IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Distributionally Robust Differential Dynamic Programming with Wasserstein Distance

Astghik Hakobyan Insoon Yang

Abstract— **Differential dynamic programming (DDP) is a popular technique for solving nonlinear optimal control problems with locally quadratic approximations. However, existing DDP methods are not designed for stochastic systems with unknown disturbance distributions. To address this limitation, we propose a novel DDP method that approximately solves the Wasserstein distributionally robust control (WDRC) problem, where the true disturbance distribution is unknown but a disturbance sample dataset is given. Our approach aims to develop a practical and computationally efficient DDP solution. To achieve this, we use the Kantrovich duality principle to decompose the value function in a novel way and derive closed-form expressions of the distributionally robust control and worst-case distribution policies to be used in each iteration of our DDP algorithm. This characterization makes our method tractable and scalable without the need for numerically solving any minimax optimization problems. The superior out-of-sample performance and scalability of our algorithm are demonstrated through kinematic car navigation and coupled oscillator problems.**

I. Introduction

Nonlinear optimal control problems are difficult to solve exactly, particularly when the state space dimension is high. Differential dynamic programming (DDP) alleviates this issue using locally-quadratic approximations of the system dynamics and cost function [1]–[6]. It efficiently computes an approximate solution with superior scalability compared to the standard dynamic programming (DP) approach. However, it is generally challenging to apply DDP to systems with random disturbances without any means to counteract them.

Although various works have extended DDP to handle stochastic systems, existing methods often rely on either the ground truth or potentially inaccurate approximate probability distributions of disturbances. For instance, the DDP algorithms proposed in [7]–[9] either consider Gaussian multiplicative noise or model the uncertain system dynamics as Gaussian processes. Another line of research is devoted to the minimax formulation of the DDP problem (e.g., [10], [11]), where the optimal control problem is solved in the face of the worst-case disturbances. However, these methods often lead to overly conservative solutions.

To address the limitations of stochastic DDP methods and handle systems with unknown disturbance distributions, we propose a novel approach inspired by distributionally robust control (DRC). The objective of DRC is to design control policies that maximize the worst-case performance over an *ambiguity set* of distributions. In the context of DRC, various techniques have been introduced to hedge against distributional uncertainties, which include moment-based and statistical distance-based approaches [12]–[19]. While moment-based approaches rely on accurate moment estimates and may not effectively capture the full distributional information, distance-based methods consider distributions that are close to a given nominal one. Many recent works have focused on Wasserstein DRC (WDRC) [20]–[24], where the ambiguity set is chosen as a statistical ball with the distance between two distributions measured by the Wasserstein metric. The Wasserstein ambiguity set offers several advantages, including finite-sample performance guarantees and the ability to avoid pathological solutions in distributionally robust optimization (DRO) problems [25]–[27].

Most of the existing WDRC methods still face challenges in terms of tractability and scalability. For instance, the DP-based approach introduced in [20] for solving the WDRC problem results in a semi-infinite program that needs to be solved for each grid point on the discretized state-space. To alleviate the computational issue, both [20] and [24] propose a relaxation technique with a penalty on the Wasserstein distance, which leads to an explicit solution in the linear-quadratic (LQ) setting. While these works primarily focus on the theoretical analyses of the distributionally robust policies, our algorithm contributes to a computationally efficient solution of nonlinear WDRC problems.

In this work, a novel DDP method is developed through a locally quadratic approximation of the nonlinear WDRC problem, where the true disturbance distribution is unknown but a disturbance sample dataset is given. By construction, the proposed distributionally robust DDP (DR-DDP) algorithm provides control policies that are robust against inevitable inaccuracies in empirical distributions of the disturbance. For tractability, we first approximate the WDRC problem with its penalty version and then apply the Kantorovich duality principle. We show that this approximation provides a suboptimal solution to the original WDRC problem with a provable performance guarantee. The value function is then decomposed in a novel way that enables deriving computationally tractable and efficient backward and forward passes. This allows us to obtain closed-form expressions for the distributionally robust control and worst-case distribution policies in each iteration of the DR-DDP algorithm. By avoiding the need for numerically solving

This work was supported in part by the National Research Foundation of Korea under MSIT2020R1C1C1009766, MSIT2021R1A4A2001824, the Information and Communications Technology Planning and Evaluation under Grants MSIT2022-0-00124, MSIT2022-0-00480, and Samsung Electronics.

A. Hakobyan, and I. Yang are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul, 08826, South Korea {astghikhakobyan, insoonyang}@snu.ac.kr

Copyright ©2023 IEEE

minimax optimization problems, our approach makes the algorithm not only tractable but also scalable. The scalability of our DDP method is a remarkable advantage because the computational complexity of the standard DP algorithm in [20] for nonlinear WDRC increases exponentially with the dimension of the state space. The experiment results on kinematic car navigation and coupled oscillator problems indicate that our algorithm outperforms existing methods in terms of out-of-sample performance and provides scalable solutions for high-dimensional nonlinear optimal control problems.

II. Preliminaries

In this section, we introduce the WDRC problem used in our development of the DR-DDP algorithm in Section III.

A. Distributionally Robust Control

Consider the following discrete-time stochastic system:

$$x_{t+1} = f(x_t, u_t, w_t), \tag{1}$$

where $x_t \in \mathbb{R}^{n_x}$ and $u_t \in \mathbb{R}^{n_u}$ are the system states and control inputs, respectively. Here, $w_t \in \mathbb{R}^{n_w}$ is a random disturbance with an unknown (true) distribution $\mathbb{Q}_t^{\text{true}} \in \mathcal{P}(\mathbb{R}^{n_w})$, where $\mathcal{P}(\mathbb{R}^{n_w})$ is the family of all Borel probability measures supported on $\mathbb{R}^{n_w}$. The nonlinear function $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_w} \to \mathbb{R}^{n_x}$ is assumed to be twice continuously differentiable.

