\mathcal{B}(\mathcal{X})$ the Borel sigma algebra over $\mathcal{X}$ and by $\mathcal{P}(\mathcal{X})$ the set of probability distributions on $\mathcal{X}$. For a random variable x_t taking values in $\mathcal{X}$, $\mathbb{P}_{x_t} \in \mathcal{P}(\mathcal{X})$ represents the probability measure associated to x_t. For $N \in \mathbb{N}$, $\Pi^N := \{\pi \in \mathbb{R}^N_{\geq 0} : \sum_{i=1}^N \pi^{(i)} = 1\}$ is the N-simplex. A discrete probability distribution $\mathbb{D} \in \mathcal{P}(\mathcal{X})$ is defined as $\mathbb{D} = \sum_{i=1}^N \pi^{(i)} \delta_{c_i}$, where δ_c is the Dirac delta function centered at location $c \in \mathcal{X}$ and $\pi \in \Pi^N$ and N is the number of locations in the support of $\mathbb{D}$. The set of discrete probability distributions on $\mathcal{X}$ with at most N

[1] The term $C(f, \mathbb{P}, \mathbb{Q})$ is a constant upper-bounding the moment under both $\mathbb{P}$ and $\mathbb{Q}$ of a function only requiring local Lipschitz continuity from f, which is a less restrictive assumption compared to the piecewise Lipschitz continuity that we need in our work.

[2] The ρ-Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ is related to how close their ρ-moments are ([40, 2]). Propagating a bound in the ρ_2-Wasserstein space to the ρ_1-Wasserstein space implies a loss of information on the difference of the higher moments of the push-forwarded measures.

locations is denoted as $\mathcal{D}_N(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$. For a probability distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and a measurable function $g : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^q$, we denote the push-forward measure of $\mathbb{P}$ by g as $g\#\mathbb{P}$ such that for all $A \subset \mathcal{B}(\mathcal{Y})$, $(g\#\mathbb{P})(A) := \mathbb{P}(g^{-1}(A))$. We note that $g\#\mathbb{P}$ is still a probability distribution such that $g\#\mathbb{P} \in \mathcal{P}(\mathcal{Y})$.

3.2. **Wasserstein (or Kantorovich) distance**

Let $\rho \geq 1$, $\mathcal{X} \subseteq \mathbb{R}^d$, and define $\mathcal{P}_\rho(\mathcal{X})$ as the set of probability distributions with finite ρ-th moments under the L_ρ-norm, i.e. all $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ such that $\int_{\mathcal{X}} \|x\|^\rho \, d\mathbb{P}(x) < \infty$.

Definition 1 (ρ-Wasserstein distance)**.** *For $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_\rho(\mathcal{X})$ the ρ-Wasserstein distance $\mathbb{W}_\rho$ between $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$\mathbb{W}_\rho(\mathbb{P}, \mathbb{Q}) := \left(\inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^\rho \, d\gamma(x, y) \right)^{\frac{1}{\rho}} \tag{1}$$

where $\Gamma(\mathbb{P}, \mathbb{Q}) \subset \mathcal{P}_\rho(\mathcal{X} \times \mathcal{X})$ represents the set of joint probability distributions with given marginals $\mathbb{P}$ and $\mathbb{Q}$ (also known as couplings *between $\mathbb{P}$ and $\mathbb{Q}$), i.e., for all $\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})$, it holds that:*

$$\gamma(A \times \mathcal{X}) = \mathbb{P}(A), \ \gamma(\mathcal{X} \times A) = \mathbb{Q}(A) \qquad \forall A \in \mathcal{B}(\mathcal{X}).$$

Additionally, we define the ρ-Wasserstein ball of radius $\theta \geq 0$ centered at the probability distribution $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$, also called ambiguity set, as:

$$\mathbb{B}_\theta(\mathbb{P}) := \Big\{ \mathbb{Q} \in \mathcal{P}_\rho(\mathcal{X}) : \mathbb{W}_\rho(\mathbb{P}, \mathbb{Q}) \leq \theta \Big\}. \tag{2}$$

That is, $\mathbb{B}_\theta(\mathbb{P})$ contains all probability measures closer than θ to $\mathbb{P}$ according to $\mathbb{W}_\rho$.

We finally state an identity that follows directly from the definition of the Wasserstein distance and will be extensively employed in the following sections:

$$\mathbb{W}_\rho(f\#\mathbb{P}, f\#\mathbb{Q}) = \left(\inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} \|f(x) - f(y)\|^\rho \, d\gamma(x, y) \right)^{\frac{1}{\rho}}. \tag{3}$$

This identity states that the ρ-Wasserstein distance between the pushforward of two probability measures under a function f is equal to a ρ-Wasserstein-like distance between the original measures, using a modified cost structure $\big(\|.\| \circ (f \times f) \big)$.

3.3. **Quantization of probability distributions**

For $\mathcal{X} \subseteq \mathbb{R}^d$, we consider a $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_i\}_{i=1}^N$ in N *regions*. Further, we denote by $\boldsymbol{\mathcal{C}} = \{c_i\}_{i=1}^N$ a set of N points in $\mathbb{R}^d$, which we refer to as *locations* henceforward.

Definition 2 (Quantization of a probability distribution)**.** *For partition $\boldsymbol{\mathcal{R}}$ and set of locations $\boldsymbol{\mathcal{C}}$, the quantization operator $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}} : \mathcal{X} \to \mathcal{X}$ is defined by*

$$\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}(x) := \sum_{i=1}^N c_i \mathbb{1}_{\mathcal{R}_i}(x). \tag{4}$$

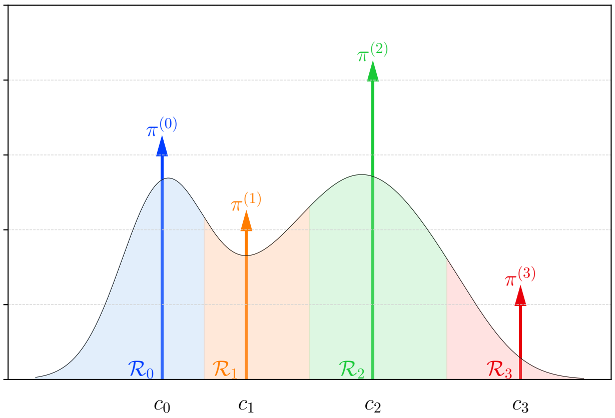

FIGURE 1. Schematic representation of the density of a continuous probability distribution $\mathbb{P}$, and its quantization $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}$, which has support of size $N = |\boldsymbol{\mathcal{C}}| = 4$, where we represent $\pi^{(i)} := \mathbb{P}(\mathcal{R}_i)$, a notation that will be commonly adopted in the rest of the paper.

That is, the quantization operator takes any point in the region $\mathcal{R}_i$ and brings it to the location c_i. For any probability distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, the quantization *(or* discretization*) of $\mathbb{P}$ is defined as the pushforward measure*

$$\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P} = \sum_{i=1}^{N} \mathbb{P}(\mathcal{R}_i)\delta_{c_i} \in \mathcal{D}_N(\mathcal{X}). \tag{5}$$

Note that in the definition of $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$, we do not assume any relationship between the partition and the locations, although it is natural to pick $c_i \in \mathcal{R}_i$. We should also stress that if, for a given set of locations $\boldsymbol{\mathcal{C}}$, one defines the partition as the Voronoi partition w.r.t. $\boldsymbol{\mathcal{C}}$, i.e., we take $\boldsymbol{\mathcal{R}}$ with each region being constructed as

$$\mathcal{R}_i := \Big\{ z \in \mathbb{R}^d : \|z - c_i\| \leq \|z - c_j\|, \forall j \neq i \Big\}, \tag{6}$$

where $\|.\|$ is the underlying norm, then the quantization operator $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ is equivalent to the signature operation described in [2]. An example of the quantization operator is shown in Figure 1.

The concept of quantizing a continuous probability distribution is well known in the literature [23, 4, 2] and the following result to compute the ρ-Wasserstein distance between $\mathbb{P}$ and $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}$ is a straightforward extension of Theorem 1 of [4], which we will employ in this work.

Proposition 3.1 (Quantization error). *Let $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$ and assume a given $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_i\}_{i=1}^N$ and set of locations $\boldsymbol{\mathcal{C}} = \{c_i\}_{i=1}^N$. Then, for any $\rho \geq 1$,*

$$\mathbb{W}_\rho(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \left(\sum_{k=1}^{N} \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) \right)^{\frac{1}{\rho}} \tag{7}$$

Furthermore, if $\boldsymbol{\mathcal{R}}$ is chosen to be the Voronoi partition w.r.t. $\boldsymbol{\mathcal{C}}$, then (7) *holds with equality.*

This result comes from the fact that the particular coupling which transports the probability mass of $\mathbb{P}$ in the region $\mathcal{R}_k$ to the location c_k belongs to $\Gamma(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$. Thus, its associated cost (right-hand side of (7)) upper-bounds the Wasserstein distance between these two distributions.

Remark 1 (Computing the quantization error)**.** *To compute the constrained ρ-moments in* (7)*, there are various approaches one can rely on. For instance, if $\mathbb{P}$ is a product measure, $\|.\|$ is the L_ρ-norm, and $\boldsymbol{\mathcal{R}}$ is a set of axis-aligned hyper-rectangles, then we have:*

$$\sum_{k=1}^{N} \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) = \sum_{k=1}^{N} \sum_{m=1}^{d} \int_{r_k^{(m)}} \left| x^{(m)} - c_k^{(m)} \right|^\rho d\mathbb{P}_m\big(x^{(m)}\big) \prod_{j \neq m} \mathbb{P}_j\big(r_k^{(m)}\big),$$

where $r_k^{(m)} := [a_k^{(m)}, b_k^{(m)}]$, and $\mathcal{R}_k = \prod_{m=1}^{d} r_k^{(m)}$. That is, we need to compute a set of constrained ρ-moment of the univariate distributions $\mathbb{P}_m$, which is analytically tractable for many distributions (especially - although not limited to - for $\rho \in \{1, 2\}$), including Gaussian (see Proposition 9 and Corollary 10 in [2] also for the general full covariance multivariate case), Uniform, Exponential, or Gamma distributions. Another particularly favorable case is when $\mathbb{P}$ is discrete, i.e., $\mathbb{P} \in \mathcal{D}_N(\mathbb{R}^d)$; in this case, the bounds can be computed directly because of the finiteness of the support of the distributions.

4. **Problem Formulation**

After having formally defined $\mathbb{W}_\rho$ and $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ we are now ready to state the main problem we consider in this paper. Given an uncertain distribution $\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})$, where $\mathbb{B}_\theta(\mathbb{P})$ is a Wasserstein ambiguity set of radius $\theta \geq 0$ centered at $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$ for $\mathcal{X} \subseteq \mathbb{R}^d$, and a possibly nonlinear measurable piecewise Lipschitz continuous function $f : \mathcal{X} \to \mathcal{Y}$, our goal is to find discrete approximations of the pushforward distribution of $\mathbb{Q}$ by f. In particular, we consider the following problem.

Problem 1. *For an error threshold $\epsilon > 0$, find a $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_i\}_{i=1}^N$ and locations $\boldsymbol{\mathcal{C}} = \{c_1, ..., c_N\}$ such that*

$$\left| \sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) - \sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P}) \right| \leq \epsilon. \tag{8a}$$

Furthermore, find a bound $\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}} \geq 0$ such that

$$\sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}. \tag{8b}$$

The goal of Problem 1 is to find arbitrarily accurate discrete approximations of the pushforward measure of an uncertain distribution and, critically, to bound the resulting uncertainty. Note that the convergence requirement in (8a) to $\sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P})$ is natural as if $\mathbb{P}$ and $\mathbb{Q}$ differ, then, in general, the distance of their pushforward distributions will not be zero. Hence, even if ϵ, the error introduced by the quantization, vanishes, then $\sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ may not. The need to compute $\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ in Problem 1 guarantees that in this paper we are not only interested in computing converging discrete approximations, but also in obtaining non-trivial error bounds for the resulting uncertainty quantification problem.

Problem 1 aims at generalizing existing methods to perform uncertainty propagation of probability distributions, such as non-linear filtering [11, 3] or sigma point methods [25], by computing formal error bounds on the error in terms of the Wasserstein distance and in selecting optimal discrete distribution approximations, which also accounts for the uncertainty in $\mathbb{P}$. Note also that for $\theta = 0$, (8b) reduces to bounding $\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$, that is, the error in the pushforward approximation of a discrete operator applied to a known distribution. While this is itself an important open problem [27], as we will illustrate in Example 1, we should already stress that in the case of uncertainty propagation in dynamical systems, which is the main application we consider in this paper, considering $\theta > 0$ in (8b) is essential.

