

Data Informed Residual Reinforcement Learning for High-Dimensional Robotic Tracking Control

Cong Li[†], Fangzhou Liu^{†*} *Member, IEEE*, Yongchao Wang, and Martin Buss *Fellow, IEEE*

Abstract—The learning inefficiency of reinforcement learning (RL) from scratch hinders its practical application towards continuous robotic tracking control, especially for high-dimensional robots. This work proposes a data-informed residual reinforcement learning (DR-RL) based robotic tracking control scheme applicable to robots with high dimensionality. The proposed DR-RL methodology outperforms common RL methods regarding sample efficiency and scalability. Specifically, we first decouple the original robot into low-dimensional robotic subsystems; and further utilize one-step backward (OSBK) data to construct incremental subsystems that are equivalent model-free representations of the above decoupled robotic subsystems. The formulated incremental subsystems allow for parallel learning to relieve computation load and offer us mathematical descriptions of robotic movements for conducting theoretical analysis. Then, we apply DR-RL to learn the tracking control policy, a combination of incremental base policy and incremental residual policy, under a parallel learning architecture. The incremental residual policy uses the guidance from the incremental base policy as the learning initialization and further learns from interactions with environments to endow the tracking control policy with adaptability towards dynamically changing environments. Our proposed DR-RL based tracking control scheme is developed with rigorous theoretical analysis of system stability and weight convergence. The effectiveness of our proposed method is validated numerically on a 7-DoF KUKA iiwa robot manipulator and experimentally on a 3-DoF robot manipulator that would fail for other counterpart RL methods.

Index Terms—Residual reinforcement learning, parallel learning, robotic tracking control.

I. INTRODUCTION

Reinforcement learning (RL) holds the promise of autonomous learning of continuous robotic control policies adaptable to varying environmental conditions [1]. However, RL from scratch requires an extensive amount of training data (interactions with surrounding environments) before learning one satisfying policy [2], [3]. This sample inefficiency inhibits RL from robotic practical applications. For example, large amounts of physical interactions between the robot and the real-world environments would cause undesirable mechanical

wear and even damage to the robot itself and the environment. The high dimensionality of robots would exacerbate the sample inefficiency problem mentioned above. Thereby, this work leverages the mechanism of model-based RL (MBRL) and residual RL (RRL) to facilitate the data-efficient training of continuous robotic control policies. Furthermore, the training is conducted in a parallel learning architecture, wherein the required computational load is distributed into multiple processors to relieve the computation complexity of high-dimensional robotic control tasks.

A. Related Works

The sample inefficiency causes the inapplicability of RL from scratch to robots given samples are often expensive to get. Reducing sample complexity motivates the advancement of MBRL and RRL fields. The common MBRL improves sample efficiency by firstly learning a latent model of environment dynamics and then using the learned latent model to simulate experience for policy learning [4]–[6]. Although the latent models learned via computation-intensive parametric [7], [8] or nonparametric [9], [10] methods contribute to reduced amounts of environmental samples, the model learning process introduces additional computation complexity. Besides, the learned latent models are often black-box input-out mappings that inhibit designers from conducting further theoretical analysis. How to efficiently learn a control-oriented latent model favoring both computation simplicity and theoretical analysis remains an open problem in the MBRL field. This motivates us to utilize the so-called one-step backward (OSBK) data to inform explicit latent models in simple incremental mathematical forms that would facilitate the theoretical analysis and ease the model learning difficulty. From a different perspective, RRL enhances data efficiency by learning upon one base policy and further optimizing performance via the residual policy trained by RL, rather than learning the solution from the very beginning [11]. The guidance from the base policy constrains the search space of the residual policy. This improves the exploration efficiency of the residual policy learning process. Thereby, the amount of training data and the consumed time is substantially reduced before a satisfying policy is learned. The current RRL-related works [11]–[16] are featured with task generality; however, the theoretical completeness remains to be further investigated. Therefore, this work designs the residual policy part from a control-theoretical RL perspective in favor of theoretical analysis.

The control-theoretical RL, namely approximate dynamic programming (ADP), is one RL branch featured with available

[†] Common first authors; * Corresponding author

This work was supported in part by the National Postdoctoral Program of China under Grants (GZC20233538).

C. Li is with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China. e-mail: lc@nudt.edu.cn.

Y. Wang, and M. Buss are with the Chair of Automatic Control Engineering, Technical University of Munich, Theresienstr. 90, 80333, Munich, Germany. e-mail: {yongchao.wang, mb}@tum.de.

F. Liu is with the Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, 150001, Harbin, China, e-mail: fangzhou.liu@hit.edu.cn.

theoretical analysis, which serves as one promising candidate that contributes to RRL with theoretical completeness. Although applicable for addressing the robotic (sub)optimal tracking control problem, ADP-based approaches [17]–[19] suffer learning inefficiency regarding different desired trajectories. In particular, the tracking control policy trained on one specific reference trajectory dynamics cannot efficiently track the unaccounted reference signals. However, reference signals are often described by different trajectory dynamics in complex tasks. Thereby, the associated training process would repeatedly restart but might not satisfy the required tracking performance in each learning period; This is unfavorable to practical real-time applications. For example, in a dynamic environment populated with moving obstacles, a robot keeps on replanning to generate safe desired trajectories, which are accounted for by multiple different trajectory dynamics. Thereby, the controller training processes in the works above [17]–[19] would repeatedly restart as replanning happens. However, the tracking performance during each limited training period is usually not satisfying for practical applications, especially considering safety issues demanding perfect tracking precision. This work utilizes the mechanism of control-theoretical RL to design the residual policy with a preference for theoretical completeness; While an assumed reference trajectory dynamics is avoided in the learning process. Thereby, the flexibility of our proposed tracking control scheme is extended.

In view of dimensionality, the works [17]–[19] discussed above are mostly restricted to low-dimensional domains due to learning inefficiency. This problem becomes even worse when referring to the robotic control with high dimensionality. A common approach in MBRL and RRL fields to solving high-dimensional control tasks is using the powerful approximation ability of deep neural networks (DNNs) to learn the associated policies and/or latent models [7], [20]. However, extensive amounts of samples are still required for training DNNs, which would negate the benefit of sample efficiency brought by MBRL and RRL. The control-theoretical RL methods [17]–[19] also suffer limited scalability towards high-dimensional robotic control tasks. The obstacle lies in the so-called curse of complexity. Specifically, the required number of activation functions to gain a sufficiently accurate approximation of a value function grows exponentially with the system dimension [21]. Even though a suitable large set of activation functions and appropriate hyperparameters are found through tedious engineering efforts, the accompanying computation load would degrade the realtime performance of the associated weight update law and the learned policy [22], [23]. Thus, experimental validations of control-theoretical RL-based (sub)optimal tracking control policy on a high-dimensional robot are seldom found in existing works. This work relieves the high sample complexity and computation load involved in high-dimensional robotic control tasks via a parallel learning architecture. In particular, parallel learners learn solutions to decomposed sub-problems independently while working toward a common goal.

B. Contributions

The contributions of our work are summarized as follows.

- A data-efficient and scalable DR-RL based robotic tracking control scheme is proposed to be applicable to high-dimensional robots, which succeeds in experimental tasks where common RL methods are intractable.
- The formulated data-informed incremental subsystems offer latent models for the learning process to improve sample efficiency; present a mathematical description of the robotic movement for theoretical analysis; and allow for the application of parallel learning architecture to relieve computation complexity.
- The proofs of system stability and weight convergence are provided on the basis of the formulated incremental subsystems and the parallel learners utilized off-policy experience data.

The organization of this paper is as follows. Section II presents the problem formulation. Then, the development of incremental subsystems is shown in Section III. Thereafter, Section IV presents the mechanism of the proposed robotic tracking control scheme. Section V elucidates the approximate solution learned in parallel. The developed robotic tracking control policy is numerically and experimentally validated in Section VI and Section VII, respectively. Finally, the conclusion is drawn in Section VIII.

Notations: Throughout this paper, $\mathbb{R}$ denotes the set of real numbers; $\mathbb{R}^n$ is the Euclidean space of n -dimensional real vector; $\mathbb{R}^{n \times m}$ is the Euclidean space of $n \times m$ real matrices; $\|\cdot\|$ represents the Euclidean norm for vectors and induced norm for matrices; The pseudo-inverse of the full column rank D is denoted as $D^\dagger := (D^\top D)^{-1} D^\top \in \mathbb{R}^{m \times n}$; $\text{diag}[x]$ is the $n \times n$ diagonal matrix with the i th diagonal entry equals x_i . For notational brevity, time dependence is suppressed without causing ambiguity.

II. PROBLEM FORMULATION

The robotic movement is assumed to be described by the control-affine nonlinear system:

$$\dot{x} = f(x) + g(x)u(x), \quad (1)$$

where $x \in \mathbb{R}^n$, $u(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are the system state and control input, respectively. Both $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $g(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ are bounded and locally Lipschitz. Assume that the explicit knowledge of $f(x)$ and $g(x)$ is unknown and $\text{rank}(g) = n$ holds for the robot (1).

This work focuses on the high-dimensional robotic tracking control task presented in Problem 1.

Problem 1. *Given a desired trajectory $x_d \in \mathbb{R}^n$, learning an efficient tracking control policy $u(x)$ applicable to the high-dimensional robot (1).*

The high-uncertainty property of Problem 1 encourages us to use RL based approaches. However, the high-dimensionality property of Problem 1 invalidates the basic RL from scratch approaches with limited scalability. In the following sections, we clarify our scalable and efficient RL methodology to solve Problem 1.

III. DATA INFORMED INCREMENTAL SUBSYSTEM

This section benefits from the decoupling technique and the OSBK data [24] to develop incremental subsystems. The formulated incremental subsystems jointly describe the movement of the original robot (1) without using explicit model information. Specifically, the high-dimensional robot is first decoupled into multiple low-dimensional subsystems in Section III-A. Then, the OSBK data is utilized to estimate the unknown model information (including coupling terms between subsystems) for constructing the incremental subsystems in Section III-B.

A. Decoupled Subsystem

The high-dimensional robot (1) is supposed to be decoupled into $N \in \mathbb{R}^+$ subsystems, wherein the i th subsystem follows

$$\dot{x}_i = f_i + g_i u_i, \quad i = 1, 2, \dots, N, \quad (2)$$

where $x_i \in \mathbb{R}^{n_i}$, $u_i \in \mathbb{R}^{m_i}$ are the local state and control input of the i th subsystem; $f_i \in \mathbb{R}^{n_i}$ is a combination of the local internal dynamics and the coupling terms of the i th subsystem. $g_i \in \mathbb{R}^{n_i \times m_i}$ is the local input gain matrix. Note that both f_i and g_i are state-dependent matrices. Besides, $n = \sum_i^N n_i$ and $m = \sum_i^N m_i$ hold.

For a better explanation, we focus on the robot manipulator case in Example 1 to show the transformation from the robot (1) into its associated decoupled subsystems (2).

Example 1. The robot manipulator could be described by the Euler-Lagrange (E-L) equation [25]:

$$M(q)\ddot{q} + N(q, \dot{q}) + F(\dot{q}) = \tau, \quad (3)$$

where $M(q) : \mathbb{R}^{n_r} \rightarrow \mathbb{R}^{n_r \times n_r}$ is the symmetric positive definite inertia matrix; $N(q, \dot{q}) := C(q, \dot{q})\dot{q} + G(q) : \mathbb{R}^{n_r} \times \mathbb{R}^{n_r} \rightarrow \mathbb{R}^{n_r}$, $C(q, \dot{q}) : \mathbb{R}^{n_r} \times \mathbb{R}^{n_r} \rightarrow \mathbb{R}^{n_r \times n_r}$ is the matrix of centrifugal and Coriolis terms, $G(q) : \mathbb{R}^{n_r} \rightarrow \mathbb{R}^{n_r}$ represents the gravitational terms; $F(\dot{q}) : \mathbb{R}^{n_r} \rightarrow \mathbb{R}^{n_r}$ denotes the viscous friction; $q, \dot{q}, \ddot{q} \in \mathbb{R}^{n_r}$ are the vectors of angles, velocities, and accelerations, respectively; $\tau \in \mathbb{R}^{m_r}$ represents the input torque vector. A fully actuated robot manipulator is considered here, thus $n_r = m_r$.