In practice, it is restrictive to assume that the true probability distribution $\mathbb{Q}_t^{\text{true}}$ is known. Instead, we are often given a sample dataset $\mathcal{D}_t := \{\hat{w}_t^{(1)}, \hat{w}_t^{(2)}, \ldots, \hat{w}_t^{(N)}\}$ drawn from the true distribution, which can be used to construct an empirical estimate about the disturbance distribution as $\mathbb{Q}_t := \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{w}_t^{(i)}}$, where $\delta_{\hat{w}_t^{(i)}}$ denotes the Dirac measure concentrated at $\hat{w}_t^{(i)}$. It is well-known that as $N \to \infty$, the empirical distribution asymptotically converges to the true distribution. However, if an inaccurate empirical estimate is used in the controller design, the resulting control performance will deteriorate due to a mismatch between the true and empirical distributions.

To mitigate the impact of distributional uncertainties, we adopt a game-theoretic approach and consider a two-player zero-sum game in which Player I is the controller and Player II is a hypothetical adversary. Let $\pi := (\pi_0, \ldots, \pi_{T-1})$ denote the control policy, where π_t maps the state x_t to a control input u_t. The adversary player selects a policy $\gamma := (\gamma_0, \ldots, \gamma_{T-1})$, where γ_t maps the current state to a probability distribution $\mathbb{P}_t$ chosen from an ambiguity set $\mathbb{D}_t \subset \mathcal{P}(\mathbb{R}^{n_w})$. The ambiguity set is a family of distributions that possess certain properties to be described.

Throughout this paper, our goal is to design an optimal finite-horizon controller with the following cost functional: $J(\pi, \gamma) := \mathbb{E}^{\pi,\gamma}\big[\ell_f(x_T) + \sum_{t=0}^{T-1} \ell(x_t, u_t)\big]$, where $\ell : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$ and $\ell_f : \mathbb{R}^{n_x} \to \mathbb{R}$ are the twice continuously differentiable running and terminal costs, respectively, and T is the time horizon. In our problem, the controller seeks a policy π^* minimizing the cost function, while the adversary aims to find a policy γ^* to maximize the same cost, which can be obtained by solving the following DRC problem:

$$\min_{\pi\in\Pi} \max_{\gamma\in\Gamma_{\mathbb{D}}} J(\pi, \gamma), \tag{2}$$

where $\Pi := \{\pi \mid \pi_t(x_t) = u_t \in \mathbb{R}^{n_u},\ \forall t\}$ and $\Gamma_{\mathbb{D}} := \{\gamma \mid \gamma_t(x_t) = \mathbb{P}_t \in \mathbb{D}_t,\ \forall t\}$ are the sets of admissible control and distribution policies, respectively.

B. Wasserstein Ambiguity Set

In problem (2), the adversary player is restricted to select a distribution from the ambiguity set $\mathbb{D}_t$, which determines the characteristics of the worst-case distribution. Therefore, it is necessary to design the ambiguity set to appropriately characterize distributional errors. Motivated by its advantages mentioned in Section I, we use the Wasserstein ambiguity set constructed around the given empirical distribution. The Wasserstein metric of order p between two distributions $\mathbb{P}$ and $\mathbb{Q}$ supported on $\mathcal{W} \subseteq \mathbb{R}^n$ represents the minimum cost of redistributing mass from one distribution to another using a small non-uniform perturbation and is defined as $W_p(\mathbb{P}, \mathbb{Q}) := \inf_{\tau\in\mathcal{P}(\mathcal{W}^2)}\{(\int_{\mathcal{W}^2} \|x - y\|^p \,\mathrm{d}\tau(x, y))^{1/p} \big|\, \Pi^1\tau = \mathbb{P}, \Pi^2\tau = \mathbb{Q}\}$, where τ is the *transport plan* with $\Pi^i\tau$ denoting its ith marginal distribution, and $\|\cdot\|$ is a norm on $\mathbb{R}^n$ which quantifies the transportation cost.

In this work, we consider the Wasserstein metric of order $p = 2$ with the transportation cost represented by the standard Euclidean norm. We design the ambiguity set as follows: $\mathbb{D}_t := \{\mathbb{P}_t \in \mathcal{P}(\mathbb{R}^{n_w}) \mid W_2(\mathbb{P}_t, \mathbb{Q}_t) \le \theta\}$ where $\theta > 0$ determines the size of $\mathbb{D}_t$. The ambiguity set is a statistical ball centered at the empirical distribution $\mathbb{Q}_t$ and contains all distributions whose Wasserstein distance from the empirical distribution is no greater than radius θ.

III. Distributionally Robust Differential Dynamic Programming

In this section, we present our main result, called DR-DDP, which efficiently finds an approximate solution to the WDRC problem. Our method exploits the Kantorovich duality principle to decompose the value function in a novel way and devise a computationally tractable algorithm.

A. Approximation with Wasserstein Penalty

In [24], the tractability and effectiveness of a penalty version of the WDRC problem are studied. Motivated by this work, we begin by replacing the Wasserstein ambiguity set constraint with a penalty term in the cost function as follows: $J_\lambda(\pi, \gamma) := \mathbb{E}^{\pi,\gamma}\big[\ell_f(x_T) + \sum_{t=0}^{T-1} \ell(x_t, u_t) - \lambda W_2(\mathbb{P}_t, \mathbb{Q}_t)^2\big]$, where $\lambda > 0$ is the penalty parameter adjusting the conservativeness of the controller. Then, the following minimax control problem approximates the original WDRC problem (2):

$$\min_{\pi\in\Pi} \max_{\gamma\in\Gamma} J_\lambda(\pi, \gamma), \tag{3}$$

where the adversary player selects policies from $\Gamma := \{\gamma := (\gamma_0, \ldots, \gamma_{T-1}) \mid \gamma_t(x_t) = \mathbb{P}_t \in \mathcal{P}(\mathbb{R}^{n_w})\}$. Note that the adversary is not restricted to select distributions from the

ambiguity set. Instead, we penalize large deviations from the empirical distribution via the penalty term, thus limiting the freedom of the adversary player.