Example 1 (Dynamical systems)**.** *Consider the general model of a discrete-time stochastic process*

$$x_{t+1} = f(x_t, w_t), \quad x_0 \sim \mathbb{P}_{x_0}, w_t \sim \mathbb{P}_{w_t},$$

where $\mathbb{P}_{x_0}$ and $\mathbb{P}_{w_t}$ represent, respectively, the distribution of the initial condition and of the noise affecting the system at time t. If f is non-linear or $\mathbb{P}_{w_t}$ is not Gaussian, then the distribution of the system at time t, $\mathbb{P}_{x_t}$, generally cannot be obtained in closed form and requires approximations [20, 27, 17]. A solution to Problem 1 allows one to approximate arbitrarily well $\mathbb{P}_{x_t}$ for any $t > 0$ with a discrete distribution $\hat{\mathbb{P}}_{x_t}$ by iteratively approximating the pushforward distribution $f\#\mathbb{P}_{x_t}$ and quantifying the approximation error. Note that for $t > 0$, the distribution of $\mathbb{P}_{x_t}$ is uncertain because of the uncertainty introduced by the quantization at the previous time steps. Consequently, approximating $\mathbb{P}_{x_{t+1}}$ requires one to consider $\theta > 0$ in Problem 1 to propagate the resulting uncertainty through f. In Section 7, we will show how a solution of Problem 1 allows us to efficiently compute approximations for $\mathbb{P}_{x_t}$ with formal guarantees of correctness in the ρ-Wasserstein metric.

We should also stress that the impact of a solution to Problem 1 is not limited to dynamical systems and, for instance, also represents a key contribution to the distributional robust uncertainty propagation quantification problem, where it is still an open question how to quantify $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P})$ when f is non-linear [6]. A solution to Problem 1 would give an efficient method to over-approximate this quantity[3].

Remark 2 (Wasserstein distance vs. divergences)**.** *A key advantage in using the ρ-Wasserstein distance to quantify the error compared to other commonly used quantities, such as KL divergence [19], is that bounds in the ρ-Wasserstein distance between two probability distributions can be used to bound their difference in moments ([2], Lemma 2), in probability ([18], Example 7), and many other further quantities of interest ([16], Section 4).*

Approach. In Section 5, we start by focusing on bounding $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$ for a given $\mathcal{R}, \mathcal{C}$ using results from stochastic optimization and properties of the Wasserstein distance and derive bounds both for $\theta > 0$ and $\theta = 0$. Then, in Section 6, we present an algorithm to efficiently select the partition $\mathcal{R}$ and locations $\mathcal{C}$, and we further prove the convergence of $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$ to $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P})$ for the resulting algorithm as the

[3] Indeed, as a corollary of (8a) and (8b), it holds that $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P}) \leq \mathcal{W}_{\mathcal{R},\mathcal{C}} + \epsilon$.

number of locations $|\boldsymbol{\mathcal{C}}|$ increases. Lastly, in Section 7, we show how our uncertainty propagation framework can be applied to approximate the state distributions in stochastic dynamical systems with formal ρ-Wasserstein guarantees for both finite and infinite prediction horizons. Section 8 provides experimental results on various benchmarks to show the effectiveness of our approach.

5. Error Bounds in Wasserstein Distance

In this section, we show how for a given quantization operator $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ one can efficiently bound $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ for any $\theta \geq 0$. Our main result is reported next and is based on a norm linearization around each location $c_k \in \boldsymbol{\mathcal{C}}$.

Theorem 5.1 (Uncertainty propagation of ambiguity sets)**.** *For $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$, assume a given $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_k\}_{k=1}^N$ and a set of locations $\boldsymbol{\mathcal{C}} = \{c_k\}_{k=1}^N$. For every $k \in \{1, ..., N\}$, call $\pi^{(k)} := \mathbb{P}(\mathcal{R}_k)$, and let $\alpha_k, \beta_k \in \mathbb{R}_+$ be such that for $x \in \mathcal{X}$*

$$\|f(x) - f(c_k)\|^\rho \leq \alpha_k \|x - c_k\|^\rho + \beta_k. \tag{9}$$

Further, denote

$$\theta_d = \left(\sum_{k=1}^N \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x)\right)^{\frac{1}{\rho}} \tag{10}$$

Then, for $\alpha_{\max} = \max_{k\in\{1,...,N\}} \alpha_k$, it holds that

$$\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \left(\alpha_{\max}(\theta + \theta_d)^\rho + \sum_{k=1}^N \pi^{(k)}\beta_k\right)^{\frac{1}{\rho}}. \tag{11}$$

The proof of Theorem 5.1 is given in Appendix A.2, where we rely on duality to relax the computation of $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ into a one-dimensional minimization problem that can be efficiently bounded by using the fact that $\mathbb{W}_\rho(\mathbb{P}, \#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ can be formulated as a semi-discrete optimal transport problem (Proposition 3.1), and on the local linearization of f given in (9). An algorithm to automatically select α_k and β_k for the various locations will be given in Section 6.1, while how to compute θ_d has already been mentioned in Remark 1. Before discussing the theoretical implications of Theorem 5.1 in the rest of this section, we should stress that a potential source of conservatism in Theorem 5.1 is in the linearization around each location c_k, which must hold for all $x \in \mathcal{X}$ and not just locally in $\mathcal{R}_k$. This is due to the uncertainty of not knowing $\mathbb{P}$ exactly. In Subsection 5.1, we show that such a requirement can be relaxed, and consequently, the bound is improved when there is no ambiguity set, i.e., $\theta = 0$.

Remark 3 (Lipschitz-based uncertainty propagation)**.** *Note that a straightforward corollary of Theorem 5.1 is that*

$$\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \mathcal{L}_f\big(\theta + \theta_d\big), \tag{12}$$

where $\mathcal{L}_f$ is the Lipschitz constant of f according to the L_ρ-norm[4]*. However, in general, as we give intuition in Example 2 below, and we will show empirically in Section 8, the bound in* (11) *is generally substantially tighter and can return bounds that are orders of magnitude smaller. The intuition is that in the regions $\mathcal{R}_i$ where the local Lipschitz constant of f is large, one can rely on a larger β_j to obtain a lower $\alpha_{\max}$. If, in these regions, the probability mass of $\mathbb{P}$ is small (and, consequently, $\pi^{(j)}$ is low), then the bound could substantially improve. Note that an exception is when f is linear, where it is easy to show that the bounds in* (11) *and* (12) *coincide. In fact, if f is linear, i.e., $f(x) := Ax + b$, Theorem 5.1 reduces to*

$$\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P}) \le \|A\| \left(\theta + \theta_d\right), \tag{13}$$

where $\|A\| := \sup_{x\in\mathcal{X}} \frac{\|Ax\|}{\|x\|}$ is the induced norm of A, which is equivalent to the global Lipschitz constant of f.

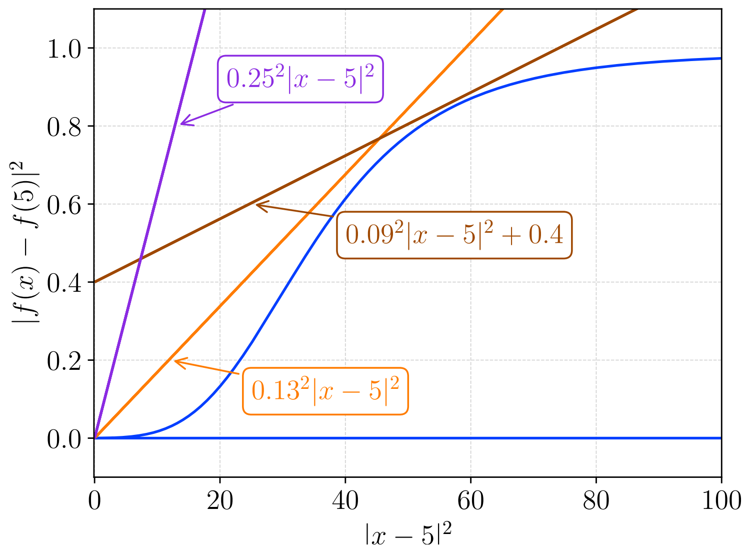

FIGURE 2. There exists infinite admissible pairs (α, β) such that (9) holds. In particular, we show three of them: $(0.25^2, 0)$ (purple line), $(0.13^2, 0)$ (orange), and $(0.09^2, 0.4)$ (brown).

Example 2 (Local vs. Lipschitz-based norm approximations)**.** *Let $\rho = 2$, and $f(x) = \frac{1}{1+e^{-x}}$ be a sigmoid function, whose Lipschitz constant w.r.t. the L_2-norm is $\mathcal{L}_f = 0.25$. We consider the location $c = 5$; see Figure 2. We can upper bound $|f(x) - f(5)|^2$ for any $x \in \mathbb{R}$ with* (9) *by choosing i) $(\alpha, \beta) = (0.25^2, 0)$, ii) $(\alpha, \beta) = (0.13^2, 0)$, or iii) $(\alpha, \beta) = (0.09^2, 0.4)$. The first observation is that since the location $c = 5$ is far from the region where the global Lipschitz is found ($x = 0$), α can be chosen to be considerably smaller than $\mathcal{L}_f^2$ even for $\beta = 0$ ($0.13^2 < 0.25^2$). Further, since f is bounded, by increasing the bias β, one can decrease α even further (see the brown line). In Section 6.1, we explain how to automatically select these parameters.*

[4] This follows by observing that for all $j \in \{1, ..., N\}$ one can select $\beta_j = 0$ and $\alpha_j = \mathcal{L}_f$. The resulting choice always satisfies (9) by the definition of Lipschitz constant.

5.1. **No ambiguity case:** $\theta = 0$.

As mentioned, when $\theta = 0$, the bound in Theorem 5.1 can be improved. In fact, in the proof of Theorem 5.1, as will be discussed in detail in Section 5.2, to obtain a tractable reformulation, we seek the worst plausible joint distribution among all couplings such that one of the marginals is $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}$ and the other is any distribution $\mathbb{Q} \in \mathbb{B}_{\theta+\theta_d}(\mathbb{P})$. Instead, when $\mathbb{P}$ is known, as in the case $\theta = 0$, we can design a specific transport plan that generally leads to improved bounds, as shown in Theorem 5.2 below.

Theorem 5.2 (Uncertainty propagation under no ambiguity)**.** *For $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$. Assume a given partition $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_k\}_{k=1}^N$ and a set of locations $\boldsymbol{\mathcal{C}} = \{c_k\}_{k=1}^N$. For every $k \in \{1, ..., N\}$, call $\pi^{(k)} := \mathbb{P}(\mathcal{R}_k)$, and let $\alpha_k, \beta_k \in \mathbb{R}_+$ be such that for $x \in \mathcal{R}_k$, it holds that*

$$\|f(x) - f(c_k)\|^\rho \leq \alpha_k \|x - c_k\|^\rho + \beta_k. \tag{14}$$

Then,

$$\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \left(\sum_{k=1}^N \alpha_k \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) + \sum_{k=1}^N \pi^{(k)}\beta_k\right)^{\frac{1}{\rho}} \tag{15}$$

The proof of Theorem 5.2 is reported in Appendix A.3. Note that, differently from Theorem 5.1, the norm overapproximation in (14) is local, i.e. for each region $\mathcal{R}_k$, (14) has to hold for every $x \in \mathcal{R}_k$, instead of $x \in \mathcal{X}$ as in Theorem 5.1. This is a consequence of the fact that, in the setting of Theorem 5.2, $\mathbb{P}$ is known with no uncertainty.

5.2. **Conservatism under ambiguity**

We should stress that, although $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ is right-continuous in $\theta = 0$, the bound in Theorem 5.1 does not generally converge to the one in 5.2 as $\theta \downarrow 0$. To see this, note that the proof of Theorem 5.1 is based on a worst-case analysis. In particular, as detailed in Appendix A.2, we define $S_\theta(\mathbb{T}) := \left\{\gamma \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) : \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_2\|^\rho \, d\gamma(x_1, x_2) \leq \theta^\rho, \mathrm{proj}_2\#\gamma = \mathbb{T}\right\}$, and show

$$\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \sup_{\gamma\in S_{\theta+\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \|f(x) - f(y)\|^\rho \, d\gamma(x, y)$$

By taking the limit $\theta \downarrow 0$ on both sides, we have

$$\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \sup_{\gamma\in S_{\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \|f(x) - f(y)\|^\rho \, d\gamma(x, y) \tag{16}$$

On the other hand, to prove Theorem 5.2, we design a specific coupling $\gamma^* \in \Gamma(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ that achieves $\int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^\rho \, d\gamma^*(x, y) = \theta_d^\rho$, as reported in (27) in the Appendix A.2, and bound

$$\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \int_{\mathcal{X}\times\mathcal{X}} \|f(x) - f(y)\|^\rho \, d\gamma^*(x, y) \tag{17}$$

While it holds by construction that γ^* is in $S_{\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$, the elements in $S_{\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ do not necessarily have $\mathbb{P}$ as one of the marginals, i.e., are not necessarily a member of $\Gamma(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$. Hence, in general, the bounds in (16) and that in (17) for $\theta = 0$ differ [5].

6. Selecting approximate discrete distributions

In this section, we provide an algorithmic approach to automatically select $\boldsymbol{\mathcal{R}}$, $\boldsymbol{\mathcal{C}}$, and the linearization coefficients in Theorems 5.1 and 5.2. First, in Section 6.1, for any $c \in \mathcal{X} \subseteq \mathbb{R}^d$, we present an algorithm to compute coefficient pairs (α, β) such that either (9) or (14) holds, and the corresponding error bound in (11) or (15) is minimized. Then, we provide a practical approach to construct a partition $\boldsymbol{\mathcal{R}}$ and a set of locations $\boldsymbol{\mathcal{C}}$ that guarantees that the approximation error defined in (8a) can be made arbitrarily small.