The high-dimensional robot (3) could be decoupled into n_r subsystems, wherein the i th subsystem reads

$$M_{ii}\ddot{q}_i + \sum_{j=1, j \neq i}^n M_{ij}\ddot{q}_j + N_i + F_i = \tau_i, \quad i = 1, 2, \dots, n_r, \quad (4)$$

where M_{ij} denotes the ij th entry of the matrix M , and N_i (F_i) is the i th entry of the vector N (F).

Let $x_{i_{r1}} := q_i \in \mathbb{R}$, $x_{i_{r2}} := \dot{q}_i \in \mathbb{R}$, $f_{i_r} := -(\sum_{j=1, j \neq i}^n M_{ij}\ddot{q}_j + N_i + F_i)/M_{ii} \in \mathbb{R}$, and $g_{i_r} := 1/M_{ii} \in \mathbb{R}$, we rewrite (4) as

$$\dot{x}_{i_{r1}} = x_{i_{r2}}, \quad (5a)$$

$$\dot{x}_{i_{r2}} = f_{i_r} + g_{i_r} \tau_i, \quad i = 1, 2, \dots, n_r \quad (5b)$$

Remark 1. The decoupled subsystems (2) are beneficial to alleviate the computation complexity induced by the high dimensionality property of Problem 1. This is because the decoupled subsystems allow for the parallel learning architecture in Section V, wherein the required intensive computational load for solving Problem 1 is distributed into multiple processors.

B. Incremental Subsystem

This subsection exploits the OSBK data to estimate the unknown model knowledge f_i and g_i in (2). This departs from common methods that identify the unknown f_i, g_i explicitly through a tedious identification process [26]–[29].

To facilitate estimation, we first introduce a predetermined constant matrix $\bar{g}_i \in \mathbb{R}^{n_i \times m_i}$ and multiply $\bar{g}_i^\dagger$ on (2),

$$\bar{g}_i^\dagger \dot{x}_i = h_i + u_i, \quad (6)$$

where $h_i := (\bar{g}_i^\dagger - g_i^\dagger)\dot{x}_i + g_i^\dagger f_i \in \mathbb{R}^{m_i}$ is a lumped term that embodies all of the unknown model knowledge in (2).

Then, with a sufficiently high sampling rate¹ [30], [31], we estimate the unknown h_i as

$$\hat{h}_i = h_{i,0} = \bar{g}_i^\dagger \dot{x}_{i,0} - u_{i,0}, \quad (7)$$

utilising the OSBK data $\dot{x}_{i,0} = \dot{x}_i(t - t_s)$, $u_{i,0} = u_i(t - t_s)$, where $t_s \in \mathbb{R}^+$ is the sampling time.

Substituting (7) into (6), we finally obtain the i th incremental subsystem

$$\dot{x}_i = \dot{x}_{i,0} + \bar{g}_i(\Delta u_i + \xi_i), \quad (8)$$

where $\Delta u_i := u_i - u_{i,0} \in \mathbb{R}^{m_i}$ is the incremental policy; and $\xi_i := h_i - \hat{h}_i \in \mathbb{R}^{m_i}$ denotes the estimation error proved to be bounded in Section IV Lemma 1.

The above formulated incremental subsystem (8) is an equivalent of the i th subsystem (2); however, no explicit model information is required.

Remark 2. The OSBK data ($\dot{x}_{i,0}$ and $u_{i,0}$ in particular) informed (8) offers us with one model-free representation of the subsystem (2). The resulting incremental subsystem benefits both efficient, parallel implementation and rigorous theoretical analysis. In particular, the informed incremental subsystem allows us to conduct the value function learning process in Section V following the MBRL mechanism in a parallel manner. Thereby, the learning efficiency is substantially improved. Furthermore, the combination of the incremental subsystems offers a mathematical form describing the robotic movement, which permits us to use the tool from the control field to conduct the rigorous theoretical analysis.

Through the aforementioned analysis (2)–(8), we could decompose the robotic tracking task in Problem 1 into sub-tasks regarding the incremental subsystems (8), as clarified in Problem 2.

Problem 2. For the incremental subsystem (8), learning the incremental policy Δu_i that drives the subsystem to track its associated desired trajectory precisely.

In the following section, we focus on Problem 2 to present our proposed tracking control scheme.

¹The so-called sufficiently high sampling rate, which is a prerequisite for estimating the unknown h_i by reusing past measurements of states and control inputs, can be chosen as the value that is larger than 30 times the system bandwidth [30].

IV. DR-RL BASED ROBOTIC TRACKING CONTROL SCHEME

This section details our proposed DR-RL based robotic tracking control scheme (see Figure. 1), wherein the incremental policies for solving Problem 2 are learned in parallel.

Under a parallel learning architecture, the learning process follows the RRL mechanism, wherein the incremental policy

$$\Delta u_i = \Delta u_{i_b} + \Delta u_{i_r}, \quad (9)$$

is an addition of the incremental base policy $\Delta u_{i_b} \in \mathbb{R}^{m_i}$ and the incremental residual policy $\Delta u_{i_r} \in \mathbb{R}^{m_i}$. The detailed procedures to design Δu_{i_b} , Δu_{i_r} and also their roles are later clarified in Section IV-A and Section IV-B, respectively.

Let $e_i := x_i - x_{d_i} \in \mathbb{R}^{n_i}$ denote the local tracking error, where $x_{d_i} \in \mathbb{R}^{n_i}$ denotes the local desired trajectory of the i th subsystem. Combining with (8) and (9), we would get the incremental error dynamics

$$\dot{e}_i = \dot{x}_{i,0} + \bar{g}_i(\Delta u_{i_b} + \Delta u_{i_r} + \xi_i) - \dot{x}_{d_i}. \quad (10)$$

In the remaining part of this section, we will focus on (10) to clarify the designed incremental base and residual policies that jointly enforce the local tracking error e_i to zero.

A. Incremental Base Policy

This subsection details the design of the incremental base policy that would offer guidance for the incremental residual policy learning process clarified in Section IV-B. The guidance would simplify the exploration. Thus, the learning difficulty of the incremental residual policy for complex and longer-horizon tasks is decreased.

Practitioners could use existing knowledge from either control or learning fields to design the incremental base policy, which is presented in detail in the following.

1) *Control Perspective*: The incremental base policy could be implemented as fine-tuned feedforward or feedback controllers such as proportional-integral-derivative, impedance control, or nonlinear dynamic inversion (NDI). In this case, the existing control knowledge is utilized to guide the learning process and offers us avenues to conduct further theoretical analysis, as illustrated later in Theorem 2. The following uses Example 2 to clarify how the NDI control technique contributes to constructing the incremental base policy and its accompanying benefits regarding the theoretical analysis.

Example 2. *This example adopts the incremental base policy designed as*

$$\Delta u_{i_b} = \bar{g}_i^\dagger(\dot{x}_{d_i} - \dot{x}_{i,0} - k_i e_i), \quad (11)$$

where $k_i \in \mathbb{R}^{n_i \times n_i}$ is a predetermined constant matrix.

The application of Δu_{i_b} (11) on (10) yields the incremental error dynamics

$$\dot{e}_i = -k_i e_i + \bar{g}_i(\Delta u_{i_r} + \xi_i), \quad (12)$$

that could be further stabilized by the incremental residual policy Δu_{i_r} . After that, focusing on (12), the incremental residual policy Δu_{i_r} would learn upon the incremental base policy (11) to further minimize the tracking error e_i . This improves the tracking precision and the robustness of the

incremental policy (9). The explicit form of the incremental error dynamics presented in (12) offers us avenues to conduct theoretical analysis using the Lyapunov tool, as illustrated later in Theorem 2.

Although the theoretical analysis on the basis of (12) is possible, practitioners should be aware of the expert knowledge and the engineering effort (debugging k_i) involved. Besides, the fine-tuned traditional controllers for one specific task often lack generalization towards other different tasks.

2) *Learning Perspective*: Alternatively, here the incremental base policy is implemented as a policy from imitation learning. In this case, the reasoning ability of human demonstrations is embedded into the incremental base policy that would guide the incremental residual policy to complete complex tasks. In the following, Example 3 is provided to exemplify the incremental base policy constructed from expert demonstrations.

Example 3. *This example uses behavioral cloning, one simple kind of imitation learning, to design the incremental base policy for an explanation. For one certain task, assume that the demonstration dataset $\mathcal{D} := \{(s_j, a_j) | j = 1, 2, \dots\}$ is available, where s_j , a_j are states and actions in suitable dimensions. The base policy $u_b(\theta)$ parameterized by θ is learned by solving the following optimization problem:*

$$\theta^* = \arg \min_{\theta} \sum_{(s_j, a_j) \sim \mathcal{D}} \|a_j - u_b(\theta)\|^2. \quad (13)$$

The required incremental base policy Δu_{i_b} for our work is gotten via the computation $\Delta u_{i_b}(kt_s) = u_{i_b}((k+1)t_s, \theta^*) - u_{i_b}(kt_s, \theta^*)$, where $k \in \mathbb{R}^+$.

The data-driven methods from the learning field are suitable for hard-to-specify tasks (challenging to be specified explicitly using rules or constraints) that are inefficient or even intractable for conventional controllers; especially in the case of available cheap (easily gotten) data. However, it is difficult to offer theoretical analysis given the non-interpretability of the utilized incremental base policy.

Remark 3. *The incremental base policy designed from different techniques either in control or learning fields implies the modularity of our proposed tracking control scheme. The property of the investigated problem, the available source (expert knowledge or data in particular), and the designers' preference (theoretical guarantee or task generalization) jointly determine the explicit method used to design the incremental base policy.*

Whether the incremental base policy is designed from control or learning perspectives, its adaptation ability towards dynamically changing environments is limited because the incremental base policy is often designed in static environments. The performance of the incremental base policy has already been determined by pre-collected data or prior-set controller parameters. This motivates us to conduct further learning on the basis of the incremental base policy to improve performance in a dynamically changing environment.

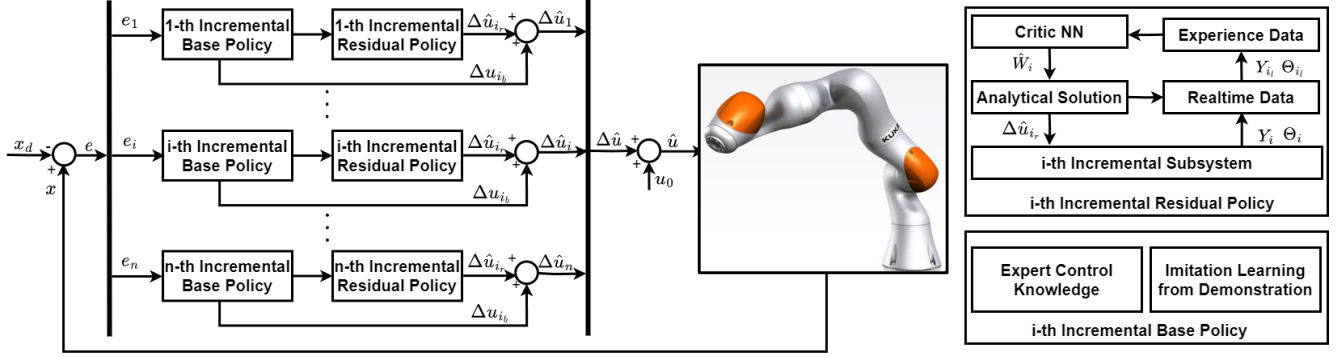


Fig. 1: Schematic of the DR-RL based robotic tracking control policy. The original high-dimensional robotic tracking task is decoupled into subtasks of incremental subsystems for efficient parallel implementation. The incremental policies, a combination of the incremental base policy and the incremental residual policy, are learned in parallel to solve the associated subtasks. The incremental base policy provides a policy initialization for the subsequent incremental residual policy learning process.