We demonstrate in the following proposition that the cost incurred by an arbitrary policy $\pi \in \Pi$ under the worst-case distributions within the Wasserstein ambiguity set has a guaranteed cost property with respect to the worst-case penalized cost. Hence, the penalty problem (3) is a reasonable approximation as it provides a suboptimal solution to (2) with a performance guarantee.

Proposition 1: Given $\lambda > 0$, let $\pi \in \Pi$ be an arbitrary admissible policy. Then, the cost incurred by π under the worst-case distribution policy in $\Gamma_{\mathbb{D}}$ is upper-bounded as follows:

$$\sup_{\gamma \in \Gamma_{\mathbb{D}}} J(\pi, \gamma) \leq \lambda T \theta^2 + \sup_{\gamma \in \Gamma} J_\lambda(\pi, \gamma). \tag{4}$$

Its proof can be found in Appendix I of the extended version [28]. The guaranteed cost property indicates the role of the penalty parameter λ in adjusting the robustness of the control policy, thereby providing a guideline on its selection. Specifically, the penalty parameter can be chosen to minimize the upper bound in (4) under the given control policy.[1]

To formalize our algorithm, we recursively define the optimal value function for problem (3) as follows: $V_t(\boldsymbol{x}) := \inf_{\pi \in \Pi} \sup_{\gamma \in \Gamma} \mathbb{E}^{\pi,\gamma}\big[\ell_f(x_T) + \sum_{s=t}^{T-1} \ell(x_s, u_s) - \lambda W_2(\mathbb{P}_s, \mathbb{Q}_s)^2 \mid x_t = \boldsymbol{x}\big]$ for $t = T-1, \ldots, 0$, with the terminal condition $V_T(\boldsymbol{x}) = \ell_f(\boldsymbol{x})$. Then, the DP principle yields

$$\begin{aligned} V_t(\boldsymbol{x}) = \inf_{\boldsymbol{u} \in \mathbb{R}^{n_u}} \sup_{\mathbb{P} \in \mathcal{P}(\mathbb{R}^{n_w})} & \ell(\boldsymbol{x}, \boldsymbol{u}) \\ & + \mathbb{E}^{w \sim \mathbb{P}}\Big[V_{t+1}(f(\boldsymbol{x}, \boldsymbol{u}, w)) - \lambda W_2(\mathbb{P}, \mathbb{Q}_t)^2\Big] \end{aligned} \tag{5}$$

with the optimal cost given by $J_\lambda^* := \inf_{\pi \in \Pi} \sup_{\gamma \in \Gamma} J_\lambda(\pi, \gamma) = V_0(x_0)$.

The standard procedure for DDP cannot be applied to the value function (5) as it constitutes an infinite-dimensional optimization problem over $\mathcal{P}(\mathbb{R}^{n_w})$. For tractability, we employ a modern DRO technique based on the Kantorovich duality principle and reformulate the value function as follows.

Proposition 2: Suppose that for each $(\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$, the value function is measurable and that the outer minimization problem in (5) has an optimal solution. Then, for any $\lambda > 0$, we have that for all $\boldsymbol{x} \in \mathbb{R}^{n_x}$

$$\begin{aligned} & V_t(\boldsymbol{x}) = \inf_{\boldsymbol{u} \in \mathbb{R}^{n_u}} \ell(\boldsymbol{x}, \boldsymbol{u}) + \\ & \mathbb{E}^{\hat{w}_t \sim \mathbb{Q}_t}\Big[\sup_{\boldsymbol{w} \in \mathbb{R}^{n_w}} V_{t+1}(f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w})) - \lambda \|\hat{w}_t - \boldsymbol{w}\|^2\Big]. \end{aligned} \tag{6}$$

Its proof can be found in Appendix II of the extended version [28]. While previous works (e.g., [20]) use similar approaches for theoretically analyzing the WDRC problem, our focus is on designing a practical and efficient method for obtaining computationally tractable solutions. For that, we let $Q_t^{(i)}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w}) := \ell(\boldsymbol{x}, \boldsymbol{u}) + V_{t+1}(f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w})) - \lambda \|\hat{w}_t^{(i)} - \boldsymbol{w}\|^2$ denote the state-action-disturbance value function or the Q-function for each sample index $i = 1, \ldots, N$ and let $Q_t^{*,(i)}(\boldsymbol{x}, \boldsymbol{u}) = \sup_{\boldsymbol{w} \in \mathbb{R}^{n_w}} Q_t^{(i)}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w})$ denote the corresponding "worst-case" state-action value function. Then,

$$V_t(\boldsymbol{x}) = \inf_{\boldsymbol{u} \in \mathbb{R}^{n_u}} \frac{1}{N} \sum_{i=1}^{N} Q_t^{*,(i)}(\boldsymbol{x}, \boldsymbol{u}). \tag{7}$$

It is worth emphasizing that the Kantorovich duality principle enables us to obtain this novel decomposition of the value function, which can be used to design a computationally tractable DR-DDP solution in the following subsection.

B. Solution via DDP

In each iteration of the original DDP algorithm, a backward pass is performed on the current estimate of the state and control trajectories, called the *nominal trajectories*, followed by a forward pass. In the backward pass, the cost function and the system dynamics are quadratically approximated around the nominal trajectories to update the policy, while in the forward pass, the nominal trajectories are recomputed by executing the latest policy to the system. We adopt this technique for our problem and derive the backward and forward passes for the value function (6). The proposed DR-DDP method is presented in Algorithm 1.[2]

1) Backward Pass: In each backward pass, we are given nominal state, control input, and disturbance trajectories $\bar{\boldsymbol{x}}_{\text{nom}} = (\bar{\boldsymbol{x}}_0, \ldots, \bar{\boldsymbol{x}}_T)$, $\bar{\boldsymbol{u}}_{\text{nom}} = (\bar{\boldsymbol{u}}_0, \ldots, \bar{\boldsymbol{u}}_{T-1})$ and $\bar{\boldsymbol{w}}_{\text{nom}} = (\bar{\boldsymbol{w}}_0, \ldots, \bar{\boldsymbol{w}}_{T-1})$, respectively. For quadratic approximations, DDP considers the following deviations of the system state, control input, and disturbance, i.e., $\delta x_t := x_t - \bar{\boldsymbol{x}}_t$, $\delta u_t := u_t - \bar{\boldsymbol{u}}_t$, and $\delta w_t := w_t - \bar{\boldsymbol{w}}_t$.