6.1. Norm approximation algorithm

As observed in Example 2, given a location $c \in \boldsymbol{\mathcal{C}}$, there exist infinite combinations of (α, β) to generate the upper-bounds for $\|f(x) - f(c)\|^\rho$ for all $x \in \mathcal{X}$. Unfortunately, due to the possibly non-linear nature of f, it is generally intractable to minimize the error bound in Theorem 5.1 or 5.2 with respect to all feasible linearization combinations. Hence, in practice, we focus on combinations of type (i) $(0, \beta)$ and type (ii) $(\alpha, 0)$, which can be computed efficiently. Specifically, for combinations of type (i), where f remains bounded in the region where the linear bounds must hold, we select $(\alpha, \beta) = (0, \sup_{x\in\mathcal{X}} \|f(x) - f(c)\|^\rho)$. For type (ii) combinations, we select $(\alpha, \beta) = (\sup_{x\in\mathcal{X}} \|f(x) - f(c)\|^\rho \,/\, \|x - c\|^\rho\,, 0)$. Due to the typically non-convex nature of these optimization problems, in practice, we utilize approximate solutions obtained via bound propagation techniques.[6] That is, for each region $\mathcal{S}_k$ of a $\mathcal{X}$-partition $\boldsymbol{\mathcal{S}}$, we compute linear maps $\check{A}_k(x - c)$ and $\hat{A}_k(x - c)$, and vectors $\check{b}_k$ and $\hat{b}_k$, that satisfy:

$$\check{A}_k(x - c) \preceq f(x) - f(c) \preceq \hat{A}_k(x - c) \tag{18}$$

$$\check{b} \preceq f(x) - f(c) \preceq \hat{b}, \tag{19}$$

for all $x \in \mathcal{S}_k$. We then use that

$$\sup_{x\in\mathcal{X}} \frac{\|f(x) - f(c)\|^\rho}{\|x - c\|^\rho} \leq \max_{k\in\{1,\ldots,N\}} \left(\|\check{A}_k\|^\rho, \|\hat{A}_k\|^\rho\right). \tag{20}$$

and

$$\sup_{x\in\mathcal{X}} \|f(x) - f(c)\|^\rho \leq \max_{k\in\{1,\ldots,N\}} \left(\|\check{b}_k\|^\rho, \|\hat{b}_k\|^\rho\right) \tag{21}$$

to set α for combinations of type (ii), and β for combinations of type (i) and respectively. Note that, for Theorem 5.2 where the norm-linearization has to hold only over a region $\mathcal{R} \subseteq \mathcal{X}$, we follow the same procedure, replacing $\mathcal{X}$ by $\mathcal{R}$.

[5] For instance, consider $\mathbb{P} = \delta_{(0,0)}$, $\boldsymbol{\mathcal{R}} := \{\mathbb{R}^2\}$, $\boldsymbol{\mathcal{C}} := \{(0, \theta_d)\}$, and $f(x) := \mathrm{diag}(2, 0.1)x$. Then, $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P} = \delta_{(0,\theta_d)}$, and $\gamma^* = \delta_{(0,0)\times(0,\theta_d)}$. Note that $\tilde{\gamma} := \delta_{(0,\theta_d)\times(\theta_d,\theta_d)} \in S_{\theta_d}(\delta_{(0,\theta_d)})$, and hence, $\int_{\mathcal{X}\times\mathcal{X}} \|f(x) - f(y)\|^\rho \, d\gamma^*(x, y) = 0.1^\rho\theta_d^\rho$, while $\int_{\mathcal{X}\times\mathcal{X}} \|f(x) - f(y)\|^\rho \, d\tilde{\gamma}(x, y) = 2^\rho\theta_d^\rho$, a significantly larger value.

[6] For our experiments, we use the linear bound propagation techniques from [29] to compute the linear maps.

Algorithm 1 details a procedure to select $(\alpha_k, 0)$ or $(0, \beta_k)$ for all $c_k \in \boldsymbol{\mathcal{C}}$ for Theorem 5.1. The case for Theorem 5.2 follows similarly. Algorithm 1 is based on the fact that Theorem 5.1 only depends on the maximum value of the α_k coefficients for all locations $c_k \in \boldsymbol{\mathcal{C}}$, so, by ordering those coefficients in descending order, we can iteratively verify whether replacing $\alpha_k \|x - c_k\|^\rho$ approximations in (9) for β_k lead to a tighter bound. As discussed in Remark 3, this will generally be the case when $\mathbb{P}(\mathcal{R}_k)$ is low. More specifically, we start by computing $\bar{\alpha} := (\alpha_1, ..., \alpha_N)$ and $\bar{\beta} := (\beta_1, ..., \beta_N)$ using (20) and (21), respectively, for each $c_k \in \boldsymbol{\mathcal{C}}$ (line 2), and θ_d in (10) as explained in Remark 1 (line 3). We then compute the first candidate for the bound, by applying Theorem 5.1 with $\bar{\alpha}$ (line 4). In line 5, we sort $\bar{\alpha}$ in descending order (and sort accordingly $\bar{\beta}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{C}}$). As mentioned above, the strategy is to try to replace the highest α_k by zero, and include instead β_k, while verifying if the bound decreases. This is implemented in the *for* loop in lines 6-12.

Algorithm 1: Compute least conservative bound in Thm 5.1 for a given quantization operator

```
   Input: $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}}$, set of locations $\boldsymbol{\mathcal{C}}$, radius $\theta$, distribution $\mathbb{P}$
   Output: Least conservative $\rho$-Wasserstein bound in Thm 5.1 given $\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{C}}$
1  function BoundGivenQuantizationOperator($\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{C}}, \theta, \mathbb{P}$):
2      $(\bar{\alpha}, \bar{\beta}) \leftarrow$ (Eqns (20) & (21) for $c_k \in \boldsymbol{\mathcal{C}})_{k=1}^{|\boldsymbol{\mathcal{C}}|}$
3      $\theta_d \leftarrow$ Eqn (10)
4      $\mathbb{W} \leftarrow \max_{\alpha \in \bar{\alpha}} \alpha(\theta + \theta_d)$
5      $\bar{\alpha}_{\text{sorted}}, \bar{\beta}_{\text{sorted}}, \boldsymbol{\mathcal{R}}_{\text{sorted}}, \boldsymbol{\mathcal{C}}_{\text{sorted}} \leftarrow$ sort descendingly according to $\bar{\alpha}$
6      for $k \in \{1, ..., |\boldsymbol{\mathcal{C}}|\}$ do
7          $b_k \leftarrow \sum_{j=1}^{k} \mathbb{P}(\mathcal{R}_{\text{sorted},j}) \bar{\beta}_{\text{sorted},j}$
8          $\tilde{\mathbb{W}} \leftarrow \left( \bar{\alpha}_{\text{sorted},k+1}(\theta + \theta_d)^\rho + b_k \right)^{\frac{1}{\rho}}$
9          if $\mathbb{W} \leq \tilde{\mathbb{W}}$ then
10             break
11         else
12             $\mathbb{W} \leftarrow \tilde{\mathbb{W}}$
13     return $\mathbb{W}$
```

6.2. Constructing a converging quantization operator

After having discussed how to select the linearization coefficients in Theorem 5.1 and 5.2, what is left to do is to explain how to effectively construct a quantization operator $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$, i.e. a $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}}$ and a set of locations $\boldsymbol{\mathcal{C}} \subset \mathcal{X}$, such that the convergence requirement in Problem 1 holds for any given $\epsilon > 0$. We start with the following lemma, which is a straightforward consequence of the triangular inequality, showing that to satisfy Problem 1, it is enough to select $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ to minimize $\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$. This result implies that to guarantee an arbitrarily small ϵ, it is enough to optimize $\boldsymbol{\mathcal{R}}$, $\boldsymbol{\mathcal{C}}$ w.r.t. to $\mathbb{P}$ even if $\theta > 0$.

Lemma 6.1. *Let $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$. For any $\mathcal{X}$-partition $\boldsymbol{\mathcal{R}}$, and set of locations $\boldsymbol{\mathcal{C}}$, it holds that*

$$\left| \sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) - \sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P}) \right| \leq \mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \tag{22}$$

Lemma 6.1 is used in the next theorem to show that even taking $\boldsymbol{\mathcal{R}}$ as a uniform partitioning of any compact set containing enough probability mass of $\mathbb{P}$ to select $\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{C}}$ can guarantee a solution to Problem 1. An improved, non-uniform, partitioning scheme will then be given in Remark 4.

Theorem 6.2 (Convergence rate)**.** *For $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$ and $\rho \geq 1$. For any $\epsilon > 0$, consider a cubic compact set $\bar{\mathcal{X}} \subseteq \mathcal{X}$ such that $\int_{\mathcal{X}\setminus\bar{\mathcal{X}}} \|x - \bar{c}\|^\rho \, d\mathbb{P}(x) \leq \frac{\epsilon^\rho}{2\mathcal{L}_f^\rho}$ for some $\bar{c} \in \mathcal{X}$. Further, consider $\boldsymbol{\mathcal{R}} \coloneqq \{\mathcal{R}_k\}_{k=1}^N$ as a uniform $\bar{\mathcal{X}}$-partition of $\bar{\mathcal{X}}$ in $N \geq \left(\frac{2^{\frac{1}{\rho}} \mathcal{L}_f d^{\frac{1}{\rho}} \|\bar{\mathcal{X}}\|_\infty}{\epsilon} \right)^d$ hypercubic regions, and $\boldsymbol{\mathcal{C}}$ as set of the centers of each hypercube $\mathcal{R}_k$. Then, for $\boldsymbol{\mathcal{R}}^* = \boldsymbol{\mathcal{R}} \cup \{\mathcal{X} \setminus \bar{\mathcal{X}}\}$ and $\boldsymbol{\mathcal{C}}^* = \boldsymbol{\mathcal{C}} \cup \{\bar{c}\}$, it holds that:*

$$\left| \sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}}^*,\boldsymbol{\mathcal{C}}^*}\#\mathbb{P}) - \sup_{\mathbb{Q} \in \mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P}) \right| \leq \epsilon. \tag{23}$$

The convergence rate reported in Theorem 6.2 is conservative due to two factors: i) it relies on linearization coefficients $(\alpha, \beta) = (\mathcal{L}_f, 0)$ in Theorem 5.2, which generally leads in over-conservative error bounds, as discussed in Remark 3, ii) Theorem 6.2 is proven w.r.t. a uniform partitioning of an appropriately selected compact set. Consequently, it is evident that non-uniform partitioning approaches that directly minimize the bounds in Theorem 5.2 would lead to improved bounds. In particular, we can rely on Lemma 6.1, which implies that the quantization error is bounded by

$$\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \leq \underbrace{\left(\sum_{k=1}^N \alpha_k \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) + \sum_{k=1}^N \pi^{(k)} \beta_k \right)^{\frac{1}{\rho}}}_{=:\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}} \leq \mathcal{L}_f \theta_d, \tag{24}$$

where $\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ is the error bound from Theorem 5.2. Consequently, by selecting $\boldsymbol{\mathcal{R}}$ and $\boldsymbol{\mathcal{C}}$ to minimize θ_d, we indirectly reduce the error bounds. This approach may lead to greatly improved bounds compared to a uniform partitioning approach, as we will illustrate empirically in Section 8.

Remark 4 (Partitioning for normalizing flows of Gaussians)**.** *When $\mathbb{P}$ is Gaussian or a normalizing flow of a latent Gaussian distribution [32],[7] we can rely on Algorithm 2 from [2] to obtain $\boldsymbol{\mathcal{C}}$ that minimize θ_d, with $\boldsymbol{\mathcal{R}}$ being the Voronoi partition w.r.t. $\boldsymbol{\mathcal{C}}$. Since Algorithm 2 from [2] guarantees that θ_d converges to zero as N increases, we can iteratively increase the number of locations N until $\mathcal{L}_f \theta_d \leq \epsilon$, and consequently, according to (24), $\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}} \leq \epsilon$, where $\epsilon > 0$ is the desired error threshold. The non-uniform partition $\boldsymbol{\mathcal{R}}$ resulting from Algorithm 2 in [2] typically leads to a convergence rate that is significantly better than the one presented in Theorem 6.2.*

[7] For normalizing Gaussian distribution flows, i.e. $\mathbb{P} \coloneqq g\#\mathcal{N}(\mu, \Sigma)$ for some known piecewise Lipschitz continuous function g, one can use that $\mathbb{W}_\rho(g\#\mathcal{N}(\mu, \Sigma), g\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathcal{N}(\mu, \Sigma)) \leq \mathcal{L}_g \mathbb{W}_\rho(\mathcal{N}(\mu, \Sigma), \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathcal{N}(\mu, \Sigma))$.