B. Incremental Residual Policy

This section utilizes the control-theoretical RL to develop the incremental residual policy that learns upon the incremental base policy to get improved task performance and robustness. In particular, we use the incremental base policy in Section IV-A to initialize the incremental residual policy learning process. Then, the incremental residual policy learns adaptations to the incremental base policy in an optimization process, where tracking errors and energy consumption are minimized for improved performance. The incremental residual policy learned from the interactions with environments endows the resulting incremental policy (9) with enhanced robustness towards dynamically changing environments.

For convenience, we represent the incremental error dynamics (10) as

$$\dot{e}_i = \bar{f}_i + \bar{g}_i \Delta u_{i_r} + \bar{g}_i \xi_i, \quad (14)$$

where $\bar{f}_i := \dot{x}_{i,0} + \bar{g}_i \Delta u_{i_b} - \dot{x}_{d_i} \in \mathbb{R}^{n_i}$. The explicit form of Δu_{i_b} has been determined in Section IV-A.

In the following, we will learn Δu_{i_r} to track the desired trajectory precisely (stabilizing the incremental error dynamics (14) interpreting from a control perspective).

Given ξ_i in (14) is unknown, thus the available incremental error dynamics for later analysis follows

$$\dot{e}_i = \bar{f}_i + \bar{g}_i \Delta u_{i_r}. \quad (15)$$

The following value function

$$V_i(t) = \int_t^\infty r_i(e_i(\nu), \Delta u_{i_r}(\nu)) d\nu, \quad (16)$$

is considered for the incremental residual policy learning process to enhance performance, where $r_i(e_i, \Delta u_{i_r}) := e_i^\top Q_i e_i + W_i(\Delta u_{i_r}) + \bar{\xi}_{oi}^2$. The quadratic term $e_i^\top Q_i e_i$, where $Q_i \in \mathbb{R}^{n_i \times n_i}$ is a positive definite matrix, is introduced to improve the tracking precision. The input penalty function $W_i(\Delta u_{i_r})$ follows

$$W_i(\Delta u_{i_r}) = 2 \int_0^{\Delta u_{i_r}} \beta \tanh^{-1}(\vartheta/\beta) d\vartheta, \quad (17)$$

which is utilized to punish and enforce the incremental residual policy as $\|\Delta u_{i_r}\| \leq \beta \in \mathbb{R}^+$. The limited Δu_{i_r} is beneficial

since a severe interruption might lead to an abrupt change of Δu_{i_r} , which might destabilize the learning process introduced in Section V. The estimation error related term follows $\bar{\xi}_{oi} = \bar{c}_i \|\Delta u_{i_r}\|$, where $\bar{c}_i \in \mathbb{R}^+$ is chosen as illustrated in Theorem 1. Note that $\bar{\xi}_{oi}$ is introduced to account for the influence of the estimation error ξ_i (temporally ignored in (15)) on the learning process. The proof given in Theorem 1 illustrates the rationality of incorporating $\bar{\xi}_{oi}$ into the value function to address the estimation error during the optimization process.

For $\Delta u_{i_r} \in \Psi$, where Ψ is the set of admissible incremental control policies [30, Definition 1], the associated optimal value function follows

$$V_i^* = \min_{\Delta u_{i_r} \in \Psi} \int_t^\infty r_i(e_i(\nu), \Delta u_{i_r}(\nu)) d\nu. \quad (18)$$

Define the Hamiltonian function as

$$H_i(e_i, \Delta u_{i_r}, \nabla V_i) = r(e_i, \Delta u_{i_r}) + \nabla V_i^\top (\bar{f}_i + \bar{g}_i \Delta u_{i_r}), \quad (19)$$

where $\nabla(\cdot) := \partial(\cdot)/\partial e_i$. Then, V_i^* satisfies the Hamilton-Jacobi-Bellman(HJB) equation

$$0 = \min_{\Delta u_{i_r} \in \Psi} [H_i(e_i, \Delta u_{i_r}, \nabla V_i^*)]. \quad (20)$$

Assume that the minimum of (18) exists and is unique [30, [32]. By using the stationary optimality condition on the HJB equation (20), we get the optimal incremental residual policy

$$\Delta u_{i_r}^* = -\beta \tanh\left(\frac{1}{2\beta} \bar{g}_i^\top \nabla V_i^*\right), \quad (21)$$

in an analytical form. To obtain $\Delta u_{i_r}^*$, we need to solve the HJB equation (20) to determine the value of ∇V_i^* , which is detailedly clarified in Section V.

In the following part of this subsection, based on the estimation error bound given in Lemma 1, we prove in Theorem 1 that the incremental residual policy $\Delta u_{i_r}^*$ (21) for (15) robustly stabilize the incremental error dynamics (14).

Lemma 1. *Given a sufficiently high sampling rate, $\exists \bar{\xi}_i \in \mathbb{R}^+$, there holds $\|\xi_i\| \leq \bar{\xi}_i$.*

Proof. The proof is available in Appendix C. $\square$

Theorem 1. Consider the incremental error dynamics (14) with a sufficiently high sampling rate, if there exists a scalar $\bar{c}_i \in \mathbb{R}^+$ such that the following inequality is satisfied

$$\bar{\xi}_i < \bar{c}_i \|\Delta u_{i,r}\|, \quad (22)$$

the optimal incremental residual policy (21) regulates the tracking error to a small neighborhood around zero while minimizing the value function (16).

Proof. See Appendix A. $\square$

Theorem 1 implies that the incremental residual policy $\Delta u_{i,r}^*$ (21) robustly stabilize (14) in an optimal manner. Thus, the combination with the incremental residual policy $\Delta u_{i,r}^*$ and the incremental base policy $\Delta u_{i,b}$ solves Problem 2 together.

Remark 4. The robot working at a sufficiently high sampling rate is the prerequisite of Lemma 1. The high sampling rate is only possible for methods with low computational complexity. To make our proposed DR-RL method work at a high sampling rate, this work uses a parallel learning architecture to distribute the intensive computation load into multiple processors; and seeks an analytical solution to the incremental residual policy that favors low computation complexity.

V. PARALLEL COMPUTATION OF APPROXIMATE SOLUTION

This section presents parallelized critic agents for learning the approximate solution to ∇V_i^* in the HJB equation (20) via a computationally efficient parallel way. Thereby, an approximation to the optimal incremental residual policy (21) is obtained. The exploitation of realtime and experience data together facilitates one simple yet efficient off-policy critic neural network (NN) weight update law with guaranteed weight convergence and improved sample efficiency.

A. Value Function Approximation

For $e_i \in \Omega$, where $\Omega \subset \mathbb{R}^{n_i}$ is a compact set, the i th continuous optimal value function (18) is approximated by i th parallelized critic agent as [32]

$$V_i^* = W_i^{*\top} \Phi_i(e_i) + \epsilon_i(e_i), \quad (23)$$

where $W_i^* \in \mathbb{R}^{N_i}$, $\Phi_i(e_i) : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{N_i}$, and $\epsilon_i(e_i) \in \mathbb{R}$ denote the NN weight, the activation function, and the approximation error of the i th parallelized critic agent, respectively.

Remark 5. The utilized decoupling technique in Section III solves the curse of complexity problem in value function approximation. In particular, the size of the constructed i th critic NN (23) relies on the dimension of the local error e_i . The n_i -D e_i allows us to construct a low-dimensional $\Phi_i(e_i)$ (easy to choose)² to approximate its associated V_i^* regardless of the value of the original robot dimension n . Otherwise, for a global approximation, i.e., $V^* = W^{*\top} \Phi(e) + \epsilon(e)$ with $e := x - x_d \in \mathbb{R}^n$, the dimension of $\Phi(e)$ increases exponentially as n increases.

²It is displayed in Sections VI and VII that 4-D activation functions $\Phi_i(e_i)$ in a fixed structure are chosen for subsystems of 2-DoF, 3-DoF, and 7-DoF robot manipulator, and also a 6-DoF quadrotor in the appendix.

To facilitate the later theoretical analysis, here we provide an assumption that is common in related works [32].

Assumption 1. There exist constants $b_{\epsilon_i}, b_{\epsilon_{ei}}, b_{\epsilon_{hi}}, b_{\Phi_i}, b_{\Phi_{ei}} \in \mathbb{R}^+$ such that $\|\epsilon_i(e_i)\| \leq b_{\epsilon_i}$, $\|\nabla \epsilon_i(e_i)\| \leq b_{\epsilon_{ei}}$, $\|\epsilon_{hi}\| \leq b_{\epsilon_{hi}}$, $\|\Phi_i(e_i)\| \leq b_{\Phi_i}$, and $\|\nabla \Phi_i(e_i)\| \leq b_{\Phi_{ei}}$.

Given a fixed i th incremental residual policy $\Delta u_{i,r}$, combining (20) with (23) yields

$$W_i^{*\top} \nabla \Phi_i(\bar{f}_i + \bar{g}_i \Delta u_{i,r}) + r_i(e_i, \Delta u_{i,r}) = \epsilon_{hi}, \quad (24)$$

where the i th residual error follows $\epsilon_{hi} := -\nabla \epsilon_i^\top(\bar{f}_i + \bar{g}_i \Delta u_{i,r}) \in \mathbb{R}$. The NN parameterized (24) is rewritten as

$$\Theta_i = -W_i^{*\top} Y_i + \epsilon_{hi}, \quad (25)$$

where $\Theta_i := r_i(e_i, \Delta u_{i,r}) \in \mathbb{R}$, and $Y_i := \nabla \Phi_i(\bar{f}_i + \bar{g}_i \Delta u_{i,r}) \in \mathbb{R}^{N_i}$. This formulated linear in parameter form simplifies the development of an efficient NN weight update law in the subsequent subsection.

B. Off-Policy Critic NN Weight Update Law

An approximation of (25) follows

$$\hat{\Theta}_i = -\hat{W}_i^\top Y_i, \quad (26)$$

where $\hat{W}_i \in \mathbb{R}^{N_i}$, $\hat{\Theta}_i \in \mathbb{R}$ are estimates of W_i^* and Θ_i , respectively. To achieve $\hat{W}_i \rightarrow W_i^*$, we design the off-policy critic NN weight update law

$$\dot{\hat{W}}_i = -\Gamma_i k_{t_i} Y_i \tilde{\Theta}_i - \sum_{l=1}^{P_i} \Gamma_i k_{e_i} Y_{i_l} \tilde{\Theta}_{i_l}, \quad (27)$$

for the i th parallelized critic agent to learn the NN weight $\hat{W}_i$ in a parallel way via minimizing $E_i := \frac{1}{2} \tilde{\Theta}_i^\top \tilde{\Theta}_i$, where $\tilde{\Theta}_i := \Theta_i - \hat{\Theta}_i \in \mathbb{R}$. Here $\Gamma_i \in \mathbb{R}^{N_i \times N_i}$ is a constant positive definite gain matrix; $k_{t_i}, k_{e_i} \in \mathbb{R}^+$ are used to trade-off the contribution of realtime and experience data to the online NN weight learning process; $P_i \in \mathbb{R}^+$ is the number of recorded experience data.

To guarantee the weight convergence of (27), as proved in Lemma 2, the exploited experience data should be sufficiently rich to satisfy the rank condition in Assumption 2, which could be easily satisfied by sequentially reusing experience data [30].

Assumption 2. Given an experience buffer $\mathfrak{B}_i = [Y_{i_1}, \dots, Y_{i_{P_i}}] \in \mathbb{R}^{N_i \times P_i}$, there holds $\text{rank}(\mathfrak{B}_i) = N_i$.

Lemma 2. Given Assumption 2, the NN weight learning error $\tilde{W}$ converges to a small neighborhood around zero.