We first consider the following second-order approximation of $V_{t+1}(x_{t+1})$:

$$\boldsymbol{V}_{t+1} + V_{t+1,x}^\top \delta x_{t+1} + \frac{1}{2} \delta x_{t+1}^\top V_{t+1,xx} \delta x_{t+1}, \tag{8}$$

for some $(\boldsymbol{V}_{t+1}, V_{t+1,x}, V_{t+1,xx}) \in \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x \times n_x}$ to be determined.[3] Let $\hat{Q}_t^{(i)}$ be an approximate Q-function, defined by replacing V_{t+1} in the definition of $Q_t^{(i)}$ with the approximate value function (8). Then, $\hat{Q}_t^{(i)}(x_t, u_t, w_t)$ is twice differentiable and its second-order Taylor expansion is given by

$$\boldsymbol{Q}_t^{(i)} + \delta Q_t^{(i)}(\delta x_t, \delta u_t, \delta w_t), \tag{9}$$

where $\delta Q_t^{(i)}(\delta x_t, \delta u_t, \delta w_t) = Q_{t,x}^\top \delta x_t + Q_{t,u}^\top \delta u_t + {Q_{t,w}^{(i)}}^\top \delta w_t + \frac{1}{2} \Delta Q_t(\delta x_t, \delta u_t, \delta w_t)$ with

$$\Delta Q_t(\delta x, \delta u, \delta w) := \begin{bmatrix} \delta x \\ \delta u \\ \delta w \end{bmatrix}^\top \begin{bmatrix} Q_{t,xx} & Q_{t,xu} & Q_{t,xw} \\ Q_{t,xu}^\top & Q_{t,uu} & Q_{t,uw} \\ Q_{t,xw}^\top & Q_{t,uw}^\top & Q_{t,ww} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \\ \delta w \end{bmatrix}$$

[1]The value of λ heavily depends on the choice of the Wasserstein ambiguity set radius θ, which is typically chosen to attain a probabilistic out-of-sample performance guarantee, given a finite dataset of disturbance samples (e.g., [25], [27]).

[2]The complexity of a single iteration of our algorithm is bounded by $O\big(T(n_x^3 + n_u^3 + (N + n_w)n_w^2)\big)$, which is polynomial in state, input and disturbance dimensions and linear in the time horizon and sample size.

[3]If V_{t+1} is twice differentiable, the parameters $(\boldsymbol{V}_{t+1}, V_{t+1,x}, V_{t+1,xx})$ can be simply determined using the second-order Taylor expansion.

and

$$\begin{cases} \boldsymbol{Q}_t^{(i)} = \ell(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) + \boldsymbol{V}_{t+1} - \lambda\|\bar{\boldsymbol{w}}_t - \hat{w}_t^{(i)}\|^2 \\ Q_{t,xx} = \ell_{t,xx} + f_{t,x}^\top V_{t+1,xx} f_{t,x} + V_{t+1,x}^\top f_{t,xx} \\ Q_{t,uu} = \ell_{t,uu} + f_{t,u}^\top V_{t+1,xx} f_{t,u} + V_{t+1,x}^\top f_{t,uu} \\ Q_{t,ww} = f_{t,w}^\top V_{t+1,xx} f_{t,w} - 2\lambda I + V_{t+1,x}^\top f_{t,ww} \\ Q_{t,xu} = \ell_{t,xu} + f_{t,x}^\top V_{t+1,xx} f_{t,u} \\ Q_{t,xw} = f_{t,x}^\top V_{t+1,xx} f_{t,w}, \quad Q_{t,uw} = f_{t,u}^\top V_{t+1,xx} f_{t,w} \\ Q_{t,x} = \ell_{t,x} + f_{t,x}^\top V_{t+1,x}, \quad Q_{t,u} = \ell_{t,u} + f_{t,u}^\top V_{t+1,x} \\ Q_{t,w}^{(i)} = f_{t,w}^\top V_{t+1,x} - 2\lambda(\bar{\boldsymbol{w}}_t - \hat{w}_t^{(i)}). \end{cases}$$

Here, $f_{t,\cdot}$ and $\ell_{t,\cdot}$ denote the partial derivatives of f and ℓ evaluated at $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t, \bar{\boldsymbol{w}}_t)$.

Let $\hat{\bar{w}}_t := \mathbb{E}^{\hat{w}_t \sim \mathbb{Q}_t}[\hat{w}_t]$ and $\hat{\Sigma}_t := \mathbb{E}^{\hat{w}_t \sim \mathbb{Q}_t}[(\hat{w}_t - \hat{\bar{w}}_t)(\hat{w}_t - \hat{\bar{w}}_t)^\top]$ denote the empirical mean vector and covariance matrix of disturbance w_t, respectively. The above approximation transforms the problem (7) into a quadratic form similar to those addressed in [20], [24]. This approximation enables us to explicitly solve the problem with respect to δu_t and δw_t, as presented in the following theorem.