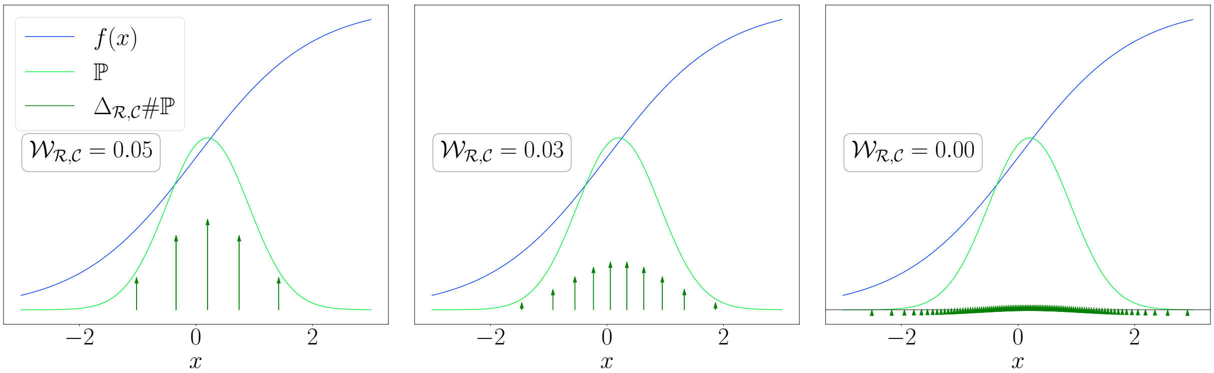

FIGURE 3. Quantization of $\mathbb{P} = \mathcal{N}(0.2, 0.5)$, $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}$, constructed as described in Remark 4, for $|\boldsymbol{\mathcal{C}}| \in \{5, 10, 10^2\}$, and the corresponding bound $\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ for f the sigmoid function.

Example 3 (Efficacy of Algorithm 1)**.** *Let $\rho = 2$. Consider again the sigmoid function $f : \mathbb{R} \to \mathbb{R}$ of Example 2. Further, let $\mathbb{P} = \mathcal{N}(0.2, 0.5)$. Figure 3, illustrates $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}$ and shows $\mathcal{W}_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}$ for different number of location $|\boldsymbol{\mathcal{C}}|$. Note how the bound monotonically decreases and reaches a value in the order of 10^{-2} with only 10 locations.*

7. Iterative predictions for stochastic dynamical systems

In this section, we show how our results can be used to generate provably correct discrete approximations for stochastic dynamical systems with formal guarantees in the ρ-Wasserstein distance. To do so, we consider the general model of a discrete-time stochastic process already introduced in Example 1:

$$x_{t+1} = f(x_t, \omega_t), \quad x_0 \sim \mathbb{P}_{x_0}, \omega_t \sim \mathbb{P}_\omega, \tag{25}$$

where the measurable function $f : \mathcal{X} \times \mathcal{W} \to \mathcal{X}$, with $\mathcal{X} \subseteq \mathbb{R}^d$ as the state space and $\mathcal{W} \subseteq \mathbb{R}^q$ as uncertainty space, represents the one-step dynamics of the system. Here, x_0 is the initial condition of the system, assumed to be distributed with distribution $\mathbb{P}_{x_0} \in \mathcal{P}(\mathcal{X})$, and ω_t is an i.i.d. process noise distributed according to $\mathbb{P}_\omega \in \mathcal{P}(\mathcal{W})$. We denote the state-noise joint distribution at time t by $\mathbb{P}_t := \mathbb{P}_{x_t} \times \mathbb{P}_\omega$. As previously mentioned in Example 1, if f is non-linear or $\mathbb{P}_w$ non-Gaussian, the distribution $\mathbb{P}_{x_t}$ of the system at time t becomes intractable. In this Section, we show how our solution of Problem 1 allows one to obtain a tractable (discrete) distribution $\hat{\mathbb{P}}_{x_t} \in \mathcal{P}(\mathcal{X})$ such that $\mathbb{W}_\rho(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t}) \leq \delta$, for a given error threshold $\delta > 0$ for any $t > 0$.

Our approach is summarized in Figure 4, where for a time t, we denote by $\boldsymbol{\mathcal{C}}_t = \{c_{t,1}, ..., c_{t,N_t}\}$, $\boldsymbol{\mathcal{R}}_t = \{\mathcal{R}_{t,1}, ..., \mathcal{R}_{t,N_t}\}$, respectively, the locations and regions for the discrete approximation of the system at time t, to emphasize how this can change over time. To describe our approach, we start with $t = 0$, assuming $\mathbb{P}_{x_0}$ is known, and setting $\hat{\mathbb{P}}_{x_0} = \mathbb{P}_{x_0}$. For $t = 1$, the true state distribution is given by $\mathbb{P}_{x_1} = f\#\mathbb{P}_0$, which, as we have previously argued, is generally intractable, Thus, as showed in Figure 4, we define the approximate state-noise joint distribution as $\hat{\mathbb{P}}_0 = \hat{\mathbb{P}}_{x_0} \times \mathbb{P}_\omega$. We then apply the quantization operation using a $(\mathcal{X} \times \mathcal{W})$-partition $\boldsymbol{\mathcal{R}}_0$ and a set of

$$\begin{tikzcd}
\textit{Actual state distr.} & \mathbb{P}_{x_0} \arrow[r] \arrow[d] & \mathbb{P}_{x_1} \arrow[r] & \mathbb{P}_{x_2} \arrow[r] & \dots \\
\textit{Joint distr.} & \underbrace{\mathbb{P}_{x_0} \times \mathbb{P}_{\omega}}_{\hat{\mathbb{P}}_0} \arrow[d, "\theta_{d,0}"] & \underbrace{\hat{\mathbb{P}}_{x_1} \times \mathbb{P}_{\omega}}_{\hat{\mathbb{P}}_1} \arrow[d, "\theta_{d,1}"] & \underbrace{\hat{\mathbb{P}}_{x_2} \times \mathbb{P}_{\omega}}_{\hat{\mathbb{P}}_2} \arrow[d, "\theta_{d,2}"] & \dots \\
\textit{Quantization} & \Delta_{\boldsymbol{\mathcal{R}}_0,\boldsymbol{\mathcal{C}}_0}\#\hat{\mathbb{P}}_0 \arrow[d] & \Delta_{\boldsymbol{\mathcal{R}}_1,\boldsymbol{\mathcal{C}}_1}\#\hat{\mathbb{P}}_1 \arrow[d] & \Delta_{\boldsymbol{\mathcal{R}}_2,\boldsymbol{\mathcal{C}}_2}\#\hat{\mathbb{P}}_2 \arrow[d] & \dots \\
\textit{Approximator} & \underbrace{f\#\Delta_{\boldsymbol{\mathcal{R}}_0,\boldsymbol{\mathcal{C}}_0}\#\hat{\mathbb{P}}_0}_{\hat{\mathbb{P}}_{x_1}} \arrow[uur] & \underbrace{f\#\Delta_{\boldsymbol{\mathcal{R}}_1,\boldsymbol{\mathcal{C}}_1}\#\hat{\mathbb{P}}_1}_{\hat{\mathbb{P}}_{x_2}} \arrow[uur] & \underbrace{f\#\Delta_{\boldsymbol{\mathcal{R}}_2,\boldsymbol{\mathcal{C}}_2}\#\hat{\mathbb{P}}_2}_{\hat{\mathbb{P}}_{x_3}} \arrow[uur] & \dots \\
\textit{Bounds} & \text{Thm 5.2} & \text{Thm 5.1} & \text{Thm 5.1} & \dots
\end{tikzcd}$$

FIGURE 4. Discrete approximation scheme for stochastic dynamical systems with formal guarantees on the ρ-Wasserstein distance, $\mathbb{W}_\rho(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t})$.

locations $\boldsymbol{\mathcal{C}}_0 \subset \mathcal{X} \times \mathcal{W}$, and propagate it through f, resulting in the approximate state (discrete) distribution $\hat{\mathbb{P}}_{x_1} = f\#\Delta_{\boldsymbol{\mathcal{R}}_0,\boldsymbol{\mathcal{C}}_0}\#\hat{\mathbb{P}}_0$. Note that the latter consists of a straightforward application of a f transformation to the support of a discrete distribution, hence providing a tractable propagation through time. This process is repeated for the next time steps, where $(\mathcal{X} \times \mathcal{W})$-partitions $\boldsymbol{\mathcal{R}}_t$ and locations $\boldsymbol{\mathcal{C}}_t$ are chosen such that the requirement in (1) is met for some predefined $\epsilon > 0$.

The next result shows how our framework can be applied to bound $\mathbb{W}_\rho(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t})$ for any $t \geq 0$. Furthermore, critically, we show that if f is contractive, the resulting error bounds propagation will reach a fixed point, allowing for infinite-time prediction horizons.

Theorem 7.1 (Approximation error dynamics)**.** *Given* $\epsilon > 0$*, let* $\boldsymbol{\mathcal{R}}_t$ *be* $(\mathcal{X} \times \mathcal{W})$*-partitions and* $\boldsymbol{\mathcal{C}}_t \subset \mathcal{X} \times \mathcal{W}$ *sets of locations such that* $\theta_{d,t} = \left(\sum_{k=1}^{N_t} \int_{\mathcal{R}_{t,k}} \|x - c_{t,k}\|^\rho \, d\hat{\mathbb{P}}_t(x) \right)^{\frac{1}{\rho}} \leq \epsilon$ *for every* $t \geq 0$*. Consider the following iterative process describing the approximation error evolution for* $t \in \mathbb{N}_{>0}$*:*

$$\theta_1 = \left(\sum_{k=1}^{N_0} \alpha_{0,k} \int_{\mathcal{R}_{0,k}} \|x - c_{0,k}\|^\rho \, d\hat{\mathbb{P}}_0(x) + \sum_{k=1}^{N_0} \hat{\mathbb{P}}(\mathcal{R}_{0,k}) \beta_{0,k} \right)^{\frac{1}{\rho}},$$

$$\theta_{t+1} = \left(\alpha_{max,t} (\theta_t + \epsilon)^\rho + \sum_{k=1}^{N_t} \hat{\mathbb{P}}(\mathcal{R}_{t,k}) \beta_{t,k} \right)^{\frac{1}{\rho}}.$$

Then, for any $t > 0$*, the following holds:*

i) $\mathbb{W}_\rho(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t}) \leq \theta_t$ *.*

ii) If the dynamics f in (25) *is Lipschitz continuous in (x, ω) with constant $\mathcal{L}_f < 1$, then*

$$\lim_{t\to\infty} \mathbb{W}_\rho(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t}) \leq \frac{\mathcal{L}_f}{1-\mathcal{L}_f}\epsilon\,.$$

The proof of Theorem 7.1 is reported in Appendix A.6 and follows from a combination of Theorem 5.1 and 5.2 with the Banach Fixed Point Theorem [21]. Theorem 7.1 has many consequences. First of all, the bound does not necessarily grow with time; it is possible that $\theta_{t+1} < \theta_t$ if the dynamics contracts sufficiently. This is a fundamental advantage with respect to existing approaches for the same problem, whose bounds tend to grow linearly with time [17]. Furthermore, Theorem 7.1 guarantees that if f is contracting w.r.t. (x, ω), i.e., $\mathcal{L}_f < 1$, then the approximation error will reach a fixed point. Note also that the bound for the fixed point of the error reported in case ii) in Theorem 7.1 is stated using the linearization coefficients from Remark 3, i.e., $(\alpha_k, \beta_k) = (\mathcal{L}_f, 0)$ for all k. Consequently, in practice, the approach in Figure 4 may yield a smaller bound. Notably, as empirically shown in Section 8, our approach can lead to a fixed point for θ_d even when $\mathcal{L}_f > 1$ if f is bounded.

Remark 5 (Separable dynamics)**.** *For a process with separable dynamics as $f(x, \omega) = g(x) + s(\omega)$, where g and s are given piecewise Lipschitz continuous functions, we observe:*

$$\begin{aligned}\mathbb{W}_\rho(\mathbb{P}_{x_{t+1}}, \hat{\mathbb{P}}_{x_{t+1}}) &= \mathbb{W}_\rho(g\#\mathbb{P}_{x_t} * s\#\mathbb{P}_\omega, g\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\hat{\mathbb{P}}_{x_t} * s\#\Delta_{\boldsymbol{\mathcal{R}}_\omega,\boldsymbol{\mathcal{C}}_\omega}\#\hat{\mathbb{P}}_\omega)\\ &\leq \mathbb{W}_\rho(g\#\mathbb{P}_{x_t}, g\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\hat{\mathbb{P}}_{x_t}) +\\ &\qquad\qquad \mathbb{W}_\rho(s\#\mathbb{P}_\omega, s\#\Delta_{\boldsymbol{\mathcal{R}}_\omega,\boldsymbol{\mathcal{C}}_\omega}\#\hat{\mathbb{P}}_\omega), \end{aligned} \tag{26}$$

where $*$ is the convolution operator, $\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{C}}$ defined in $\mathcal{X}$-space, and $\boldsymbol{\mathcal{R}}_\omega, \boldsymbol{\mathcal{C}}_\omega$ in $\mathcal{W}$. When $\mathbb{P}_\omega$ is known, the right term in (26) *is constant for all t and only needs to be computed only once.*

Remark 6 (Ambiguous noise)**.** *Although we consider both $\mathbb{P}_{x_0}$ and $\mathbb{P}_\omega$ are known in this section, the framework can be easily extended to case where one has uncertain $\mathbb{P}_{x_0} \in \mathbb{B}_{\theta_0}(\tilde{\mathbb{P}})$ and $\mathbb{P}_\omega \in \mathbb{B}_{\theta_\omega}(\tilde{\mathbb{T}})$, where $\theta_0, \theta_\omega > 0$, $\tilde{\mathbb{P}} \in \mathcal{P}_\rho(\mathcal{X})$, and $\tilde{\mathbb{T}} \in \mathcal{P}_\rho(\mathcal{W})$ are given. In this case, we note that $\mathbb{P}_{x_0} \times \mathbb{P}_\omega \in \mathbb{B}_{\theta_0+\theta_\omega}(\tilde{\mathbb{P}} \times \tilde{\mathbb{T}})$ and then we use Theorem 5.1 to bound the first time-step $\mathbb{W}_\rho(f\#(\mathbb{P}_{x_0} \times \mathbb{P}_\omega), f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#(\mathbb{P}_{x_0} \times \mathbb{P}_\omega))$.*

8. Experimental results

In this Section, we empirically evaluate the performance of our ρ-Wasserstein uncertainty propagation framework on various benchmarks taken from the literature[8]. We consider the following piecewise Lipschitz continuous functions f: a *Bounded Linear* f adapted from [36] with state space dimension d ranging from 1 to 4, an instance of the *Quadruple-Tank* from [24], the *Mountain Car* dynamics [37], and the *Dubins Car* [8]. Additionally, we consider the *Sigmoid* function introduced in Example 2, and a 10-dimensional *Neural Network layer*. In Section 8.4, we consider stochastic dynamical systems variants of the Mountain Car [37], and Quadruple-Tank [24] with additive Gaussian

[8] Our code is available at `https://github.com/sjladams/DUQviaWasserstein`

noise, and of a 3D-NN dynamics affected by non-Gaussian process noise. Additional details about the functions, dynamical systems, and probability distributions are available in the Appendix B.0.1.