Proof. The proof is similar to our previous work [30], which relates with the optimal regulation control problem. Thus, it is omitted here due to page limits. $\square$

The guaranteed weight convergence of $\hat{W}_i$ to W_i^* presented in Lemma 2 permits us to adopt a computation-simple single critic NN structure, where the estimated critic NN weight $\hat{W}_i$ is directly used to construct the approximate optimal incremental residual policy:

$$\Delta \hat{u}_{i,r} = -\beta \tanh\left(\frac{1}{2\beta} \bar{g}_i^\top \nabla \Phi_i^\top \hat{W}_i\right). \quad (28)$$

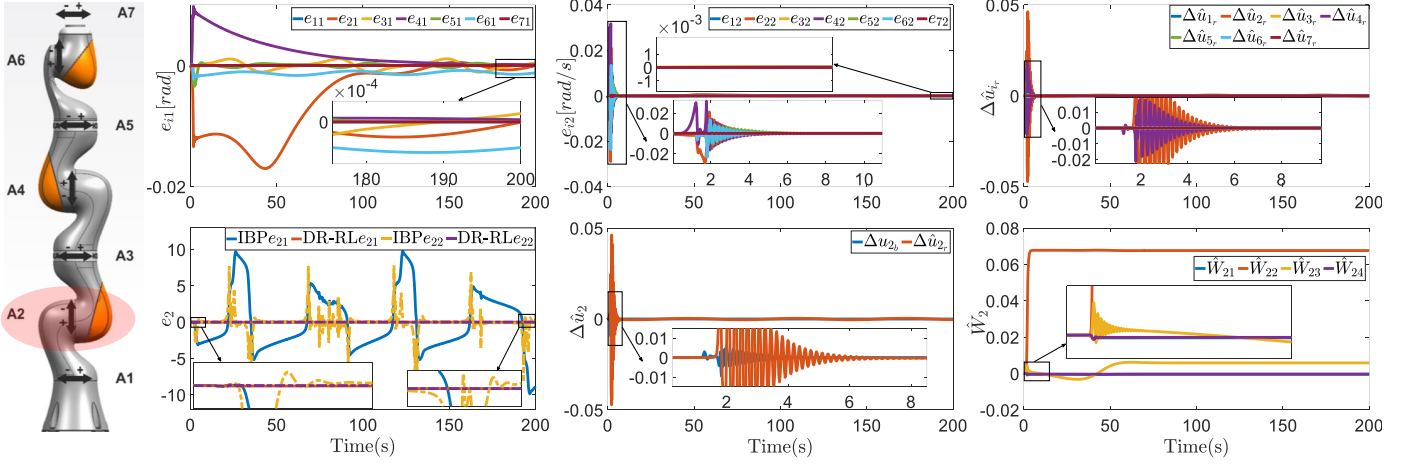


Fig. 2: The numerical validation results on the 7-DoF KUKA iiwa robot manipulator. Top: the evolution trajectories of the i th subsystem's tracking error e_{i1} , e_{i2} and the learned IRP $\Delta \hat{u}_{i,r}$, $i = 1, 2, \dots, 7$; Bottom: the comparative simulation results focusing on the 2nd subsystem, including the associated error trajectories of both incremental base policy (IBP) and DR-RL cases, the evolution trajectories of the IBP $\Delta \hat{u}_{i,r}$ and the incremental residual policy $\Delta \hat{u}_{i,r}$, and the weight convergence result.

Finally, combining with (9) and (28) gets the robotic policy

$$\hat{u}_i = u_{i,0} + \Delta u_{i,b} + \Delta \hat{u}_{i,r}, \quad (29)$$

applied at the i th subsystem (2). Based on the theoretical analysis mentioned above, we provide the main conclusions in Theorem 2.

Theorem 2. *Given Assumptions 1–2, for a sufficiently large N_i , the off-policy critic NN weight update law (27), and the approximate optimal incremental residual policy (28) guarantee the tracking error and the NN weight learning error uniformly ultimately bounded (UUB).*

Proof. See Appendix B. $\square$

VI. COMPARATIVE NUMERICAL SIMULATION

This section presents comparative numerical validations to show the superiority of our proposed DR-RL method over other methods. We firstly compare the performance of the incremental base policy and the DR-RL based control strategies on the basis of the high-dimensional 7-DoF KUKA iiwa robot manipulator and 6-DoF quadrotor, which clearly present the improved task performance induced by the online learning part. Then, we compare with two fine-tuned baselines [17], [18] focusing on a 2-DoF benchmark robot manipulator to show the superiority of our proposed DR-RL over the baseline methods regarding sample efficiency, tracking performance, and task flexibility.

A. Validations on high-dimensional robotics

This subsection fully validates the performance of our proposed DR-RL method based on the high-dimensional robotic systems: the 7-DoF KUKA iiwa robot manipulator in Matlab Robotics System Toolbox and the commonly used benchmark 6-DoF quadrotor platform.

A desired trajectory $x_d = [q_d^\top, \dot{q}_d^\top]^\top \in \mathbb{R}^{14}$, wherein $q_d = (1 + \sin(\frac{t}{10} - \frac{\pi}{2}))k_{p1} \in \mathbb{R}^7$ with $k_{p1} = 0.1\mathbf{I}_7$ is adopted to simulate a manipulation task for the KUKA iiwa

robot manipulator. We adopt the 4-D activation function $\Phi_i(e_i) = [e_{i1}^2, e_{i2}^2, e_{i1}e_{i2}, e_{i2}^3]^\top$ for the accurate value function approximation of the i th subsystem, $i = 1, 2, \dots, 7$. The detailed parameter settings of the incremental base policy and the DR-RL based control policies are referred to in Table I in the appendix.

The error trajectories of the joint angles and angle velocities present in Fig. 2 fully validate the performance of our proposed method. Furthermore, the comparative results focusing the 2th subsystem (difficulty in control due to gravity) exemplify the enhanced task performance from the online learning part, which realizes the desired weight convergence as presented in Fig. 2.

Furthermore, we test the generality of our proposed DR-RL method on different high-dimensional robotic systems via the trajectory tracking task of one 6-DoF benchmark quadrotor platform. The detailed parameter settings and results are referred to in Appendix E, which are omitted here due to the page limit.

B. Validations of Sample Efficiency and Task Flexibility

This subsection considers a task-space circle tracking task (centered at $c = (1, 1)$ with radius $r = 0.5$) of a 2-DoF robot manipulator (the detailed model information is referred to in Appendix F) for our proposed DR-RL method and one learning-from-scratch RL approach that uses a discounted factor suppressed value function [17] to learn the approximate optimal tracking control policy (referred to as DF-RL for simplicity). The associated joint-space reference trajectory $x_d = [q_d^\top, \dot{q}_d^\top]^\top \in \mathbb{R}^4$ is calculated through analytical inverse kinematics. The DF-RL³ approach adopts a 10-D activation function for the accurate value function approximation of each subsystem, which is referred to in Appendix F. While our proposed DR-RL approach only requires a 4-D activation function $\Phi_i(e_i) = [e_{i1}^2, e_{i2}^2, e_{i1}e_{i2}, e_{i2}^3]^\top$ for each subsystem,

³We introduce the parallel learning architecture to reformulate the method proposed in [17] to make it work on a 2-DoF robot manipulator.

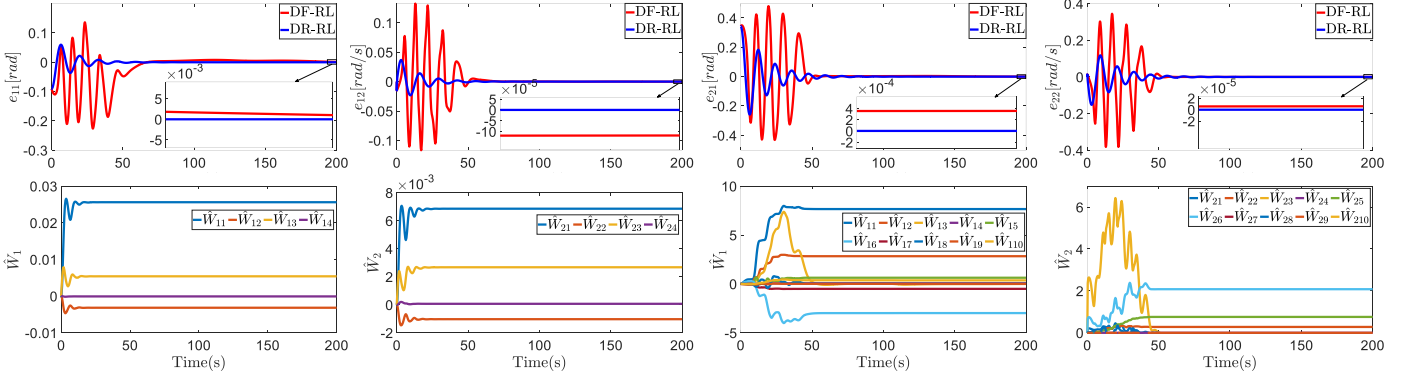


Fig. 3: The comparative numerical simulation results about the task-space task of the 2-DoF robot manipulator. Top: the evolution trajectories of the i th subsystem's tracking errors e_{i1} , e_{i2} based on the DF-RL and DR-RL methods, $i = 1, 2$; Bottom: the evolution trajectories of the i th subsystem's 4-D estimated NN weight $\hat{W}_i = [\hat{W}_{i1}, \dots, \hat{W}_{i4}]^\top$ for the DR-RL method, and the i th subsystem's 10-D estimated NN weight $\hat{W}_i = [\hat{W}_{i1}, \dots, \hat{W}_{i10}]^\top$ for the DF-RL method, $i = 1, 2$.

same as the ones used for KUKA iiwa robot manipulator and the quadrotor. The sampling rate is chosen as 1kHz. The parameter settings for the DR-RL and DF-RL approaches are referred to in Table III in the Appendix F.

A broader spatial variance during the initial learning period is observed for the DF-RL approach in the top four subfigures in Fig. 3. This is undesirable for hardware deployments. While our developed DR-RL approach explores a smaller set of states for the online learning process and converges faster. This is because the guidance from the incremental base policy helps decrease the exploration space. Besides, we observe that our proposed DR-RL approach realizes higher tracking accuracy than the DF-RL approach which learns from scratch. This benefits from the residual learning mechanism. We observe from the bottom four subfigures in Fig. 3 that the weight trajectories of our developed DR-RL approach converge faster than the DF-RL approach. The faster convergence rate implies less required data for learning one satisfying policy. This validates the improved sample efficiency brought by the residual formulation and the MBRL mechanism utilized in our work.

To validate the enhanced task flexibility of our proposed DR-RL method over the baseline RL-based tracking control strategy [18] developed under an augmented system, we focuses on a complex robotic manipulation task composed of trajectories represented as different trajectory dynamics. The detailed results are referred to in Appendix G, which is omitted due to page limits.

VII. EXPERIMENTAL VALIDATION

This section experimentally validates the efficiency of our proposed DR-RL based tracking control policy on a 3-DoF robot manipulator (see Fig. 4). The detailed hardware information is provided in our previous work [25].

During the experiment, the robot manipulator is driven to track the desired piecewise trajectory $x_d = [q_d^\top, \dot{q}_d^\top]^\top \in \mathbb{R}^6$ with $q_d = (1 + \sin(\frac{t}{2} - \frac{\pi}{2}))k_{p2} \in \mathbb{R}^3$, where $k_{p2} = [0.3, 0.6, 1]^\top$ for $t \in [0, 5]$, and $k_{p2} = [0.2, 0.5, 0.8]^\top$ for $t \in [5, 10]$. The sampling rate is set as 1kHz. Note that neither DF-RL nor AS-RL based tracking control policy is intractable

to complete the tracking task provided here. This is because it is nontrivial to find the high-dimensional activation function required for accurate value function approximation of DF-RL and AS-RL methods. Even though a high-dimensional activation function is available, the realtime performance of the corresponding weight update law is poor for practical experiments.

Regarding the incremental residual policy, we choose the 4-D activation function $\Phi_i(e_i) = [e_{i1}^2, e_{i2}^2, e_{i1}e_{i2}, e_{i2}^3]^\top$ (same as the one for the 7-DoF KUKA iiwa robot manipulator) for the value function approximation of the i th subsystem, $i = 1, 2, 3$. The utilized low-dimensional activation function $\Phi_i(e_i)$ in a fixed structure exemplifies our method's scalability and practicability towards different robotic systems. The parameters for subsystems 1-3 are set as: $Q_i = \text{diag}[300, 40000]$, $\bar{c}_i = 200$, $\Gamma_i = \text{diag}[100, 4, 0.1, 16]$, $k_{ti} = 0.2$, $k_{ei} = 0.01$, $P_i = 10$, $k_i = \text{diag}[8, 8]$, $i = 1, 2, 3$; and $\beta = 0.1$, $\bar{g}_1 = 40$, $\bar{g}_2 = 46$, and $\bar{g}_3 = 54$.