Theorem 1: Let $Q_{t,ww} \prec 0$ and $\ell_{t,uu} \succ 0$. Suppose the value function at time $t+1$ is approximated as (8). Then, the outer minimization problem in (7) with $Q_t^{(i)}(x_t, u_t, w_t)$ replaced by the approximation (9) has the following unique minimizer:

$$\delta u_t^* = K_t \delta x_t + k_t, \tag{10}$$

where $K_t = -\tilde{Q}_t(Q_{t,xu}^\top - Q_{t,uw}Q_{t,ww}^{-1}Q_{t,xw}^\top)$, $k_t = -\tilde{Q}_t(Q_{t,u} - Q_{t,uw}Q_{t,ww}^{-1}\bar{Q}_{t,w})$ with $\tilde{Q}_t := (Q_{t,uu} - Q_{t,uw}Q_{t,ww}^{-1}Q_{t,uw}^\top)^{-1}$ and $\bar{Q}_{t,w} := f_{t,w}^\top V_{t+1,x} - 2\lambda(\bar{\boldsymbol{w}}_t - \hat{\bar{w}}_t)$.

Moreover, for each $i = 1, \ldots, N$, the maximization problem in (7) with $Q_t^{(i)}(x_t, u_t, w_t)$ replaced by the approximation (9) has the following unique solution:

$$\delta w_t^{*,(i)} = H_t \delta x_t + h_t^{(i)}, \tag{11}$$

where $H_t = -Q_{t,ww}^{-1}[Q_{t,uw}^\top K_t + Q_{t,xw}^\top]$ and $h_t^{(i)} = -Q_{t,ww}^{-1}[Q_{t,uw}^\top k_t + Q_{t,w}^{(i)}]$.

Proof: Let $\delta w_t^{(i)} := w_t^{(i)} - \bar{\boldsymbol{w}}_t$. Evaluating the approximate Q-function (9) for $\delta w_t^{(i)}$, we see that it is strictly concave in $\delta w_t^{(i)}$ as $Q_{t,ww} \prec 0$. Then, the first-order optimality condition yields the following unique maximizer:

$$\delta w_t^{*,(i)} = -Q_{t,ww}^{-1}(Q_{t,xw}^\top \delta x_t + Q_{t,uw}^\top \delta u_t + Q_{t,w}^{(i)}). \tag{12}$$

Replacing $Q_t^{(i)}(x_t, u_t, w_t)$ with the approximation (9), the objective function in (7) is quadratically approximated as

$$\bar{\boldsymbol{Q}}_t + Q_{t,x}^\top \delta x_t + Q_{t,u}^\top \delta u_t + \bar{Q}_{t,w}^\top \overline{\delta w}_t^* + \frac{1}{2}\Delta Q_t(\delta x_t, \delta u_t, \overline{\delta w}_t^*)$$

where $\overline{\delta w}_t^* := \frac{1}{N}\sum_{i=1}^N \delta w_t^{*,(i)} = -Q_{t,ww}^{-1}(Q_{t,xw}^\top \delta x_t + Q_{t,uw}^\top \delta u_t + \bar{Q}_{t,w})$ and $\bar{\boldsymbol{Q}}_t := \ell_t(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) + \boldsymbol{V}_{t+1} - \lambda\|\bar{\boldsymbol{w}}_t - \hat{\bar{w}}_t\|^2 - \lambda \mathrm{Tr}[\hat{\Sigma}_t] - 2\lambda^2 \mathrm{Tr}[Q_{t,ww}^{-1}\hat{\Sigma}_t]$. To minimize this approximated objective function with respect to δu_t, the following first-order optimality condition can be used:

$$\begin{aligned} 0 = {} & Q_{t,u} + Q_{t,uu}\delta u_t + Q_{t,xu}^\top \delta x_t + Q_{t,uw}\overline{\delta w}_t^* \\ & + \frac{\partial \overline{\delta w}_t^{*\top}}{\partial \delta u_t}(\bar{Q}_{t,w} + Q_{t,xw}^\top \delta x_t + Q_{t,uw}^\top \delta u_t + Q_{t,ww}\overline{\delta w}_t^*). \end{aligned}$$

Algorithm 1: DR-DDP algorithm

Input: $x_0, \pi_{\text{init}}, \gamma_{\text{init}}, T, \lambda$
Apply $(\pi_{\text{init}}, \gamma_{\text{init}})$ to generate $(\bar{\boldsymbol{x}}_{\text{nom}}, \bar{\boldsymbol{u}}_{\text{nom}}, \bar{\boldsymbol{w}}_{\text{nom}})$
while *not converged* **do**
  `// Backward Pass`
  $\boldsymbol{V}_T \leftarrow \ell_f(\bar{\boldsymbol{x}}_T), V_{T,x} \leftarrow \ell_{f,x}, V_{T,xx} \leftarrow \ell_{f,xx}$
  for $t = T-1$ **to** 0 **do**
    Construct $(\bar{\pi}_t^*, \bar{\gamma}_t^*)$ using (13a) and (13b)
    Update $\boldsymbol{V}_t, V_{t,x}, V_{t,xx}$ according to (14)
  `// Forward Pass`
  Perform line-search to update α
  for $t = 0$ **to** $T-1$ **do**
    Compute $u_t = \bar{\boldsymbol{u}}_t + \alpha k_t + K_t(x_t - \bar{\boldsymbol{x}}_t)$
    Sample $w_t \sim \frac{1}{N}\sum_{i=1}^N \delta_{\bar{\boldsymbol{w}}_t + \alpha h_t^{(i)} + H_t(x_t - \bar{\boldsymbol{x}}_t)}$
    Execute u_t and w_t to (1) and observe x_{t+1}
  $\bar{\boldsymbol{x}}_{\text{nom}} \leftarrow x_{0:T}, \bar{\boldsymbol{u}}_{\text{nom}} \leftarrow u_{0:T-1}, \bar{\boldsymbol{w}}_{\text{nom}} \leftarrow w_{0:T-1}$
return $(\bar{\pi}^*, \bar{\gamma}^*)$

By the strong convexity of the quadratic approximation, its minimizer is uniquely given by $\delta u_t^* = -\tilde{Q}_t\big(Q_{t,u} - Q_{t,uw}Q_{t,ww}^{-1}\bar{Q}_{t,w} + [Q_{t,xu}^\top - Q_{t,uw}Q_{t,ww}^{-1}Q_{t,xw}^\top]\delta x_t\big)$, which is equivalent to (10). By substituting δu_t^* into (12), we obtain the maximizer defined in (11). ∎