In what follows, first, in Sections 8.1 and 8.2 we investigate the impact of the placement and the number of quantization locations $\mathcal{C}$ on the error bounds, respectively. Then, in Section 8.3, we analyze the effect of the linearization coefficients in Theorems 5.1 and 5.2 in case of non-linear functions f for different radii of uncertainty θ. Lastly, in Section 8.4, we apply the approximation scheme presented in Section 7 to stochastic dynamical systems. For all the experiments, we fix $\rho = 2$. All the experiments were conducted on an Intel Core i7-1365U CPU with 16GB of RAM using a single-core implementation.

		$\lvert\mathcal{C}\rvert$			
Dim. d	Algorithm	5	10	100	1000
1	Optimized grid	0.5085	0.2731	0.0280	0.0087
	Uniform grid	0.5420	0.3363	0.0487	0.0169
2	Optimized grid	0.7867	0.1935	0.0723	0.0248
	Uniform grid	0.7867	0.3826	0.1566	0.0539
3	Optimized grid	0.7940	0.1982	0.0818	0.0410
	Uniform grid	0.7940	0.5428	0.3532	0.1801
4	Optimized grid	1.8681	0.8043	0.4078	0.2111
	Uniform grid	1.8681	1.8465	0.7935	0.6161

TABLE 1. Comparison of error bounds from Theorem 5.2 for $\mathcal{R}, \mathcal{C}$ obtained as described in Remark 4 (*Optimized grid*) and the uniform partition (*Uniform grid*) for the Bounded Linear benchmark defined in the Appendix B.0.1.

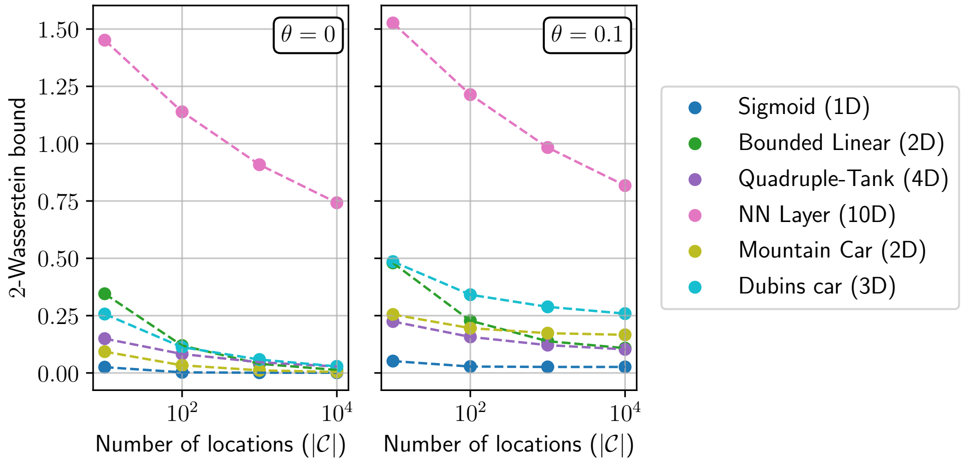

FIGURE 5. Upper bounds on $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_2(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$ for various benchmarks computed using Theorem 5.2 for $\theta = 0$ and Theorem 5.1 for $\theta = 0.1$.

8.1. **Improving on uniform grids of quantization locations**

In this Section, we analyze the effect of optimizing the locations $\mathcal{C}$ used for the quantization operator $\Delta_{\mathcal{R},\mathcal{C}}$ using the approach in Remark 4 compared to taking a uniform grid. In particular, in Table 1, for a bounded linear $f : \mathbb{R}^d \to \mathbb{R}^d$ defined in Appendix B.1 for each $d \in \{1, 2, 3, 4\}$, for different quantization sizes $|\mathcal{C}|$, we compare the error bound from Theorem 5.2, obtained using the the procedure described in Remark 4, with that obtained from a uniform partition of a subset $\tilde{\mathcal{X}} \subset \mathcal{X}$ containing most of the probability mass of $\mathbb{P}$[9]. From Table 1, we observe that as the dimensionality of the problem increases, the restrictiveness of placing locations in an equidistant fashion also augments. In fact, note that while for $d = 1$ the error bound in Theorem 5.2 is similar regardless of the heuristics used to place the locations, for $d = 4$ the selection performed by employing Remark 4 leads to bound 2-3 times smaller than the uniform partition approach.

8.2. **Error bound convergence**

In the previous Section, we focused on the placement of the quantization operator's locations. Here, we analyze how the 2-Wasserstein bounds decrease as the number of optimized locations for Theorems 5.1 and 5.2 grows. More precisely, given a distribution $\mathbb{P}$ and a ambiguity set $\mathbb{B}_\theta(\mathbb{P})$ of radius $\theta = 0$ or $\theta = 0.1$, we report the bound of $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_2(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$ for different quantization sizes $|\mathcal{C}|$.

From Figure 5, we observe that for all benchmarks increasing the number of locations in the quantization leads to a decreasing bound. This is expected due to the reduction of θ_d guaranteed by the discussion in Section 6.2. In the case where $\theta = 0$, as there is no uncertainty around $\mathbb{P}$, the bounds converge to zero. In contrast, with $\theta_d = 0.1$, the bounds do not converge to zero, but to different values for each system. Both observations empirically confirm Theorem 6.2. It is also important to note that the error bounds are impacted both by the geometry of the probability space of $\mathbb{P}$ as well as the system dynamics f. For instance, for the Dubins car, the upper bound on $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_2(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$ is consistently larger than that of the Quadruple-Tank, even though the Quadruple-Tank is higher dimensional. This can be explained because the Dubins car is not a stable system, and, consequently, the resulting uncertainty in terms of 2-Wasserstein distance is amplified.

8.3. **Analysis of ambiguity set propagation using global and local linearization**

We continue our analysis by investigating the impact of the linearization coefficients on our 2-Wasserstein bounds for different uncertainty radii θ. Specifically, we compare the bounds constructed using the trivial linearization coefficients $(\mathcal{L}_f, 0)$, with those derived from the coefficients described

[9] This uniform partition is defined as follows. We first get $\mathcal{C}$ from Remark 4. We then move the locations $c_k \in \mathcal{C}$ such that they are equally spaced in all axes (also forming a grid), obtaining $\mathcal{C}_{\text{unif}}$. Finally, we compute $\mathcal{R}_{\text{unif}}$ as the Voronoi partition w.r.t. $\mathcal{C}_{\text{unif}}$.

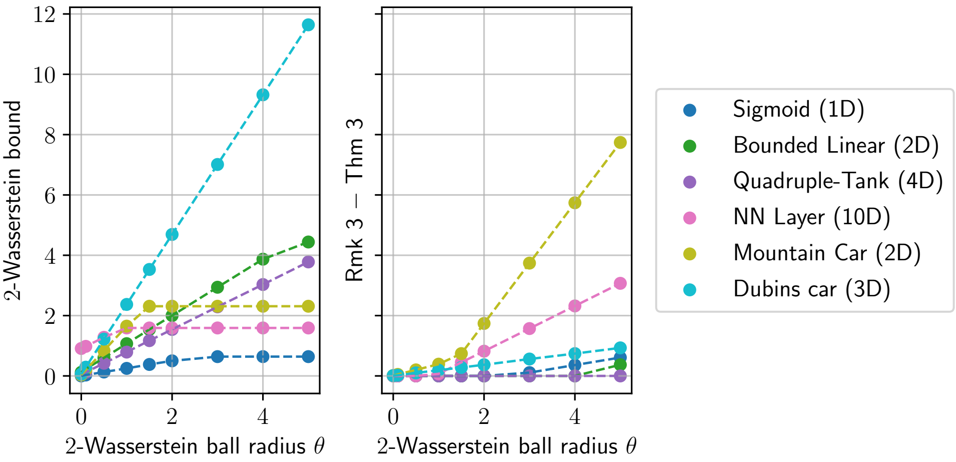

FIGURE 6. Analysis of the upper bounds on $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_2(f\#\mathbb{Q}, f\#\Delta_{\mathcal{R},\mathcal{C}}\#\mathbb{P})$ computed using Theorem 5.2 for $\theta = 0$ and Theorem 5.1 for $\theta > 0$. In the left plot are the absolute bounds using the linearization coefficients from Section 6.1; on the right, the absolute difference between the bounds using the more conservative global Lipschitz coefficients.

	NN Layer (3D)			**Mountain Car**			**Quadruple-Tank**		
t	**Emp.**	**Rmk 1**	**Thm 4**	**Emp.**	**Rmk 1**	**Thm 4**	**Emp.**	**Rmk 1**	**Thm 4**
1	0.0116	0.2020	0.1214	0.0256	0.0627	0.0547	0.0821	0.1517	0.1517
2	0.0090	0.2436	0.1358	0.0302	0.2364	0.1826	0.0790	0.2748	0.2748
3	0.0102	0.2732	0.1464	0.0498	0.6183	0.4178	0.0757	0.3670	0.3670
4	0.0102	0.2941	0.1522	0.0371	1.3944	0.8088	0.0680	0.4308	0.4308
5	0.0104	0.3097	0.1555	0.0413	2.9423	1.4388	0.0643	0.4751	0.4751
6	0.0105	0.3213	0.1574	0.0433	6.0399	2.4609	0.0621	0.5031	0.5031
7	0.0106	0.3301	0.1586	0.0407	12.2351	2.9560	0.0618	0.5185	0.5185
8	0.0102	0.3366	0.1593	0.0507	24.6256	2.9748	0.0659	0.5260	0.5260
9	0.0105	0.3414	0.1595	0.0505	49.4063	2.9910	0.0793	0.5242	0.5242
10	0.0099	0.3451	0.1598	0.0758	98.9695	3.0035	0.0769	0.5174	0.5174
50	0.0100	0.3562	0.1601	0.0676	1.1×10^{14}	3.1819	0.0767	0.4794	0.4794

TABLE 2. Formal bounds on $\mathbb{W}_2(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t})$ from Theorem 5.1 using the linearization coefficients described in Section 6.1, as shown in column *Thm 4*, or employing coefficients $(\mathcal{L}_f, 0)$, as in column *Rmk 1*. Column *Emp.* presents a Monte Carlo approximation of $\mathbb{W}_\rho(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t})$, calculated using 5×10^5 samples.

in Section 6.1, as per Theorem 5.1 and 5.2[10]. We set $|\mathcal{C}| = 10^2$ for functions with dimension of at most three, and use $|\mathcal{C}| = 10^3$ otherwise. The $\mathbb{R}^d$-partitions $\mathcal{R}$ and locations $\mathcal{C} \subset \mathbb{R}^d$ are selected as outlined in Remark 4.

[10] More specifically, we report $\mathcal{L}_f(\theta + \theta_d) - \left(\alpha_{\max}(\theta + \theta_d)^\rho + \sum_{k=1}^N \pi^{(k)}\beta_k\right)^{\frac{1}{\rho}}$ for $\theta > 0$, and $\mathcal{L}_f(\theta + \theta_d) - \left(\sum_{k=1}^N \alpha_k \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) + \sum_{k=1}^N \pi^{(k)}\beta_k\right)^{\frac{1}{\rho}}$ for $\theta = 0$.