The trajectories of e_i , $i = 1, 2, 3$ under different payloads (installed to the end effector of the robot manipulator) are displayed in the top three subfigures in Fig. 4. It is shown that our developed tracking control scheme efficiently tracks the desired trajectories with satisfying tracking precision and robustness against varying payloads. Three subsystems' $\hat{W}_i$ of the 500 g payload case, which are trained in parallel using realtime and experience data together, are displayed in the bottom three subfigures in Fig. 4. We obtain the desired weight convergence for each subsystem. This validates the realtime performance of our developed weight update law (27) and the efficiency of the parallel learning mechanism.

VIII. CONCLUSION

This work develops a sample-efficient and scalable DR-RL method applicable to high-dimensional robotic tracking control tasks. The sample efficiency is improved by guiding the learning process with expert knowledge or human demonstration, and using both implicit model information and off-policy experience data for the learning process. The scalability towards high-dimensional robots is through the parallel learning architecture,

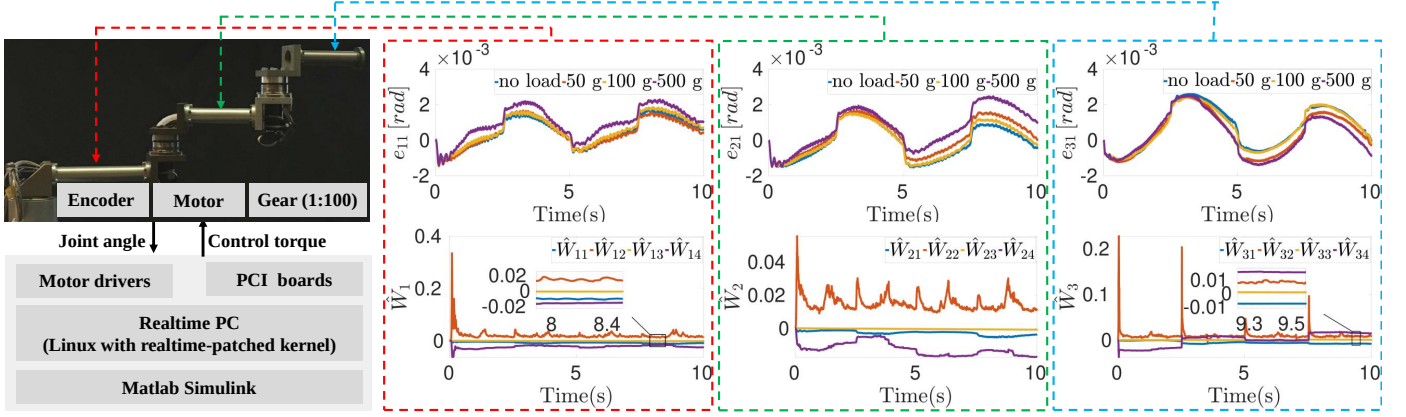


Fig. 4: The experimental validation results about the 3-DoF robot manipulator. Top: the evolution trajectories of the i th subsystem's tracking error e_{i1} under different payloads, $i = 1, 2, 3$; Bottom: the evolution trajectories of the i th subsystem's 4-D estimated weight $\hat{W}_i = [\hat{W}_{i1}, \dots, \hat{W}_{i4}]^T$, $i = 1, 2, 3$ for the 500 g payload case.

wherein critic agents learn in parallel to jointly solve the task. Comparative numerical and experimental results validate the effectiveness of our proposed DR-RL approach. In the future, the proposed method remains to be further improved from the following perspectives: addressing the input saturation problem, extending to control nonaffine systems, investigating estimation methods for indirectly measurable states, and online learning of the gain matrix $\bar{g}_i$ in (8).

REFERENCES

- [1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [2] S. Kim, D. D. Kim, and B. W. Anthony, "Dynamic control of a fiber manufacturing process using deep reinforcement learning," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 2, pp. 1128–1137, 2021.
- [3] R. Ma, Y. Wang, C. Tang, S. Wang, and R. Wang, "Position and attitude tracking control of a biomimetic underwater vehicle via deep reinforcement learning," *IEEE/ASME Transactions on Mechatronics*, 2023.
- [4] R. S. Sutton, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM Sigart Bulletin*, vol. 2, no. 4, pp. 160–163, 1991.
- [5] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *Journal of Intelligent & Robotic Systems*, vol. 86, no. 2, pp. 153–173, 2017.
- [6] B. Eysenbach, A. Khazatsky, S. Levine, and R. R. Salakhutdinov, "Mismatched no more: Joint model-policy optimization for model-based rl," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 230–23 243, 2022.
- [7] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," *Advances in neural information processing systems*, vol. 32, 2019.
- [8] Z. Yao, X. Liang, G.-P. Jiang, and J. Yao, "Model-based reinforcement learning control of electrohydraulic position servo systems," *IEEE/ASME Transactions on Mechatronics*, 2022.
- [9] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, "Gaussian processes for data-efficient learning in robotics and control," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 408–423, 2013.
- [10] M. Deisenroth and C. E. Rasmussen, "Pilco: A model-based and data-efficient approach to policy search," in *Proceedings of the 28th International Conference on machine learning (ICML-11)*, 2011, pp. 465–472.
- [11] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine, "Residual reinforcement learning for robot control," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6023–6029.
- [12] T. Silver, K. Allen, J. Tenenbaum, and L. Kaelbling, "Residual policy learning," *arXiv preprint arXiv:1812.06298*, 2018.
- [13] M. Alakuijala, G. Dulac-Arnold, J. Mairal, J. Ponce, and C. Schmid, "Residual reinforcement learning from demonstrations," *arXiv preprint arXiv:2106.08050*, 2021.
- [14] P. Kulkarni, J. Kober, R. Babuška, and C. Della Santina, "Learning assembly tasks in a few minutes by combining impedance control and residual recurrent reinforcement learning," *Advanced Intelligent Systems*, vol. 4, no. 1, p. 2100095, 2022.
- [15] K. Rana, M. Xu, B. Tidd, M. Milford, and N. Sünderhauf, "Residual skill policies: Learning an adaptable skill-based action space for reinforcement learning for robotics," in *Conference on Robot Learning*. PMLR, 2023, pp. 2095–2104.
- [16] R. Zhang, J. Hou, G. Chen, Z. Li, J. Chen, and A. Knoll, "Residual policy learning facilitates efficient model-free autonomous racing," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 625–11 632, 2022.
- [17] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, 2014.
- [18] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, 2015.
- [19] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2226–2236, 2011.
- [20] A. Plaata, W. Kusters, and M. Preuss, "Deep model-based reinforcement learning for high-dimensional problems, a survey," *arXiv preprint arXiv:2008.05598*, 2020.
- [21] R. Kamalapurkar, J. A. Rosenfeld, and W. E. Dixon, "Efficient model-based reinforcement learning for approximate online optimal control," *Automatica*, vol. 74, pp. 247–258, 2016.
- [22] W. Zhao, H. Liu, and F. L. Lewis, "Robust formation control for cooperative underactuated quadrotors via reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [23] T. Yang, N. Sun, H. Chen, and Y. Fang, "Adaptive optimal motion control of uncertain underactuated mechatronic systems with actuator constraints," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 1, pp. 210–222, 2022.
- [24] T. S. Hsia, "A new technique for robust control of servo systems," *IEEE Transactions on Industrial Electronics*, vol. 36, no. 1, pp. 1–7, 1989.
- [25] C. Li, F. Liu, Y. Wang, and M. Buss, "Concurrent learning-based adaptive control of an uncertain robot manipulator with guaranteed safety and performance," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.
- [26] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 3, pp. 753–758, 2016.
- [27] Y. Liu, W. Sun, and H. Gao, "High precision robust control for periodic tasks of linear motor via b-spline wavelet neural network observer," *IEEE Transactions on Industrial Electronics*, 2021.
- [28] T. Beckers, J. Umlauf, D. Kulic, and S. Hirche, "Stable gaussian process based tracking control of lagrangian systems," in *2017 IEEE 56th Annual*

Conference on Decision and Control (CDC). IEEE, 2017, pp. 5180–5185.

- [29] J. Na, G. Herrmann, and K. G. Vamvoudakis, “Adaptive optimal observer design via approximate dynamic programming,” in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 3288–3293.
- [30] C. Li, Y. Wang, F. Liu, Q. Liu, and M. Buss, “Model-free incremental adaptive dynamic programming based approximate robust optimal regulation,” *Int J Robust Nonlinear Control*, 2021; 1–21. doi:10.1002/rnc.5964.
- [31] K. Youcef-Toumi and S.-T. Wu, “Input/output linearization using time delay control,” *Journal of dynamic systems, measurement, and control*, vol. 114, no. 1, pp. 10–19, 1992.
- [32] K. G. Vamvoudakis and F. L. Lewis, “Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem,” *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.
- [33] X. Yang, D. Liu, H. Ma, and Y. Xu, “Online approximate solution of hji equation for unknown constrained-input nonlinear continuous-time systems,” *Information Sciences*, vol. 328, pp. 435–454, 2016.

APPENDIX A PROOF OF THEOREM 1

Proof. Considering that $V_i^*(e_i = 0) = 0$, and $V_i^* > 0$ for $\forall e_i \neq 0$, V_i^* in (18) could serve as a candidate Lyapunov function. Taking time derivative of V_i^* along the i th incremental error subsystem (14) yields

$$\dot{V}_i^* = \nabla V_i^*(\bar{f}_i + \bar{g}_i \Delta u_{i_r}^*) + \nabla V_i^* \bar{g}_i \xi_i. \quad (30)$$

According to (19) and (20), the following equations establish

$$\begin{aligned} \nabla V_i^*(\bar{f}_i + \bar{g}_i \Delta u_{i_r}^*) &= -e_i^\top Q_i e_i - W_i(\Delta u_{i_r}^*) - \bar{\xi}_{oi}^2 \\ \nabla V_i^* \bar{g}_i &= -2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta). \end{aligned} \quad (31)$$

Substituting (31) into (30) yields

$$\dot{V}_i^* = -e_i^\top Q_i e_i - W_i(\Delta u_{i_r}^*) - \bar{\xi}_{oi}^2 - 2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) \xi_i. \quad (32)$$

As for the $W_i(\Delta u_{i_r}^*)$ in (32), according to our previous result [30, Theorem 1], it follows that

$$W_i(\Delta u_{i_r}^*) = \beta^2 \sum_{j=1}^m (\tanh^{-1}(\Delta u_{i_r}^*/\beta))^2 - \epsilon_{ui}, \quad (33)$$

where $\epsilon_{ui} \leq \frac{1}{2} \bar{g}_i^2 \nabla V_i^* \nabla V_i^*$. Given that there exists $b_{\nabla V_i^*} \in \mathbb{R}^+$ such that $\|\nabla V_i^*\| \leq b_{\nabla V_i^*}$. Thus, we could rewrite the bound of ϵ_{ui} as $\epsilon_{ui} \leq b_{\epsilon_{ui}} \leq \frac{1}{2} \bar{g}_i^2 b_{\nabla V_i^*}^2$.