Theorem 1 provides the remarkable advantage that a DR-DDP policy pair $(\bar{\pi}^*, \bar{\gamma}^*)$ is constructed in the following closed-form without numerically solving any infinite-dimensional minimax optimization problems:

$$\bar{\pi}_t^*(x_t) = \bar{\boldsymbol{u}}_t + K_t(x_t - \bar{\boldsymbol{x}}_t) + k_t \tag{13a}$$

$$\bar{\gamma}_t^*(x_t) = \frac{1}{N}\sum_{i=1}^N \delta_{(\bar{\boldsymbol{w}}_t + h_t^{(i)} + H_t(x_t - \bar{\boldsymbol{x}}_t))}. \tag{13b}$$

As a result of the backward pass, we also obtain the following equations for updating the parameters of the approximate value function (8):

$$\begin{aligned} \boldsymbol{V}_t = {} & \bar{\boldsymbol{Q}}_t + Q_{t,u}^\top k_t + \bar{Q}_{t,w}^\top h_t \\ & + \frac{1}{2}k_t^\top Q_{t,uu}k_t + \frac{1}{2}h_t^\top Q_{t,ww}h_t + k_t^\top Q_{t,uw}h_t \\ V_{t,x} = {} & Q_{t,x} + Q_{t,xu}k_t + K_t^\top(Q_{t,u} + Q_{t,uu}k_t + Q_{uw}h_t) \\ & + Q_{xw}h_t + H_t^\top(\bar{Q}_{t,w} + Q_{t,ww}h_t + Q_{t,uw}^\top k_t) \\ V_{t,xx} = {} & Q_{t,xx} + K_t^\top Q_{t,uu}K_t + H_t^\top Q_{t,ww}H_t + 2Q_{t,xu}K_t \\ & + 2K_t^\top Q_{t,uw}H_t + 2Q_{t,xw}H_t, \end{aligned} \tag{14}$$

where $h_t := \frac{1}{N}\sum_{i=1}^N h_t^{(i)}$.

In the next step, the nominal trajectories have to be reconstructed using the DR-DDP policy pair $(\bar{\pi}^*, \bar{\gamma}^*)$ to update the quadratically approximated models, which is performed during the forward pass introduced in what follows.

2) Forward Pass: In the original DDP algorithm, the forward pass is performed by executing the control policy to the system. However, due to the disturbance term in the system dynamics and lack of knowledge about its true distribution, it is not trivial to perform forward rollouts for the ambiguous

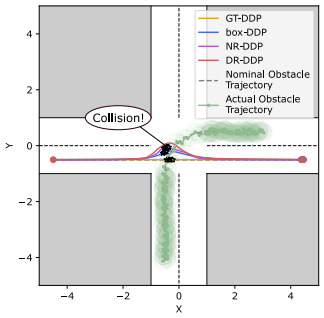


Fig. 1: Trajectories of the kinematic car, controlled by GT-DDP, box-DDP, NR-DDP, and DR-DDP, in the presence of a randomly moving obstacle. Star marks represent collisions.

stochastic system (1). Instead, we choose to execute the control and distribution policy pair $(\bar{\pi}^*, \bar{\gamma}^*)$ in the following manner. First, using (13a) and (13b), we construct a control input $u_t = \bar{\boldsymbol{u}}_t + \alpha k_t + K_t(x_t - \bar{\boldsymbol{x}}_t)$ and sample a disturbance realization as $w_t \sim \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{\boldsymbol{w}}_t + \alpha h_t^{(i)} + H_t(x_t - \bar{\boldsymbol{x}}_t)}$, where $\alpha \in (0,1)$ is a line-search parameter.[4] Then, both the control input and the disturbance sample are executed to the system for $t = 0, \ldots, T-1$ starting from the initial state x_0.

IV. Numerical Experiments

In this section, we compare the empirical performance of our DR-DDP method with three baseline algorithms: *GT-DDP* [10], which uses a minimax approach to consider the worst-case disturbances, *box-DDP* [2], a deterministic DDP algorithm that ignores uncertainties in the controller design but considers box constraints on control inputs, and *NR-DDP*, the non-robust version of our DR-DDP algorithm that utilizes the empirical distribution.[5] [6]

A. Kinematic Car Navigation

In the first experiment, we consider an autonomous navigation task for a kinematic car in an intersection where a randomly moving obstacle obstructs navigation. Consider the following system: $x_{t+1} = \begin{bmatrix} x_{t+1}^{\text{car}} \\ \boldsymbol{p}_{t+1}^{\text{obs}} \end{bmatrix} = \begin{bmatrix} f_{\text{car}}(x_t^{\text{car}}, u_t) \\ \boldsymbol{p}_t^{\text{obs}} + \Delta \boldsymbol{p}_t^{\text{obs}} + w_t \end{bmatrix}$ with system state $x_t \in \mathbb{R}^5$ and control input $u_t \in \mathbb{R}^2$. Here, $x_t^{\text{car}} \in \mathbb{R}^3$ represents the car's state evolving according to the differential-drive kinematics $f_{\text{car}} : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3$ and consists of the car's center position $\boldsymbol{p}$ and its heading angle ϕ. The control input vector comprises the velocity and steering angle of the car and has a lower limit of $\underline{\mathbf{u}} = [0, -0.6]^\top$ and an upper limit of $\overline{\mathbf{u}} = [10, 0.6]^\top$. The state component $\boldsymbol{p}_t^{\text{obs}}$ represents the position vector of a random circular obstacle with radius $r_{\text{obs}} = 0.2$. It is assumed that in each time instance, the obstacle has a nominal deterministic motion represented by $\Delta \boldsymbol{p}_t^{\text{obs}} \in \mathbb{R}^2$, which is obstructed with a positional disturbance vector $w_t \in \mathbb{R}^2$. Each component of the disturbances follows a uniform distribution $\mathcal{U}(-0.1, 0.1)$. Our DR-DDP algorithm uses only $N = 10$ samples drawn from the true distribution. The goal is to safely pass the intersection by tracking the reference trajectory x^{ref} and avoiding the obstacle in $T = 800$ steps. For this purpose, we design a time-varying cost function as $\ell_t(x, u) := \|x^{\text{car}} - x_t^{\text{ref}}\|_Q^2 + \|u\|_R^2 + Q_{\text{obs}} \exp[-\|\boldsymbol{p} - \boldsymbol{p}^{\text{obs}}\|^2/(2r^2)]$, where the last term is a soft constraint for avoiding the obstacle with a safe margin of $r = 2r_{\text{obs}}$. The weights are chosen as $Q = 10I, R = 0.1I$ and $Q_{\text{obs}} = 20$. The terminal cost is similar to the running cost with no control cost. The penalty parameter is set to $\lambda = 9000$ that is found as the minimizer of the upper bound in (4).