The left plot of Figure 6 shows that for the optimized coefficients in case of bounded functions (NN Layer, Mountain Car and Bounded Linear), the bounds saturate from a certain θ onwards. This saturation occurs because, for large θ, we select $(\alpha_k, \beta_k) = (0, \sup_{x \in \mathcal{X}} \|f(x) - f(c_k)\|^\rho)$ for most regions, as explained in Section 6.1. Consequently, the error bound from Theorem 5.1 becomes independent of θ. Furthermore, it is important to note that in many cases, the error bounds are smaller than θ, which indicates a contraction of the ambiguity set. An exception is the Dubins car example, where instability in the system dynamics causes the ambiguity set to expand.

The right plot of Figure 6 confirms that, as discussed in Remark 3, for nonlinear systems, the bounds constructed using the optimized coefficients are consistently and substantially tighter that the bound resulting from using the global Lipschitz coefficients. Note that for linear systems (Quadruple-Tank), the two coefficients are equivalent and lead to the same linearizations in (9) and (14).

8.4. **Uncertainty Propagation in Stochastic Dynamical Systems**

In this Section, we apply the discrete approximation scheme presented in Section 7 and illustrated in Figure 4 to three stochastic dynamical systems. We analyze both an empirical estimation of, and our formal bounds on, $\mathbb{W}_2(\mathbb{P}_{x_t}, \hat{\mathbb{P}}_{x_t})$, where $\mathbb{P}_{x_t}$ represents the true unknown state distribution at time t and $\hat{\mathbb{P}}_{x_t}$ is our discrete approximator. In Table 2, we observe that the empirical ρ-Wasserstein distance remains low over longer time horizons, demonstrating the effectiveness of the approximation in practical scenarios. For the contracting NN Layer and Quadruple-Tank dynamics, the Monte Carlo estimates of the approximation error converge to fixed points, supporting Theorem 7.1. For the non-contracting ($\mathcal{L}_f > 1$) but bounded Mountain Car dynamics, the bounds from Theorem 5.1 obtained using coefficients $(\mathcal{L}_f, 0)$ quickly explode. In contrast, using Theorem 5.1 results in bounds that converge to a fixed point due to the boundedness of the dynamics. From Figure 7, we can visually confirm that our discrete approximators (right column) closely match an empirical estimate of true distributions (left column). We highlight that the discrete approximator is able to capture the fact that the state distribution becomes bimodal at $t = 10$. Such characteristics are challenging to identify using techniques like moment matching [14], for instance, which only focus on approximating, commonly with no guarantees, the first few moments of the distribution.

9. **Conclusion and future direction**

We introduced a novel framework to approximate the push-forward measure of uncertain distributions with discrete distributions with formal quantification of the resulting uncertainty in terms of ρ-Wasserstein distance, allowing for a tractable propagation of ρ-Wasserstein ambiguity sets. We see at least three interesting future research directions. First, the development of efficient ways to compute the norm approximations in (9). Further, in the context of multi-step propagation of ambiguity sets, such as for dynamical systems, it may be of interest to directly rely on properties of the compositions of $f \circ ... \circ f$, instead of the sequential application of our framework, as we propose in Section 7. Lastly, we indicate that this framework could be directly applied as the prediction mechanism for distributionally-robust non-linear Model Predictive Control.

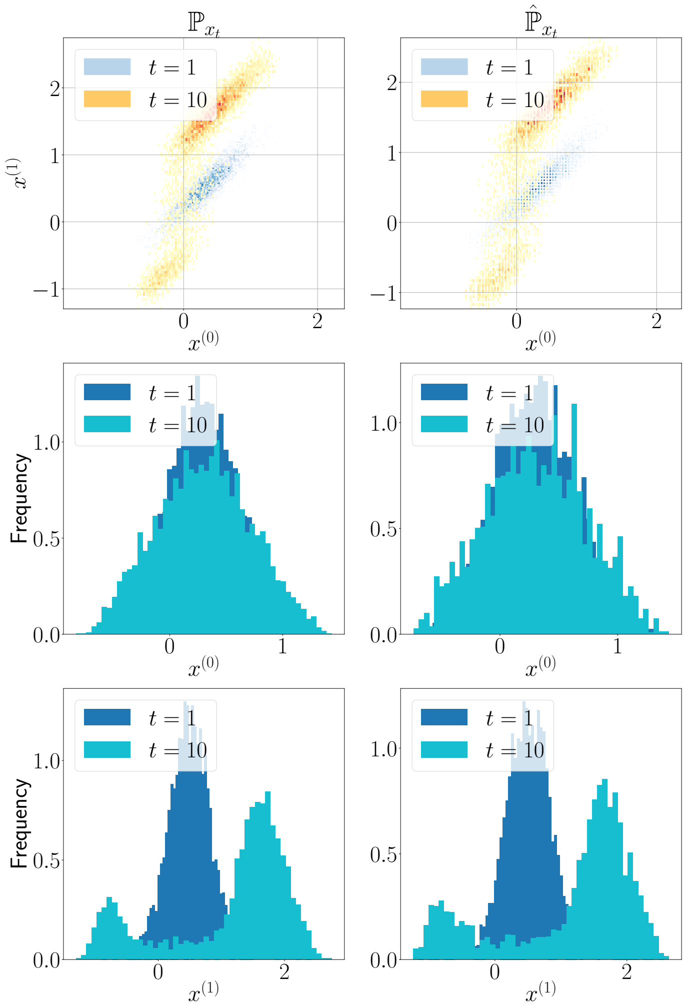

FIGURE 7. Monte Carlo simulation of the true state distribution (left plots) - with 5×10^3 samples - and our discrete approximation from Section 7 (right plots) - with $|\mathcal{C}| = 100$ - for the Mountain Car system from $t = 1$ to $t = 10$. The upper plots display the joint distribution of the first two state dimensions, $x_t^{(1)}$ and $x_t^{(2)}$, for all time steps, and the lower plots illustrate the initial and final marginal distributions.

Appendix A. Proofs

In this section, we present the proofs for all the results discussed in the paper's main text.

A.1. Proof of Proposition 3.1

Before proving Proposition 3.1, we prove an auxiliary Lemma.

Lemma A.1. *For $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$. Further, let $\boldsymbol{\mathcal{R}}$ be a $\mathcal{X}$-partition and $\boldsymbol{\mathcal{C}}$ a set of locations. Then, for $\gamma^* \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ defined as*

$$d\gamma^*(x_1, x_2) := \sum_{i=1}^{N} \mathbb{1}_{\mathcal{R}_i}(x_1) d\mathbb{P}(x_1) d\delta_{c_i}(x_2) \tag{27}$$

it holds that:

i) γ^* *is a valid coupling between* $\mathbb{P}$ *and* $\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}$, *i.e.* $\gamma^* \in \Gamma(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$,

ii) if $\bar{\boldsymbol{\mathcal{R}}}$ *is the Voronoi partition w.r.t.* $\boldsymbol{\mathcal{C}}$ *then*

$$\gamma^* = \underset{\gamma\in\Gamma(\mathbb{P},\Delta_{\bar{\boldsymbol{\mathcal{R}}},\mathcal{C}}\#\mathbb{P})}{\operatorname{arginf}} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^\rho \, d\gamma(x,y).$$

Proof. We start by proving that $\gamma^* \in \Gamma(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$. For any $A, B \in \mathcal{B}(\mathcal{X})$, we have:

$$\begin{aligned}\gamma^*(A,B) &= \int_A\int_B d\gamma^*(x_1,x_2)\\ &= \int_A\int_B \sum_{i=1}^N \mathbb{1}_{\mathcal{R}_i}(x_1)d\mathbb{P}(x_1)d\delta_{c_i}(x_2)\\ &= \sum_{i=1}^N \mathbb{P}(A\cap\mathcal{R}_i)\mathbb{1}_B(c_i),\end{aligned}$$

which is a value in $[0,1]$ since $\boldsymbol{\mathcal{R}}$ is a partition of $\mathcal{X}$. Further, note that:

$$\gamma(A,\mathcal{X}) = \sum_{i=1}^N \mathbb{P}(A\cap\mathcal{R}_i)\mathbb{1}_{\mathcal{X}}(c_i) = \sum_{i=1}^N \mathbb{P}(A\cap\mathcal{R}_i) = \mathbb{P}(A)$$

and

$$\gamma(\mathcal{X},B) = \sum_{i=1}^N \mathbb{P}(\mathcal{X}\cap\mathcal{R}_i)\mathbb{1}_B(c_i) = \sum_{i=1}^N \mathbb{P}(\mathcal{R}_i)\mathbb{1}_B(c_i) = \big(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}\big)(B)$$

and, consequently, $\gamma(\mathcal{X},\mathcal{X}) = 1$. Thus, $\gamma \in \Gamma(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$. This proves item i). To prove item ii), it suffices to note that if $x \in \mathcal{R}_i$, then by the definition of the Voronoi partition w.r.t. $\boldsymbol{\mathcal{C}}$, the cost of transporting $d\mathbb{P}(x)$ to c_i is smaller than any other $c_j, j \neq i$ since $\|x-c_i\| \leq \|x-c_j\|$. □

We are now ready to prove Proposition 3.1. Let γ^* be defined as in (27). Using item i) from Lemma A.1, we have:

$$\begin{aligned}\mathbb{W}_\rho(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})^\rho &\leq \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^\rho \, d\gamma^*(x,y) \qquad (28)\\ &= \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^\rho \sum_{k=1}^N \mathbb{1}_{\mathcal{R}_k}(x)d\mathbb{P}(x)d\delta_{c_k}(y)\\ &= \sum_{k=1}^N \int_{\mathcal{R}_k} \|x-c_k\|^\rho \, d\mathbb{P}(x)\end{aligned}$$

If $\bar{\boldsymbol{\mathcal{R}}}$ is the Voronoi partition w.r.t. $\boldsymbol{\mathcal{C}}$, by applying item ii) from Lemma A.1, the inequality in (28) becomes an equality. □

A.2. **Proof of Theorem 5.1**

Before proving Theorem 5.1, we show that $\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})$ can be upper bounded by a one-dimensional minimization program, using duality techniques from the DRO literature [18, 31, 42].

Proposition A.2. *For $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$, $\boldsymbol{\mathcal{R}}$ be a $\mathcal{X}$-partition, and $\boldsymbol{\mathcal{C}}$ be a set of locations. Further, denote $\theta_d := \Big(\sum_{k=1}^N \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x)\Big)^{\frac{1}{\rho}}$, and call $\pi^{(i)} := \mathbb{P}(\mathcal{R}_i)$ for every $\mathcal{R}_i \in \boldsymbol{\mathcal{R}}$. Then,*

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \\
&\quad\le \left(\inf_{\lambda\ge 0} \left\{ \lambda(\theta+\theta_d)^\rho + \sum_{j=1}^N \pi^{(j)} \sup_{\xi\in\mathcal{X}} \big(\|f(\xi) - f(c_j)\|^\rho - \lambda \|\xi - c_j\|^\rho \big) \right\} \right)^{\frac{1}{\rho}} \qquad (29)
\end{aligned}$$

Proof. We define $S_\theta(\mathbb{P})$ as a subspace of $\mathcal{P}(\mathcal{X}\times\mathcal{X})$ containing all the couplings for which one of the marginals is $\mathbb{P}$ and the other implied marginal is at most θ far in ρ-Wasserstein distance from $\mathbb{P}$, i.e.

$$S_\theta(\mathbb{P}) := \Big\{ \gamma \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) : \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_2\|^\rho \, d\gamma(x_1,x_2) \le \theta^\rho, \mathrm{proj}_2\#\gamma = \mathbb{P} \Big\},$$

where $\mathrm{proj}_2\#\gamma$ returns the marginal distribution of γ in the second component, i.e. $\mathrm{proj}_2\#\gamma := \int_{\mathcal{X}} \gamma(dx_1, .)$. We then have:

$$\begin{aligned}
&\Big(\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \Big)^\rho \\
&\qquad \text{(By monotonicity of } x^\rho \text{ for } x \ge 0) \\
&= \sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})^\rho \\
&\qquad \text{(Using Eqn (3))} \\
&= \sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \inf_{\gamma\in\Gamma(\mathbb{Q},\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \|f(x_1) - f(x_2)\|^\rho \, d\gamma(x_1,x_2) \\
&\text{(As } \mathbb{B}_\theta(\mathbb{P}) \subseteq \mathbb{B}_{\theta+\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \text{ for } \mathbb{W}_\rho(\mathbb{P}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \le \theta_d) \\
&\le \sup_{\mathbb{Q}\in\mathbb{B}_{\theta+\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \inf_{\gamma\in\Gamma(\mathbb{Q},\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \|f(x_1) - f(x_2)\|^\rho \, d\gamma(x_1,x_2) \\
&\qquad \text{(By the fact that } \Gamma(\mathbb{Q}, \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \subseteq S_{\theta+\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})) \\
&\le \sup_{\gamma\in S_{\theta+\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \|f(x_1) - f(x_2)\|^\rho \, d\gamma(x_1,x_2) \qquad (30)
\end{aligned}$$