Then, substituting (33) into (32), we get

$$\begin{aligned} \dot{V}_i^* &= -e_i^\top Q_i e_i - [\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) + \xi_i]^2 \\ &\quad - (\bar{\xi}_{oi}^2 - \xi_i^\top \xi_i) + b_{\epsilon_{ui}}. \end{aligned} \quad (34)$$

We choose $\bar{\xi}_{oi} = \bar{c}_i \|\Delta u_{i_r}\|$, and $\bar{c}_i$ is picked to satisfy $\bar{c}_i \|\Delta u_{i_r}\| > \xi_i$, where ξ_i is defined in Lemma 1. Then, the following equation holds

$$\dot{V}_i^* \leq -e_i^\top Q_i e_i + b_{\epsilon_{ui}}. \quad (35)$$

Thus, if $-\lambda_{\min}(Q_i) \|e_i\|^2 + b_{\epsilon_{ui}} < 0$, $\dot{V}_i^* < 0$ holds. Here $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalues of a symmetric real matrix. Finally, it concludes that states of the i th incremental error subsystem (14) converges to the residual set

$$\Omega_{e_i} = \{e_i \mid \|e_i\| \leq \sqrt{b_{\epsilon_{ui}}/\lambda_{\min}(Q_i)}\}. \quad (36)$$

This concludes the proof. $\square$

APPENDIX B PROOF OF THEOREM 3

Proof. Consider the candidate Lyapunov function for the i th incremental error subsystem (14) as

$$L_i = V_i^* + \frac{1}{2} \tilde{W}_i^\top \Gamma_i^{-1} \tilde{W}_i. \quad (37)$$

By denoting $L_{i_1} = V_i^*$, its derivative follows

$$\begin{aligned} \dot{L}_{i_1} &= \nabla V_i^* (\bar{f}_i + \bar{g}_i \Delta \hat{u}_{i_r} + \bar{g}_i \xi_i) \\ &= \nabla V_i^* (\bar{f}_i + \bar{g}_i \Delta u_{i_r}^*) + \nabla V_i^* \bar{g}_i \xi_i \\ &\quad + \nabla V_i^* \bar{g}_i (\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*). \end{aligned} \quad (38)$$

Substituting (31) into (38) reads

$$\begin{aligned} \dot{L}_{i_1} &= -e_i^\top Q_i e_i - W(\Delta u_{i_r}^*) - \bar{\xi}_{oi}^2 - 2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) \xi_i \\ &\quad - 2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) (\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*). \end{aligned} \quad (39)$$

Combining with (33) and (34), (39) follows

$$\begin{aligned} \dot{L}_{i_1} &\leq -e_i^\top Q_i e_i - (\bar{\xi}_{oi}^2 - \|\xi_i\|^2) - [\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) + \xi_i]^2 \\ &\quad + \frac{1}{2} \nabla V_i^* \bar{g}_i \bar{g}_i^\top \nabla V_i^* - 2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) (\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*). \end{aligned} \quad (40)$$

The term $-2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) (\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*)$ in (40) follows

$$\begin{aligned} -2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) (\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*) &\leq \beta^2 \|\tanh^{-1}(\Delta u_{i_r}^*/\beta)\|^2 \\ &\quad + \|\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*\|^2. \end{aligned} \quad (41)$$

According to (21) and (23), and the mean-value theorem, the optimal incremental control is rewritten as

$$\Delta u_{i_r}^* = -\beta \tanh\left(\frac{1}{2\beta} \bar{g}_i^\top \nabla \Phi_i^\top W_i^*\right) - \epsilon_{\Delta u_i^*}, \quad (42)$$

where $\epsilon_{\Delta u_i^*} = \frac{1}{2}(1 - \tanh^2(\eta_i)) \bar{g}_i^\top \nabla \epsilon_i$, and $\eta_i \in \mathbb{R}$ is chosen between $\frac{1}{2\beta} \bar{g}_i^\top \nabla \Phi_i^\top W_i^*$ and $\frac{1}{2\beta} \bar{g}_i^\top \nabla V_i^*$. According to $\|\nabla \epsilon_i\| \leq b_{\epsilon_{ei}}$ in Assumption 1, $\|\epsilon_{\Delta u_i^*}\| \leq \frac{1}{2} \|\bar{g}_i\| b_{\epsilon_{ei}}$ holds. Then, by combining (28) with (42), we get

$$\Delta \hat{u}_{i_r} - \Delta u_{i_r}^* = \beta (\tanh(\mathcal{G}_i^*) - \tanh(\hat{\mathcal{G}}_i)) + \epsilon_{\Delta u_i^*}. \quad (43)$$

where $\mathcal{G}_i^* = \frac{1}{2\beta} \bar{g}_i^\top \nabla \Phi_i^\top W_i^*$, and $\hat{\mathcal{G}}_i = \frac{1}{2\beta} \bar{g}_i^\top \nabla \Phi_i^\top \dot{W}$. Based on (21) and (28), the Taylor series of $\tanh(\mathcal{G}_i^*)$ follows

$$\begin{aligned} \tanh(\mathcal{G}_i^*) &= \tanh(\hat{\mathcal{G}}_i) + \frac{\partial \tanh(\hat{\mathcal{G}}_i)}{\partial \hat{\mathcal{G}}_i} (\mathcal{G}_i^* - \hat{\mathcal{G}}_i) + \mathcal{O}((\mathcal{G}_i^* - \hat{\mathcal{G}}_i)^2) \\ &= \tanh(\hat{\mathcal{G}}_i) - \frac{1}{2\beta} (1 - \tanh^2(\hat{\mathcal{G}}_i)) \bar{g}_i^\top \nabla \Phi_i^\top \tilde{W}_i \\ &\quad + \mathcal{O}((\mathcal{G}_i^* - \hat{\mathcal{G}}_i)^2), \end{aligned} \quad (44)$$

where $\mathcal{O}((\mathcal{G}_i^* - \hat{\mathcal{G}}_i)^2)$ is a higher order term of the Taylor series. By following [33, Lemma 1], this higher order term is bounded as

$$\|\mathcal{O}((\mathcal{G}_i^* - \hat{\mathcal{G}}_i)^2)\| \leq 2 + \frac{1}{\beta} \|\bar{g}_i\| b_{\Phi_{ei}} \|\tilde{W}_i\|. \quad (45)$$

Based on (44), we rewrite (43) as

$$\begin{aligned} \Delta \hat{u}_{i_r} - \Delta u_{i_r}^* &= \beta (\tanh(\mathcal{G}_i^*) - \tanh(\hat{\mathcal{G}}_i)) + \epsilon_{\Delta u_i^*} \\ &= -\frac{1}{2} (1 - \tanh^2(\hat{\mathcal{G}}_i)) \bar{g}_i^\top \nabla \Phi_i^\top \tilde{W}_i \\ &\quad + \beta \mathcal{O}((\mathcal{G}_i^* - \hat{\mathcal{G}}_i)^2) + \epsilon_{\Delta u_i^*}. \end{aligned} \quad (46)$$

Then, by combining (45) with (46), and given that $\|1 - \tanh^2(\hat{\mathcal{G}}_i)\| \leq 2$, $\|\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*\|^2$ in (41) follows

$$\begin{aligned} \|\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*\|^2 &\leq 3\beta^2 \left\| \mathcal{O}((\mathcal{G}_i^* - \hat{\mathcal{G}}_i)^2) \right\|^2 + 3 \|\epsilon_{\Delta u_i^*}\|^2 \\ &\quad + 3 \left\| -\frac{1}{2}(1 - \tanh^2(\hat{\mathcal{G}}_i)) \bar{g}_i^\top \nabla \Phi_i^\top \tilde{W}_i \right\|^2 \\ &\leq 6 \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 \|\tilde{W}_i\|^2 + 12\beta^2 + \frac{3}{4} \|\bar{g}_i\|^2 b_{\epsilon_{ei}}^2 \\ &\quad + 12\beta \|\bar{g}_i\| b_{\Phi_{ei}} \|\tilde{W}_i\|. \end{aligned} \quad (47)$$

Based on (23), (31), Assumption 1, and the fact that $\|W_i^*\| \leq b_{W_i^*}$, $\|\tanh^{-1}(\Delta u_{i_r}^*/\beta)\|^2$ in (41) follows

$$\begin{aligned} \|\tanh^{-1}(\Delta u_{i_r}^*/\beta)\|^2 &= \left\| \frac{1}{4\beta^2} \nabla V_i^{*\top} \bar{g}_i \bar{g}_i^\top \nabla V_i^* \right\|^2 \\ &\leq \frac{1}{4\beta^2} \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 b_{W_i^*}^2 + \frac{1}{4\beta^2} b_{\epsilon_{ei}}^2 \|\bar{g}_i\|^2 \\ &\quad + \frac{1}{2\beta^2} \|\bar{g}_i\|^2 b_{\Phi_{ei}} b_{\epsilon_{ei}} b_{W_i^*}. \end{aligned} \quad (48)$$

Using (47) and (48), (41) reads

$$\begin{aligned} -2\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) (\Delta \hat{u}_{i_r} - \Delta u_{i_r}^*) &\leq \frac{1}{4} \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 b_{W_i^*}^2 \\ &\quad + \frac{1}{4} b_{\epsilon_{ei}}^2 \|\bar{g}_i\|^2 + \frac{1}{2} \|\bar{g}_i\|^2 b_{\Phi_{ei}} b_{\epsilon_{ei}} b_{W_i^*} + 6 \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 \|\tilde{W}_i\|^2 \\ &\quad + 12\beta^2 + \frac{3}{4} \|\bar{g}_i\|^2 b_{\epsilon_{ei}}^2 + 12\beta \|\bar{g}_i\| b_{\Phi_{ei}} \|\tilde{W}_i\|. \end{aligned} \quad (49)$$

Substituting (49) into (40), finally the first term $\dot{L}_{i_1}$ follows

$$\begin{aligned} \dot{L}_{i_1} &\leq -e_i^\top Q_i e_i - (\bar{\xi}_{oi}^2 - \xi_i^\top \xi_i) - [\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) + \xi_i]^\top \\ &\quad + \frac{3}{4} \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 b_{W_i^*}^2 + \frac{3}{4} b_{\epsilon_{ei}}^2 \|\bar{g}_i\|^2 + \frac{3}{2} \|\bar{g}_i\|^2 b_{\Phi_{ei}} b_{\epsilon_{ei}} b_{W_i^*} \\ &\quad + 6 \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 \|\tilde{W}_i\|^2 + 12\beta^2 + \frac{3}{4} \|\bar{g}_i\|^2 b_{\epsilon_{ei}}^2 \\ &\quad + 12\beta \|\bar{g}_i\| b_{\Phi_{ei}} \|\tilde{W}_i\|. \end{aligned} \quad (50)$$

As for the second term $\dot{L}_W = \frac{1}{2} \tilde{W}_i^\top \Gamma_i^{-1} \tilde{W}_i$, based on (27) and Theorem 1 in our previous work [30], it follows

$$\dot{L}_{i_2} \leq -\tilde{W}_i^\top \mathcal{Y}_i \tilde{W}_i + \tilde{W}_i^\top \epsilon_{\tilde{W}_i}. \quad (51)$$

where $\mathcal{Y}_i = \sum_{l=1}^{P_i} k_{ei} Y_{il} Y_{il}^\top \in \mathbb{R}^{N_i \times N_i}$, and $\epsilon_{\tilde{W}_i} = -k_{ti} Y_i \epsilon_{h_i} - \sum_{l=1}^{P_i} k_{ei} Y_{il} \epsilon_{h_{il}} \in \mathbb{R}^{N_i}$. The boundness of Y_i and ϵ_{h_i} results in bounded $\epsilon_{\tilde{W}_i}$. Thus, there exists $\bar{\epsilon}_{\tilde{W}_i} \in \mathbb{R}^+$ such that $\|\epsilon_{\tilde{W}_i}\| \leq \bar{\epsilon}_{\tilde{W}_i}$. According to Assumption 2, $\mathcal{Y}_i$ is positive definite. Thus, (51) could be rewritten as

$$\dot{L}_{i_2} \leq -\lambda_{\min}(\mathcal{Y}_i) \|\tilde{W}_i\|^2 - \bar{\epsilon}_{\tilde{W}_i} \|\tilde{W}_i\|. \quad (52)$$