Fig. 1 shows the trajectories of the kinematic car for a single realization of the disturbances. Only DR-DDP successfully avoids the obstacle and accomplishes the task, resulting in the lowest total cost. Even though both box-DDP and NR-DDP drive the car away from the reference path, they collide with the obstacle, leading to increased total costs due to the soft constraint for collision avoidance. This is because box-DDP completely disregards uncertainties, while NR-DDP relies solely on inaccurate disturbance information. Meanwhile, GT-DDP incurs extremely high costs as it fails to drive the car away from the obstacle. Despite the distinct behaviors exhibited by the two algorithms, the average total computation times for DR-DDP and GT-DDP are quite similar (less than 25 *sec.*), indicating their comparable computational efficiency. To validate our results, we conducted 1,000 independent simulation runs to measure the *out-of-sample performance* of each method.[7] The proposed DR-DDP algorithm achieves an out-of-sample cost as low as 176.713, while box-DDP, NR-DDP, and GT-DDP demonstrate worse out-of-sample performance costs of 225.335, 211,461, and 198.611, respectively. These findings demonstrate the effectiveness of our algorithm in addressing distributional ambiguity in nonlinear stochastic systems.

B. Synchronization of Coupled Oscillators

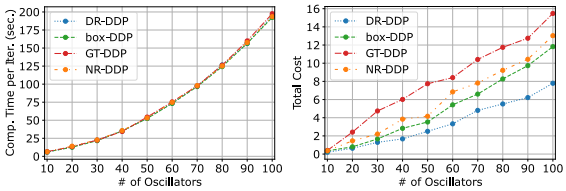
In the second experiment, we demonstrate the scalability of our algorithm through a synchronization problem with L coupled noisy oscillators using the following discrete-time Kuramoto model [29]: $\eta_{t+1}^{(i)} = \eta_t^{(i)} + \Delta t[\omega_i + \mathcal{K} u_t \sum_{j=1}^{L} \sin(\eta_t^{(j)} - \eta_t^{(i)})] + w_t^{(i)}$, $i = 1, \ldots, L$. Here, $x_t = [\eta_t^{(1)}, \ldots, \eta_t^{(L)}]^\top \in \mathbb{R}^L$ is the system state, and $u_t \in \mathbb{R}$ is the control input. For each ith oscillator, $\eta_t^{(i)}$

[4] Since DDP is a second-order method and potentially takes large steps, regularization is required to prevent the blow-up of the value. Therefore, we multiply k_t and $h_t^{(i)}$ by scaling a parameter $\alpha \in (0,1)$ and perform a line-search. In particular, the line-search parameter α is iteratively reduced to improve the performance of the DDP policy pair.

[5] In our experiments, we choose the penalty parameter λ that minimizes the cost upper bound in (4) for $\theta = 0.1$ under the DR-DDP policy pair $(\bar{\pi}^*, \bar{\gamma}^*)$. We estimate the upper bound by conducting 1,000 independent Monte Carlo simulations and computing the Wasserstein distance via a linear program. The optimal penalty parameter is then found via numerical optimization. This procedure does not require the true disturbance distribution.

[6] All simulations were performed on a PC with a 3.70 GHz Intel Core i7-8700K processor and 32 GB RAM. The source code of our DR-DDP implementation is available online: `https://github.com/CORE-SNU/DR-DDP`.

[7] The out-of-sample performance of the controller is defined as $\mathbb{E}^{w_t \sim \mathbb{Q}_t^{\text{true}}}[\ell_f(x_T) + \sum_{t=0}^{T-1} \ell(x_t, \pi_t^*(x_t))]$, which is evaluated using 10,000 disturbance samples drawn from the true distribution $\mathbb{Q}_t^{\text{true}}$ and averaged over 200 simulations. It represents the expected total cost under a new disturbance sample generated according to the true disturbance distribution $\mathbb{Q}_t^{\text{true}}$, *independent* of the sample dataset used in DR-DDP.



(a) (b)

Fig. 2: (a) Computation time per iteration (in seconds) and (b) out-of-sample cost depending on the number of oscillators calculated with 1,000 simulations.

represents its phase, $\omega^{(i)} \sim \mathcal{N}(0, 0.004)$ is its natural frequency, $\mathcal{K}$ is the coupling strength, and $\Delta t = 0.03\,sec.$ is the discretization step. Assuming disturbances $w_t^{(i)}$ follow a Gaussian distribution $\mathcal{N}(0.001, 0.001)$, and having only $N = 50$ samples, our objective is to synchronize oscillators over $T = 100$ steps using the cost function $\ell(x_t, u_t) := \sum_{i,j=1}^{L} \sin^2(\eta_t^{(j)} - \eta_t^{(i)}) + 0.0001 u_t^2$. The penalty parameter $\lambda = 10000$ is chosen to minimize the upper bound in (4).