Applying Lagrangian duality to (30):

$$\begin{aligned}
&\sup_{\gamma \in S_{\theta+\theta_d}(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \|f(x_1)-f(x_2)\|^\rho \, d\gamma(x_1,x_2) \\
&\qquad \text{(By Lagrangian strong duality)} \\
&= \inf_{\lambda\geq 0} \sup_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{X})} \Bigg\{ \int_{\mathcal{X}\times\mathcal{X}} \|f(x_1)-f(x_2)\|^\rho \, d\gamma(x_1,x_2) + \\
&\lambda\Bigg((\theta+\theta_d)^\rho - \int_{\mathcal{X}\times\mathcal{X}} \|x_1-x_2\|^\rho \, d\gamma(x_1,x_2)\Bigg) : \mathrm{proj}_2\#\gamma = \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}\Bigg\} \qquad (31) \\
&\qquad \text{(Reorganizing the terms)} \\
&= \inf_{\lambda\geq 0} \sup_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{X})} \Bigg\{ \lambda(\theta+\theta_d)^\rho + \\
&\int_{\mathcal{X}\times\mathcal{X}} \big(\|f(x_1)-f(x_2)\|^\rho - \lambda\|x_1-x_2\|^\rho \big) d\gamma(x_1,x_2) : \mathrm{proj}_2\#\gamma = \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}\Bigg\} \\
&\qquad \text{(By applying Theorem 1 from [18])} \\
&= \inf_{\lambda\geq 0} \Bigg\{ \lambda(\theta+\theta_d)^\rho + \int_{\mathcal{X}} \sup_{\xi\in\mathcal{X}} \big(\|f(\xi)-f(\varsigma)\|^\rho - \lambda\|\xi-\varsigma\|^\rho \big) d\big(\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}\big)(\varsigma) \Bigg\} \qquad (32) \\
&\qquad \text{(By using the definition of } \Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \\
&= \inf_{\lambda\geq 0} \Bigg\{ \lambda(\theta+\theta_d)^\rho + \sum_{j=1}^N \pi^{(j)} \sup_{\xi\in\mathcal{X}} \big(\|f(\xi)-f(c_j)\|^\rho - \lambda\|\xi-c_j\|^\rho \big) \Bigg\}
\end{aligned}$$

□

We are now ready to prove Theorem 5.1. By Proposition A.2,

$$\begin{aligned}
&\left(\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \right)^\rho \\
&\leq \inf_{\lambda\geq 0} \Bigg\{ \lambda(\theta+\theta_d)^\rho + \sum_{j=1}^N \pi^{(j)} \sup_{\xi\in\mathcal{X}} \big(\|f(\xi)-f(c_j)\|^\rho - \lambda\|\xi-c_j\|^\rho \big) \Bigg\} \\
&\qquad \text{(By the norm linearization in (9))} \\
&\leq \inf_{\lambda\geq 0} \Bigg\{ \lambda(\theta+\theta_d)^\rho + \sum_{j=1}^N \pi^{(j)} \sup_{\xi\in\mathcal{X}} \big(\alpha_j\|\xi-c_j\|^\rho + \beta_j - \lambda\|\xi-c_j\|^\rho \big) \Bigg\} \\
&= \inf_{\lambda\geq 0} \Bigg\{ \lambda(\theta+\theta_d)^\rho + \sum_{j=1}^N \pi^{(j)}\beta_j + \sum_{j=1}^N \pi^{(j)} \sup_{\xi\in\mathcal{X}} \big((\alpha_j-\lambda)\|\xi-c_j\|^\rho \big) \Bigg\} \qquad (33)
\end{aligned}$$

First, consider the case where $\mathcal{X}$ is unbounded (e.g. $\mathcal{X}=\mathbb{R}^d$). If there exists α_ℓ such that $\alpha_\ell > \lambda$, then the correspondent inner supremum returns ∞. Thus, in the outer minimization program, it is enough to search for $\lambda \geq \max_{j\in\{1,\ldots,N\}} \alpha_j$. Moreover, we note that for any $\lambda \geq \max_{j\in\{1,\ldots,N\}} \alpha_j$, the inner supremum returns 0. Hence, the solution of the whole optimization program is given by

$\lambda^* = \max_{j\in\{1,\dots,N\}} \alpha_j$, so that (33) becomes $\lambda^*(\theta+\theta_q)^\rho + \sum_{j=1}^N \pi^{(j)}\beta_j$. For bounded $\mathcal{X}$, one may find a solution $\lambda^* < \max_{j\in\{1,\dots,N\}} \alpha_j$ for the entire optimization program. However, we remark that choosing $\tilde{\lambda} = \max_{j\in\{1,\dots,N\}} \alpha_j$ still provides a valid upper bound on (33). □

A.3. **Proof of Theorem 5.2**

Let γ^* be defined as in (27). Then, from statement i) from Lemma A.1, we have:

$$\begin{aligned}
\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P})^\rho &\le \int_{\mathcal{X}\times\mathcal{X}} \|f(x)-f(y)\|^\rho \, d\gamma^*(x,y) \qquad (34)\\
&= \int_{\mathcal{X}\times\mathcal{X}} \|f(x)-f(y)\|^\rho \sum_{k=1}^N \mathbb{1}_{\mathcal{R}_k}(x) d\mathbb{P}(x) d\delta_{c_k}(y)\\
&= \sum_{k=1}^N \int_{\mathcal{R}_k} \|f(x)-f(c_k)\|^\rho \, d\mathbb{P}(x)\\
&\quad \text{(By the norm linearization in (14)))}\\
&\le \sum_{k=1}^N \int_{\mathcal{R}_k} \big(\alpha_k \|x-c_k\|^\rho + \beta_k\big) d\mathbb{P}(x)\\
&= \sum_{k=1}^N \alpha_k \int_{\mathcal{R}_k} \|x-c_k\|^\rho \, d\mathbb{P}(x) + \sum_{k=1}^N \beta_k \mathbb{P}(\mathcal{R}_k)
\end{aligned}$$

In the case where $\boldsymbol{\mathcal{R}}$ is the Voronoi partition w.r.t. $\boldsymbol{\mathcal{C}}$, by item ii) of Lemma A.1, the inequality in (34) can be replaced by equality. The rest of the proof remains the same. □

A.4. **Proof of Lemma 6.1**

By straightforward applications of the triangle inequality:

$$\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) \le \sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P}) + \mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}),$$

$$\sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\mathbb{P}) \le \sup_{\mathbb{Q}\in\mathbb{B}_\theta(\mathbb{P})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}) + \mathbb{W}_\rho(f\#\Delta_{\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{C}}}\#\mathbb{P}, f\#\mathbb{P}).$$

We conclude by combining both inequalities. □

A.5. **Proof of Theorem 6.2**

From Lemma 6.1, to prove this theorem, it is enough to show that $\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}}^*,\boldsymbol{\mathcal{C}}^*}\#\mathbb{P}) \leq \epsilon$. From Theorem 5.2, by taking $(\alpha_k, \beta_k) = (\mathcal{L}_f, 0)$ as discussed in Remark 3, we have that:

$$\begin{aligned}\mathbb{W}_\rho(f\#\mathbb{P}, f\#\Delta_{\boldsymbol{\mathcal{R}}^*,\boldsymbol{\mathcal{C}}^*}\#\mathbb{P})^\rho \leq \mathcal{L}_f^\rho \sum_{k=1}^{N+1} \int_{\mathcal{R}_k^*} \|x - c_k^*\|^\rho \, d\mathbb{P}(x)&\\ &= \mathcal{L}_f^\rho \sum_{k=1}^{N} \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) + \mathcal{L}_f^\rho \int_{\mathcal{X}\setminus\bar{\mathcal{X}}} \|x - \bar{c}\|^\rho \, d\mathbb{P}(x)\\ &\leq \mathcal{L}_f^\rho \sum_{k=1}^{N} \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x) + \frac{\epsilon^\rho}{2} \qquad (35)\end{aligned}$$

where we use the fact that, by construction, $\boldsymbol{\mathcal{R}}^* := \boldsymbol{\mathcal{R}} \cup \{\mathcal{X} \setminus \bar{\mathcal{X}}\}$ and $\boldsymbol{\mathcal{C}}^* := \boldsymbol{\mathcal{C}} \cup \{\bar{c}\}$, and also $\int_{\mathcal{X}\setminus\bar{\mathcal{X}}} \|x - \bar{c}\|^\rho \, d\mathbb{P}(x) \leq \frac{\epsilon^\rho}{2\mathcal{L}_f^\rho}$ (which, we must highlight, is always possible as $\mathbb{P} \in \mathcal{P}_\rho(\mathcal{X})$). Then, what is left to show is that the left term in (35) can also be upper-bounded by $\frac{\epsilon^\rho}{2}$. Indeed, because $\|R_k\|_\infty = \frac{\|\bar{\mathcal{X}}\|_\infty}{N^{\frac{1}{d}}}$ (as all compact regions are hypercubic), it holds that: $\|R_k\|_\infty = \frac{\|\bar{\mathcal{X}}\|_\infty}{N^{\frac{1}{d}}} \leq \frac{\epsilon}{2^{\frac{1}{\rho}} d^{\frac{1}{\rho}} \mathcal{L}_f}$, where we use the fact that again by construction, $N \geq \left(\frac{2^{\frac{1}{\rho}} \mathcal{L}_f d^{\frac{1}{\rho}} \|\bar{\mathcal{X}}\|_\infty}{\epsilon}\right)^d$. Thus,

$$\mathcal{L}_f^\rho \sum_{k=1}^{N} \int_{\mathcal{R}_k} \|x - c_k\|^\rho \, d\mathbb{P}(x)$$

(From the L_ρ-norm definition)

$$= \mathcal{L}_f^\rho \sum_{k=1}^{N} \int_{\mathcal{R}_k} \sum_{i=1}^{d} |x^{(i)} - c_k^{(i)}|^\rho d\mathbb{P}(x)$$

(From the $\|.\|_\infty$ definition)

$$\leq \mathcal{L}_f^\rho \sum_{k=1}^{N} \int_{\mathcal{R}_k} \sum_{i=1}^{d} \|\mathcal{R}_k\|_\infty^\rho \, d\mathbb{P}(x)$$

(Using that $\|R_k\|_\infty \leq \frac{\epsilon}{2^{\frac{1}{\rho}} d^{\frac{1}{\rho}} \mathcal{L}_f}$)

$$\leq \mathcal{L}_f^\rho \sum_{k=1}^{N} \int_{\mathcal{R}_k} \sum_{i=1}^{d} \frac{\epsilon^\rho}{2d\mathcal{L}_f^\rho} d\mathbb{P}(x) = \mathcal{L}_f^\rho \sum_{k=1}^{N} \frac{\epsilon^\rho}{2\mathcal{L}_f^\rho} \mathbb{P}(\mathcal{R}_k) = \frac{\epsilon^\rho}{2} \sum_{k=1}^{N} \mathbb{P}(\mathcal{R}_k) \leq \frac{\epsilon^\rho}{2}.$$

□

A.6. **Proof of Theorem 7.1**

We use the same notation as in Figure 4. The proof follows by induction. The base case is $t = 1$, for which we have

$$\mathbb{W}_\rho(\mathbb{P}_{x_1}, \hat{\mathbb{P}}_{x_1}) = \mathbb{W}_\rho(f\#\mathbb{P}_0, f\#\Delta_{\boldsymbol{\mathcal{R}}_0,\boldsymbol{\mathcal{C}}_0}\#\hat{\mathbb{P}}_0) = \mathbb{W}_\rho(f\#\hat{\mathbb{P}}_0, f\#\Delta_{\boldsymbol{\mathcal{R}}_0,\boldsymbol{\mathcal{C}}_0}\#\hat{\mathbb{P}}_0)$$

			Section					
			7.1	7.4				
System	d	$f(x)$	$\mathbb{P}$	$f(x,\omega)$	$\mathbb{P}_{x_0}$	$\mathbb{P}_{\omega}$	$\lvert\boldsymbol{\mathcal{C}}\rvert$	T
Sigmoid	1	f_{Sigm}	$\mathcal{N}(0.2, 0.5)$					
Bounded Linear	2	f_{BoundLin}	$\mathcal{N}\left(\begin{bmatrix}1.5\\2.5\end{bmatrix}, \begin{bmatrix}0.4 & 0.0\\0.0 & 0.5\end{bmatrix}\right)$					
Quadruple-Tank	4	f_{QuadTank}	$\mathcal{N}\left(\begin{bmatrix}1.5\\2.5\\-0.5\\-1.0\end{bmatrix}, \begin{bmatrix}0.001 & 0 & 0 & 0\\0 & 0.02 & 0 & 0\\0 & 0 & 0.4 & 0\\0 & 0 & 0 & 0.01\end{bmatrix}\right)$					
NN Layer	10	f_{NNLay}	$\bar{\mathbb{P}}$	$\sigma(Ax + B\omega)$	$\bar{\mathbb{P}}_{x_0}$	$\bar{\mathbb{P}}_{\omega}$	10^2	50
Mountain Car	2	f_{MountCar}	$\mathcal{N}\left(\begin{bmatrix}0.3\\0.2\end{bmatrix}, \begin{bmatrix}10^{-1} & 0\\0 & 10^{-3}\end{bmatrix}\right)$	$f(x) + \omega$	$\mathbb{P}$ (from 7.1)	$\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, 10^{-2}I\right)$	10^2	50
Dubins Car	3	$f_{\text{DubinsCar}}$	$\mathcal{N}\left(\begin{bmatrix}0.3\\0.2\\0.01\end{bmatrix}, \begin{bmatrix}10^{-1} & 0 & 0\\0 & 10^{-2} & 0\\0 & 0 & 10^{-3}\end{bmatrix}\right)$	$f(x) + \omega$	$\mathbb{P}$ (from 7.1)	$\mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, 10^{-2}I\right)$	10^2	50