Finally, as for $\dot{L}_i$, substituting (50) and (51) into (37), we get

$$\dot{L}_i \leq -\mathcal{A}_i - \mathcal{B}_i \|\tilde{W}_i\|^2 + \mathcal{C}_i \|\tilde{W}_i\| + \mathcal{D}_i, \quad (53)$$

where $\mathcal{A}_i = e_i^\top Q_i e_i + (\bar{\xi}_{oi}^2 - \xi_i^\top \xi_i) + [\beta \tanh^{-1}(\Delta u_{i_r}^*/\beta) + \xi_i]^\top$, $\mathcal{B}_i = \lambda_{\min}(\mathcal{Y}_i) - 6 \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2$, $\mathcal{C}_i = 12\beta \|\bar{g}_i\| b_{\Phi_{ei}} + \bar{\epsilon}_{\tilde{W}_i}$, and $\mathcal{D}_i = \frac{3}{4} \|\bar{g}_i\|^2 b_{\Phi_{ei}}^2 b_{W_i^*}^2 + \frac{3}{2} b_{\epsilon_{ei}}^2 \|\bar{g}_i\|^2 + \frac{3}{2} \|\bar{g}_i\|^2 b_{\Phi_{ei}} b_{\epsilon_{ei}} b_{W_i^*} + 12\beta^2$. Let the parameters

be chosen such that $\mathcal{B}_i > 0$. Since $\mathcal{A}_i$ is positive definite, the above Lyapunov derivative (53) is negative if

$$\|\tilde{W}_i\| > \frac{\mathcal{C}_i}{2\mathcal{B}_i} + \sqrt{\frac{\mathcal{C}_i^2}{4\mathcal{B}_i^2} + \frac{\mathcal{D}_i}{\mathcal{B}_i}}. \quad (54)$$

Thus, the weight learning error of the critic agent converges to the residual set

$$\tilde{\Omega}_{\tilde{W}_i} = \left\{ \tilde{W}_i \mid \|\tilde{W}_i\| \leq \frac{\mathcal{C}_i}{2\mathcal{B}_i} + \sqrt{\frac{\mathcal{C}_i^2}{4\mathcal{B}_i^2} + \frac{\mathcal{D}_i}{\mathcal{B}_i}} \right\}. \quad (55)$$

This completes the proof. $\square$

APPENDIX C PROOF OF LEMMA 1

Proof. Combing (6) with (7), the estimation error for the i th subsystem (8) follows

$$\begin{aligned} \xi_i &= h_i - \hat{h}_i = h_i - h_{i,0} \\ &= (\bar{g}_i^{-1} - g_i^{-1}) \Delta \dot{x}_i + (g_{i,0}^{-1} - g_i^{-1}) \dot{x}_{i,0} \\ &\quad + g^{-1}(f_i - f_{i,0}) + (g_i^{-1} - g_{i,0}^{-1}) f_{i,0}, \end{aligned} \quad (56)$$

where $\Delta \dot{x}_i = \dot{x}_i - \dot{x}_{i,0}$. Based on (2), (8) and (9), an equivalent form of $\Delta \dot{x}_i$ follows

$$\begin{aligned} \Delta \dot{x}_i &= f_i + g_i u_i - f_{i,0} - g_{i,0} u_{i,0} \\ &= g_i \Delta u_i + (g_i - g_{i,0}) u_{i,0} + f_i - f_{i,0} \\ &= g_i (\Delta u_{i_b} + \Delta u_{i_r}) + (g_i - g_{i,0}) u_{i,0} + f_i - f_{i,0}. \end{aligned} \quad (57)$$

Substituting (57) into (56), we get

$$\xi_i = (g_i \bar{g}_i^{-1} - 1) \Delta u_{i_b} + (g_i \bar{g}_i^{-1} - 1) \Delta u_{i_r} + \delta_{1i}, \quad (58)$$

where $\delta_{1i} = \bar{g}_i^{-1}(g_i - g_{i,0}) u_0 + \bar{g}_i^{-1}(f_i - f_{i,0})$.

In the following, we implement the incremental base policy as (11) in favor of theoretical completeness. For simplicity, denoting $\mu_i = \dot{x}_{d_i} - k_i e_i \in \mathbb{R}^{n_i}$. According to (7) and (11), Δu_{i_b} in (58) follows

$$\begin{aligned} \Delta u_{i_b} &= \bar{g}_i^{-1}(\mu_i - \bar{g}_i h_{i,0} - \bar{g}_i u_{i,0}) \\ &= \bar{g}_i^{-1} \mu_i - (\bar{g}_i^{-1} - g_{i,0}^{-1}) \dot{x}_{i,0} - g_{i,0}^{-1} f_{i,0} - u_{i,0} \\ &= \bar{g}_i^{-1} \mu_i - (\bar{g}_i^{-1} - g_{i,0}^{-1})(f_{i,0} + g_{i,0} u_{i,0}) - g_{i,0}^{-1} f_{i,0} - u_{i,0} \\ &= \bar{g}_i^{-1} \mu_i - \bar{g}_i^{-1}(f_{i,0} + g_{i,0} u_{i,0}) \\ &= \bar{g}_i^{-1}(\mu_i - \mu_{i,0}) - \bar{g}_i^{-1}(\dot{x}_{i,0} - \mu_{i,0}), \end{aligned} \quad (59)$$

where $\mu_{i,0} = \dot{x}_{d_{i,0}} - k_i e_{i,0}$. Besides, combining (8) with (9), we get

$$\begin{aligned} \dot{x}_i &= \dot{x}_{i,0} + \bar{g}_i (\Delta u_{i_b} + \Delta u_{i_r}) + \bar{g}_i \xi_i \\ &= \dot{x}_{i,0} + \bar{g}_i \bar{g}_i^{-1}(\mu_i - \dot{x}_{i,0}) + \bar{g}_i \Delta u_{i_r} + \bar{g}_i \xi_i \\ &= \mu_i + \bar{g}_i \Delta u_{i_r} + \bar{g}_i \xi_i. \end{aligned} \quad (60)$$

Based on the result shown in (60), we get

$$\xi_i = \bar{g}_i^{-1}(\dot{x}_i - \mu_i - \bar{g}_i \Delta u_{i_r}). \quad (61)$$

Accordingly, the following equation establishes

$$\xi_{i,0} = \bar{g}_i^{-1}(\dot{x}_{i,0} - \mu_{i,0} - \bar{g}_i \Delta u_{i_b,0}). \quad (62)$$

Based on the result given in (62), (59) is rewritten as

$$\begin{aligned}\Delta u_{ib} &= \bar{g}_i^{-1}(\mu_i - \mu_{i,0}) - \bar{g}_i^{-1}(\dot{x}_{i,0} - \mu_{i,0} \\ &\quad - \bar{g}_i \Delta u_{ib,0}) - \Delta u_{ib,0} \\ &= \bar{g}_i^{-1}(\mu_i - \mu_{i,0}) - \xi_{i,0} - \Delta u_{ib,0}.\end{aligned}\quad (63)$$

Substituting (63) into (58) yields

$$\begin{aligned}\xi_i &= (1 - g_i \bar{g}_i^{-1})\xi_{i,0} + (1 - g_i \bar{g}_i^{-1})\bar{g}_i^{-1}(\mu_{i,0} - \mu_i) \\ &\quad + (1 - g_i \bar{g}_i^{-1})(\Delta u_{ib,0} - \Delta u_{i,r}) + \delta_{1i}.\end{aligned}\quad (64)$$

In discrete-time domain, (64) can be represented as

$$\begin{aligned}\xi_i(k) &= (1 - g_i(k)\bar{g}_i^{-1})\xi_i(k-1) + (1 - g_i(k)\bar{g}_i^{-1})\Delta \tilde{u}_{ib} \\ &\quad + \delta_{1i} + \delta_{2i},\end{aligned}\quad (65)$$

where $\Delta \tilde{u}_{i,r} = \Delta u_{i,r}(k-1) - \Delta u_{i,r}(k)$, $\delta_{2i} = (1 - g_i(k)\bar{g}_i^{-1})\bar{g}_i^{-1}(\mu_i(k-1) - \mu_i(k))$.

The constrained input $\|\Delta u_{i,r}(k)\| \leq \beta$ implies that the following equation holds

$$\|\Delta \tilde{u}_{i,r}\| \leq \|\Delta u_{i,r}(k-1)\| + \|\Delta u_{i,r}(k)\| \leq 2\beta. \quad (66)$$

We choose the value of $\bar{g}_i$ to meet $\|1 - g_i(k)\bar{g}_i^{-1}\| \leq \iota_i < 1$, where $\iota_i \in \mathbb{R}^+$. Under a sufficiently high sampling rate, it is reasonable to assume that there exists $\bar{\delta}_{1i}, \bar{\delta}_{2i} \in \mathbb{R}^+$ such that $\|\delta_{1i}\| \leq \bar{\delta}_{1i}$, and $\|\delta_{2i}\| \leq \iota_i \bar{\delta}_{2i}$. Then, the following equations hold:

$$\begin{aligned}\|\xi_i(k)\| &\leq \iota_i \|\xi_i(k-1)\| + \iota_i \|\Delta \tilde{u}_{i,r}\| + \bar{\delta}_{1i} + \iota_i \bar{\delta}_{2i} \\ &\leq \iota_i^2 \|\xi_i(k-2)\| + (\iota_i^2 + \iota_i) \|\Delta \tilde{u}_{i,r}\| \\ &\quad + (\iota_i + 1)(\bar{\delta}_{1i} + \iota_i \bar{\delta}_{2i}) \\ &\leq \dots \\ &\leq \iota_i^k \|\xi_i(0)\| + \frac{\bar{\delta}_{1i} + \iota_i \bar{\delta}_{2i}}{1 - \iota_i} + \frac{\iota_i \|\Delta \tilde{u}_{i,r}\|}{1 - \iota_i} \\ &\leq \iota_i^k \|\xi_i(0)\| + \frac{\bar{\delta}_{1i} + \iota_i \bar{\delta}_{2i}}{1 - \iota_i} + \frac{2\iota_i \beta}{1 - \iota_i} := \bar{\xi}_i.\end{aligned}\quad (67)$$

As $k \rightarrow \infty$, $\bar{\xi}_i \rightarrow \frac{\bar{\delta}_{1i} + \iota_i \bar{\delta}_{2i}}{1 - \iota_i} + \frac{2\iota_i \beta}{1 - \iota_i}$. This concludes the proof. $\square$

APPENDIX D

SETTINGS FOR 7-DOF ROBOT MANIPULATOR

The parameter settings for the joint-space tracking task of the 7-DoF KUKA iiwa robot manipulator is presented in Table I. The associated simulation results of the 7 subsystems are presented in Fig. 5, which fully validate the superiority of our proposed DR-RL method.

TABLE I: The parameter settings for KUKA iiwa.