To assess the scalability of our method, we evaluate the computation time for one iteration of our DR-DDP algorithm with varying number of oscillators. The computation times required for our method and the three baselines, along with the corresponding total costs, are presented in Fig. 2. As expected, the computation time increases with the number of oscillators. However, consistent with the theoretical complexity, the computation time grows as a polynomial function of the state dimension, showing the superiority of our method over the DP algorithm. Notably, the computation time required to perform a single iteration of DR-DDP is almost identical to the computation times required by box-DDP, NR-DDP, and GT-DDP. Furthermore, our DR-DDP algorithm consistently returns the lowest out-of-sample cost for any number of oscillators considered, successfully synchronizing the oscillators despite the disturbances.

V. Conclusions

In this work, we have proposed a practical DR-DDP algorithm for solving nonlinear stochastic optimal control problems with unknown disturbance distributions. We reformulated the quadratic approximation of value functions for WDRC using the Kantorovich duality principle and then solved it in a DDP fashion to obtain closed-form expressions of the distributionally robust control and distribution policies in each iteration. Our simulation results demonstrate the superior out-of-sample performance of the proposed method compared to existing DDP methods, as well as its notable scalability to high-dimensional state spaces.

References

[1] L.-Z. Liao and C. A. Shoemaker, "Convergence in unconstrained discrete-time differential dynamic programming," *IEEE Trans. Autom. Control*, vol. 36, no. 6, pp. 692–706, 1991.

[2] Y. Tassa, N. Mansard, and E. Todorov, "Control-limited differential dynamic programming," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014.

[3] A. Pavlov, I. Shames, and C. Manzie, "Interior point differential dynamic programming," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 6, pp. 2720–2727, 2021.

[4] W. Jallet, N. Mansard, and J. Carpentier, "Implicit differential dynamic programming," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022.

[5] O. So, Z. Wang, and E. A. Theodorou, "Maximum entropy differential dynamic programming," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022.

[6] V. Roulet, S. Srinivasa, M. Fazel, and Z. Harchaoui, "Iterative linear quadratic optimization for nonlinear control: Differentiable programming algorithmic templates," *arXiv preprint arXiv:2207.06362*, 2022.

[7] E. Todorov and W. Li, "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems," in *Proc. IEEE Am. Control Conf.*, 2005.

[8] E. Theodorou, Y. Tassa, and E. Todorov, "Stochastic differential dynamic programming," in *Proc. IEEE Am. Control Conf.*, 2010.

[9] Y. Pan, G. I. Boutselis, and E. A. Theodorou, "Efficient reinforcement learning via probabilistic trajectory optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5459–5474, 2018.

[10] W. Sun, Y. Pan, J. Lim, E. A. Theodorou, and P. Tsiotras, "Min-max differential dynamic programming: Continuous and discrete time formulations," *J. Guid. Control Dyn.*, vol. 41, no. 12, pp. 2568–2580, 2018.

[11] J. Morimoto, G. Zeglin, and C. G. Atkeson, "Minimax differential dynamic programming: Application to a biped walking robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2003.

[12] I. Yang, "A dynamic game approach to distributionally robust safety specifications for stochastic systems," *Automatica*, vol. 94, pp. 94–101, 2018.

[13] P. Coppens, M. Schuurmans, and P. Patrinos, "Data-driven distributionally robust LQR with multiplicative noise," in *Proc. Conf. Learn. Dyn. Control*, 2020.

[14] C. Mark and S. Liu, "Data-driven distributionally robust MPC: An indirect feedback approach," *arXiv preprint arXiv:2109.09558*, 2021.

[15] J. Coulson, J. Lygeros, and F. Dörfler, "Distributionally robust chance constrained data-enabled predictive control," *IEEE Trans. Autom. Control*, 2021.

[16] A. Zolanvari and A. Cherukuri, "Data-driven distributionally robust iterative risk-constrained model predictive control," in *Proc. IEEE Eur. Control Conf.*, 2022.

[17] A. Hakobyan and I. Yang, "Wasserstein distributionally robust motion control for collision avoidance using conditional value-at-risk," *IEEE Trans. Robot.*, vol. 38, no. 2, pp. 939–957, 2022.

[18] A. Dixit, M. Ahmadi, and J. W. Burdick, "Distributionally robust model predictive control with total variation distance," *arXiv preprint arXiv:2203.12062*, 2022.

[19] F. Micheli, T. Summers, and J. Lygeros, "Data-driven distributionally robust MPC for systems with uncertain dynamics," in *Proc. IEEE Conf. Decis. Control*, 2022.

[20] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *IEEE Trans. Autom. Control*, vol. 66, no. 8, pp. 3863–3870, 2021.

[21] Z. Zhong, E. A. del Rio-Chanona, and P. Petsagkourakis, "Data-driven distributionally robust MPC using the Wasserstein metric," *arXiv preprint arXiv:2105.08414*, 2021.

[22] A. B. Kordabad, R. Wisniewski, and S. Gros, "Safe reinforcement learning using Wasserstein distributionally robust MPC and chance constraint," *IEEE Access*, vol. 10, pp. 130 058–130 067, 2022.

[23] A. Hakobyan and I. Yang, "Wasserstein distributionally robust control of partially observable linear systems: Tractable approximation and performance guarantee," in *Proc. IEEE Conf. Decis. Control*, 2022.

[24] K. Kim and I. Yang, "Distributional robustness in minimax linear quadratic control with Wasserstein distance," *SIAM J. Control Optim.*, 2022.

[25] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Math. Prog.*, vol. 171, no. 1, pp. 115–166, 2018.

[26] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *Math. Oper. Res.*, 2022.

[27] D. Boskos, J. Cortés, and S. Martínez, "Data-driven ambiguity sets with probabilistic guarantees for dynamic processes," *IEEE Trans. Autom. Control*, vol. 66, no. 7, pp. 2991–3006, 2020.

[28] A. Hakobyan and I. Yang, "Distributionally robust differential dynamic programming with Wasserstein distance," *arXiv preprint arXiv:2305.09760*, 2023.

[29] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*. Springer Science & Business Media, 2012, vol. 19.