Table 3. Summary of implementation details

since $\mathbb{P}_0 = \hat{\mathbb{P}}_0$ (as $\mathbb{P}_{x_0} = \hat{\mathbb{P}}_{x_0}$). Thus, the bound θ_1 comes from the application of Theorem 5.2. For the induction case (i.e., $t > 1$) we have:

$$\mathbb{W}_\rho(\mathbb{P}_{x_{t+1}}, \hat{\mathbb{P}}_{x_{t+1}}) = \mathbb{W}_\rho(f\#\mathbb{P}_t, f\#\Delta_{\boldsymbol{\mathcal{R}}_t,\boldsymbol{\mathcal{C}}_t}\#\hat{\mathbb{P}}_t) \leq \sup_{\mathbb{Q}\in\mathbb{B}_{\theta_t}(\hat{\mathbb{P}}_{x_t})} \mathbb{W}_\rho(f\#\mathbb{Q}, f\#\Delta_{\boldsymbol{\mathcal{R}}_t,\boldsymbol{\mathcal{C}}_t}\#\hat{\mathbb{P}}_t),$$

from which the bound θ_{t+1} follows from applying Theorem 5.1 for $\theta = \theta_t$, and using that $\theta_{d,t} \leq \epsilon$. This proves statement i) in the Theorem. For statement ii), we first note that by Remark 3, for $t > 1$:

$$\mathbb{W}_\rho(\mathbb{P}_{x_{t+1}}, \hat{\mathbb{P}}_{x_{t+1}}) \leq \theta_{t+1} \leq \left(\alpha_{\max,t}(\theta_t + \epsilon)^\rho + \sum_{k=1}^{N_t} \mathbb{P}(\mathcal{R}_{t,k})\beta_{t,k}\right)^{\frac{1}{\rho}} \leq \mathcal{L}_f(\theta_t + \epsilon)$$

Let $T : \mathbb{R} \to \mathbb{R}$ be a map given by $T(\theta) \coloneqq \mathcal{L}_f(\theta + \epsilon)$. Note that T is contractive since $|T(\theta) - T(\tilde{\theta})| \leq |\mathcal{L}_f(\theta - \tilde{\theta})| \leq \mathcal{L}_f|\theta - \tilde{\theta}|$. One can easily find a fixed point θ^* for T, i.e.

$$\theta^* = T(\theta^*) \iff \theta^* = \mathcal{L}_f(\theta^* + \epsilon) \iff \theta^* = \frac{\mathcal{L}_f}{1 - \mathcal{L}_f}\epsilon$$

Then, by the Banach fixed-point theorem, for the sequence $\theta_{t+1} = T(\theta_t)$, it holds that $\lim_{t\to\infty} \theta_t = \theta^*$, which concludes the proof. □

Appendix B. Implementation Details

In the following, we present the implementation details of the experiments in Section 8. First, we introduce the piecewise Lipschitz continuous functions f that we consider. Then, in Table 3, we show the probability distributions used in the experiments.

B.0.1. *Functions.*

Sigmoid (Example 2). $f_{\text{Sigm}} = \frac{1}{1+e^{-x}}$.

Bounded Linear (adapted from [36]). $f_{\text{BoundLin}} : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$f_{\text{BoundLin}}(x) = \text{clamp}\left(\begin{bmatrix} 0.0 & 0.4 \\ 0.3 & 0.8 \end{bmatrix} x, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix} \right). \tag{36}$$

Quadruple-Tank (instance of [24]). $f_{\text{QuadTank}} : \mathbb{R}^4 \to \mathbb{R}^4$ such that:

$$f_{\text{QuadTank}}(x) = \begin{bmatrix} 0.721 & 0 & 0.041 & 0 \\ 0 & 0.718 & 0 & 0.033 \\ 0 & 0 & 0.724 & 0 \\ 0 & 0 & 0 & 0.737 \end{bmatrix} x \tag{37}$$

NN Layer. $f_{\text{NNLay}} : \mathbb{R}^{10} \to \mathbb{R}^{10}$ such that $f_{\text{NNLay}}(x) = \sigma(Ax)$, where $A = \text{diag}(3 \times 10^0, 10^{-3}, 5 \times 10^{-3}, 7 \times 10^{-3}, 3 \times 10^{-2}, 10^{-3}, 10^{-3}, 10^{-3}, 10^{-3}, 10^{-3})$.

Mountain Car (adapted from [37]). $f_{\text{MountCar}} : \mathbb{R}^2 \to \mathbb{R}^2$ such that:

$$f_{\text{MountCar}}(x) = \text{clamp}\left(\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} x + \begin{bmatrix} 10^{-3} \\ 0 \end{bmatrix}, \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 1.2 \\ 1.2 \end{bmatrix} \right) - 2.5 \times 10^{-3} \begin{bmatrix} \cos\left(3x^{(2)}\right) \\ 0 \end{bmatrix} \tag{38}$$

Dubins Car [8]. $f_{\text{DubinsCar}} : \mathbb{R}^3 \to \mathbb{R}^3$ such that

$$f_{\text{DubinsCar}}(x) = \begin{bmatrix} x^{(1)} + 1.5 \sin\left(x^{(3)}\right) \\ x^{(2)} + 1.5 \cos\left(x^{(3)}\right) \\ x^{(3)} + 0.6 \end{bmatrix}. \tag{39}$$

B.1. **Further details**

In Section 8.2, for the NN Layer function, we consider (see Table 3) $\bar{\mathbb{P}}_{x_0} = \mathcal{N}(\mu_{NN}, \Sigma_{NN})$, with:

$$\mu_{NN} = [0.0, 1.0, 0.5, -0.7, 0.3, 2.0, -3.0, 0.4, -0.1, 4.0]^T,$$

and

$$\Sigma_{NN} = \text{diag}([0.0001, 0.5, 0.7, 0.2, 1.5, 2.5, 0.1, 0.5, 0.8, 0.2]).$$

Alternatively, in Section 8.4, we consider a 3D NN Layer, where (using the same notation as in Table 3) $A = \text{diag}([3.0, 1.5, 1.2]), B = \text{diag}([0.5, 1.0, 0.9])$,

$$\bar{\mathbb{P}}_{x_0} = \mathcal{N}\left(\begin{bmatrix} 1.5 \\ -1.2 \\ 2.4 \end{bmatrix}, \begin{bmatrix} 10^{-1} & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} \right), \qquad \bar{\mathbb{P}}_{\omega} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10^{-1} & 0 & 0 \\ 0 & 10^{-1} & 0 \\ 0 & 0 & 10^{-2} \end{bmatrix} \right).$$

Finally, in Section 8.1, we consider the following functions and distributions, for $d \in \{1, 2, 3, 4\}$: For $d = 1$, $f(x) = \text{clamp}(3x, -1, 1), \mathbb{P} = \mathcal{N}(0, 1)$. For $d = 2$, $f(x) = \text{clamp}(\text{diag}([3, 0.001])x, -2, 2)$,

$$\mathbb{P} = \mathcal{N}\left(\begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.02 & 0 \\ 0 & 0.5 \end{bmatrix} \right).$$

For $d = 3,\ f(x) = \mathrm{clamp}(\mathrm{diag}([3, 0.001, 1.1])x, -2, 2)$,

$$\mathbb{P} = \mathcal{N}\left(\begin{bmatrix} 3 \\ 1 \\ -0.9 \end{bmatrix}, \begin{bmatrix} 0.02 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.001 \end{bmatrix}\right).$$

For $d = 4,\ f(x) = \mathrm{clamp}(\mathrm{diag}([3, 0.001, 1.1, 2.2])x, -2, 2)$,

$$\mathbb{P} = \mathcal{N}\left(\begin{bmatrix} 3 \\ 1 \\ -0.9 \\ 0.4 \end{bmatrix}, \begin{bmatrix} 0.02 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.001 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix}\right).$$

References

[1] Alessandro Abate, Joost-Pieter Katoen, John Lygeros, and Maria Prandini. Approximate model checking of stochastic hybrid systems. *European Journal of Control*, 16(6):624–641, 2010.

[2] Steven Adams, Andrea Patane, Morteza Lahijanian, and Luca Laurenti. Finite neural networks as mixtures of gaussian processes: From provable error bounds to prior selection. *arXiv preprint arXiv:2407.18707*, 2024.

[3] Daniel Alspach and Harold Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.

[4] Luca Ambrogioni, Umut Guclu, and Marcel van Gerven. Wasserstein variational gradient descent: From semi-discrete optimal transport to ensemble variational inference. *arXiv preprint arXiv:1811.02827*, 2018.

[5] Liviu Aolaritei, Marta Fochesato, John Lygeros, and Florian Dörfler. Wasserstein tube mpc with exact uncertainty propagation. In *IEEE Conference on Decision and Control (CDC)*, pages 2036–2041, 2023.

[6] Liviu Aolaritei, Nicolas Lanzetti, Hongruyu Chen, and Florian Dörfler. Distributional uncertainty propagation via optimal transport. *arXiv preprint arXiv:2205.00343*, 2022.

[7] Ludwig Arnold, Christopher Jones, Konstantin Mischaikow, Geneviève Raugel, and Ludwig Arnold. *Random Dynamical Systems*. Springer, 1995.

[8] Devin Balkcom, Andrei Furtuna, and Weifu Wang. The dubins car and other arm-like mobile robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 380–386, 2018.

[9] Aharon Ben-Tal, Arkadi Nemirovski, and Laurent El Ghaoui. *Robust Optimization*. Princeton University Press, 2009.

[10] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

[11] Richard Bucy and Kenneth Senne. Digital synthesis of non-linear filters. *Automatica*, 7(3):287–298, 1971.

[12] Giuseppe Carlo Calafiore and L El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130:1–22, 2006.

[13] Richard Cheng, Richard Murray, and Joel Burdick. Limits of probabilistic safety guarantees when considering human uncertainty. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3182–3189, 2021.

[14] Marc Deisenroth and Carl Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *International Conference on machine learning (ICML)*, pages 465–472, 2011.

[15] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[16] Oliver Ernst, Alois Pichler, and Björn Sprungk. Wasserstein sensitivity of risk and uncertainty propagation. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):915–948, 2022.

[17] Eduardo Figueiredo, Andrea Patane, Morteza Lahijanian, and Luca Laurenti. Uncertainty propagation in stochastic systems via mixture models with error quantification. *IEEE Conference on Decision and Control (CDC)*, 2024.

[18] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

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

[20] Agathe Girard, Carl Rasmussen, Joaquin Candela, and Roderick Murray-Smith. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. *Advances in neural information processing systems (NeurIPS)*, 15, 2002.

[21] Kazimierz Goebel and William Kirk. *Topics in metric fixed point theory.* Cambridge University Press, 1990.

[22] Ibon Gracia, Dimitris Boskos, Morteza Lahijanian, Luca Laurenti, and Manuel Mazo Jr. Efficient strategy synthesis for switched stochastic systems with distributional uncertainty. *Nonlinear Analysis: Hybrid Systems*, 55:101554, 2025.

[23] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions.* Springer Science & Business Media, 2000.

[24] Karl Henrik Johansson. The quadruple-tank process: A multivariable laboratory process with an adjustable zero. *IEEE Transactions on Control Systems Technology*, 8(3):456–465, 2000.

[25] Simon Julier and Jeffrey Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[26] Harold Kushner. Numerical methods for stochastic control problems in continuous time. *SIAM Journal on Control and Optimization*, 28(5):999–1048, 1990.

[27] Daniel Landgraf, Andreas Völz, Felix Berkel, Kevin Schmidt, Thomas Specker, and Knut Graichen. Probabilistic prediction methods for nonlinear systems with application to stochastic model predictive control. *Annual Reviews in Control*, 56:100905, 2023.

[28] Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, 2010.

[29] Frederik Mathiesen, Simeon Calvert, and Luca Laurenti. Safety certification for stochastic systems via neural barrier functions. *IEEE Control Systems Letters*, 7:973–978, 2022.

[30] Robert McAllister and Peyman Mohajerin Esfahani. Distributionally robust model predictive control: Closed-loop guarantees and scalable algorithms. *IEEE Transactions on Automatic Control*, 2024.

[31] 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.

[32] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[33] Grigorios Pavliotis. Stochastic Processes and Applications. *Texts in applied mathematics*, 60, 2014.

[34] Gabriel Peyré, Marco Cuturi, et al. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[35] Ioana Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112, 2007.

[36] Cesar Santoyo, Maxence Dutreix, and Samuel Coogan. A barrier function approach to finite-time stochastic system verification and control. *Automatica*, 125:109439, 2021.

[37] Satinder Singh and Richard Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1):123–158, 1996.

[38] Robert Stengel. *Optimal Control and Estimation.* Courier Corporation, 1994.

[39] Anastasios Tsiamis and George Pappas. Finite sample analysis of stochastic system identification. In *IEEE Conference on Decision and Control (CDC)*, pages 3648–3654, 2019.

[40] Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[41] Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems (NeurIPS)*, 8, 1995.

[42] Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. On linear optimization over Wasserstein balls. *Mathematical Programming*, 195(1):1107–1122, 2022.