Initial value conditions	Incremental Base Policy (IBP)	DR-RL
	$q(0) = [0, 0, 0, 0, 0, 0, 0]^T$, $\dot{q}(0) = [0, 0, 0, 0, 0, 0, 0]^T$, $u(0) = [0, 0, 0, 0, 0, 0, 0]^T$.	$q(0) = [0, 0, 0, 0, 0, 0, 0]^T$, $\dot{q}(0) = [0, 0, 0, 0, 0, 0, 0]^T$, $u(0) = [0, 0, 0, 0, 0, 0, 0]^T$, $\dot{W}(0) = 0_{4 \times 1}, i = 1, 2, \dots, 7$.
parameters	$g_1 = 200, g_2 = 80, g_3 = 100, g_4 = 800,$ $g_5 = 800, g_6 = 800, g_7 = 800,$ $k_1 = \text{diag}[1, 1], k_2 = \text{diag}[0.1, 0.1],$ $k_3 = \text{diag}[1, 1], k_4 = \text{diag}[0.1, 0.1],$ $k_5 = \text{diag}[1, 1], k_6 = \text{diag}[0.1, 0.1],$ $k_7 = \text{diag}[1, 1],$ $\Gamma_1 = \text{diag}[10, 10, 10, 10], \Gamma_2 = \text{diag}[1, 1, 1, 1],$ $\Gamma_3 = \text{diag}[10^3, 10^3, 10^3, 10^3], \Gamma_4 = \text{diag}[1, 1, 1, 1],$ $\Gamma_5 = \text{diag}[10, 10, 10, 10], \Gamma_6 = \text{diag}[10, 10, 10, 10],$ $\Gamma_7 = \text{diag}[10^7, 10^7, 10^7, 10^7],$	$g_1 = 200, g_2 = 80, g_3 = 100, g_4 = 800,$ $g_5 = 800, g_6 = 800, g_7 = 800,$ $k_1 = \text{diag}[1, 1], k_2 = \text{diag}[0.1, 0.1],$ $k_3 = \text{diag}[1, 1], k_4 = \text{diag}[0.1, 0.1],$ $k_5 = \text{diag}[1, 1], k_6 = \text{diag}[0.1, 0.1],$ $k_7 = \text{diag}[1, 1],$ $\Gamma_1 = \text{diag}[10, 10, 10, 10], \Gamma_2 = \text{diag}[1, 1, 1, 1],$ $\Gamma_3 = \text{diag}[10^3, 10^3, 10^3, 10^3], \Gamma_4 = \text{diag}[1, 1, 1, 1],$ $\Gamma_5 = \text{diag}[10, 10, 10, 10], \Gamma_6 = \text{diag}[10, 10, 10, 10],$ $\Gamma_7 = \text{diag}[10^7, 10^7, 10^7, 10^7],$ $Q_1 = \text{diag}[1, 1], Q_2 = \text{diag}[2000, 2000],$ $Q_3 = \text{diag}[10, 10], Q_4 = \text{diag}[20, 20],$ $Q_5 = \text{diag}[1, 1], Q_6 = \text{diag}[20, 20],$ $Q_7 = \text{diag}[1, 1],$ $\beta = 1, P_i = 10, c_i = 2, k_{t_i} = 1, k_{e_i} = 1, i = 1, 2, \dots, 7.$

APPENDIX E

SETTINGS FOR 6-DOF QUADROTOR

This section further certifies the effectiveness of DR-RL based tracking control policy under a high-dimensional quadrotor tracking task. The quadrotor is driven to track the desired spiral reference trajectory $x_r = [\frac{3}{10 \sin(t)}, \cos(t), \frac{t}{10\pi}, 0]^T \in \mathbb{R}^3$, $t \in [0, 50]$. The associated parameter settings to conduct numerical simulations are referred to Table II. The detailed procedures to decouple the 6-DoF quadrotor into 6 subsystems (2) are referred to Section H. For subsystem 1-6, we adopt the same activation functions as the ones used for the KUKA iiwa robot manipulator case. As displayed in Fig. 6, we obtain a satisfying tracking performance via the DR-RL based tracking control policy.

TABLE II: The parameter settings of a quadrotor task.

Initial value conditions	$\xi(0) = [0.1, 1.1, 0]^T$, $\eta(0) = [0, 0, 0]^T$, $u(0) = [0, 0, 0.5]^T$, $\bar{g}_i = 300, i = 1, 2, 3; \bar{g}_i = 60000, i = 4, 5, 6,$ $k_i = \text{diag}[3, 3], \dot{W}_i(0) = 0_{4 \times 1}, i = 1, \dots, 6.$
Cost function parameters	$Q_i = \text{diag}(1, 1), \bar{c}_i = 4,$ $\beta = 0.1, i = 1, \dots, 6.$
Weight learning parameters	$k_{t_i} = 1, k_{e_i} = 0.01, P_i = 6,$ $\Gamma_i = 0.01 \text{diag}(I_{1 \times 4}), i = 1, \dots, 6.$

APPENDIX F

SETTINGS FOR 2-DOF ROBOT MANIPULATOR

The 2-DoF robot manipulator's dynamics follows [18]

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + F_d\dot{q} + F_s = \tau,$$

where $q = [q_1, q_2]^T, \dot{q} = [\dot{q}_1, \dot{q}_2]^T, \tau \in \mathbb{R}^2$; Let $c_2 = \cos q_2$, $s_2 = \sin q_2$, then $M(q) = \begin{bmatrix} a_1 + 2a_3c_2 & a_2 + a_3c_2 \\ a_2 + a_3c_2 & a_2 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$, and $C(q, \dot{q}) = \begin{bmatrix} -a_3\dot{q}_2s_2 & -a_3(\dot{q}_1 + \dot{q}_2)s_2 \\ a_3\dot{q}_1s_2 & 0 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$, wherein $a_1 = 3.473 \text{ kg m}^2$, $a_2 = 0.196 \text{ kg m}^2$, $a_3 = 0.242 \text{ kg m}^2$; The static friction follows $F_d = \text{diag}[5.3, 1.1] \text{ N m s}$, and the dynamic friction is $F_s = [8.45 \tanh(\dot{q}_1), 2.35 \tanh(\dot{q}_2)]^T \text{ N m s}$.

TABLE III: The parameter settings for the task-space task.

Initial value conditions	DF-RL [17]	DR-RL
	$x(0) = [-0.18, \frac{\pi}{2}, 0]^T$, $u(0) = [0, 0]^T$, $\dot{W}(0) = 0_{10 \times 1}$	$x(0) = [-0.18, \frac{\pi}{2}, 0]^T$, $u(0) = [0, 0]^T$, $\dot{W}_i(0) = 0_{4 \times 1}$, $\bar{g}_1 = 3.8, \bar{g}_2 = 8.3$, $k_i = \text{diag}[10, 10]$
Cost function parameters	$Q_1 = \text{diag}(10.4, 23.4)$, $Q_2 = \text{diag}(15.6, 18.2)$, $R = \text{diag}(1, 1)$	$Q_1 = \text{diag}(21, 47)$, $Q_2 = \text{diag}(30, 40)$, $R_i = \text{diag}(1, 1)$, $\beta_i = 2 \bar{c}_i = 1$
Weight learning parameters	$k_t = 0.9, k_e = 0.08$, $\Gamma = 3.2 \text{diag}(I_{1 \times 10})$, $P = 15$	$k_{t_i} = 200, k_{e_i} = 0.2$, $\Gamma_i = 0.01 \text{diag}(I_{1 \times 4})$, $P_i = 12 \ i = 1, 2.$

The 10-D activation function used in for task-space tracking task of the 2-DoF robot manipulator is :

$$\begin{aligned}\Phi_i(\sigma_i) &= [\sigma_{i1}^2, \sigma_{i2}^2, \sigma_{i3}^2, \sigma_{i4}^2, \sigma_{i1}\sigma_{i2}, \\ &\quad \sigma_{i1}\sigma_{i3}, \sigma_{i1}\sigma_{i4}, \sigma_{i2}\sigma_{i3}, \sigma_{i2}\sigma_{i4}, \sigma_{i3}\sigma_{i4}]^T,\end{aligned}$$

where $\sigma_i = [e_i^T(t), e_i^T(t - t_s)]^T \in \mathbb{R}^4$, $e_i = x_i - x_{d_i} \in \mathbb{R}^2$.

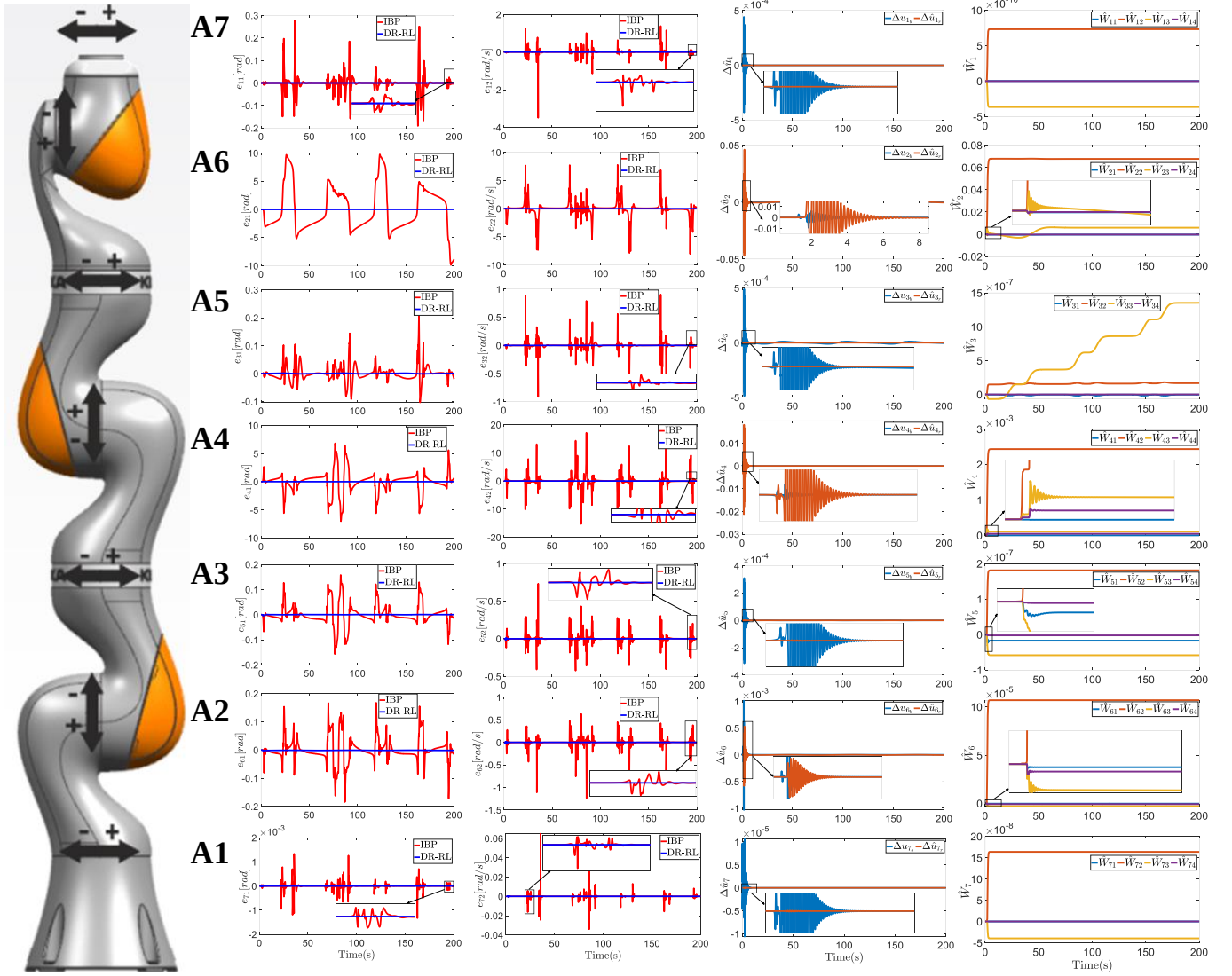


Fig. 5: The numerical validation results on the 7-DoF KUKA iiwa robot manipulator. The simulation results of the subsystem 1 to the subsystem 7 are displayed from bottom to top. In each row, each subsystem's evolution trajectories of angle error e_{i1} , angle velocity error e_{i2} , incremental control inputs Δu_{ib} and Δu_{ir} , and weight $\hat{W}_i$ are presented from left to right in sequence.

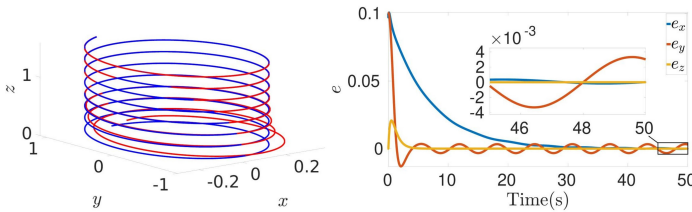


Fig. 6: The numerical validations about a 6-DoF quadrotor. Left: the quadrotor trajectory in 3-D space; Right: the evolution trajectories of the position tracking error.

APPENDIX G

COMPARISON BETWEEN DR-RL AND AS-RL METHODS

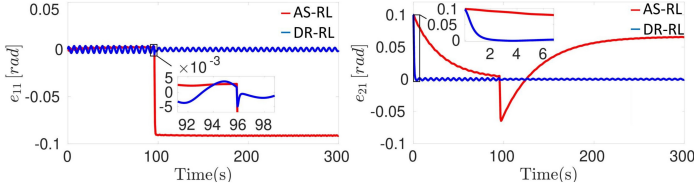
This section focuses on a complex robotic manipulation task composed of trajectories represented as different trajectory dynamics to validate the enhanced task flexibility of our

proposed DR-RL method. Note that the baseline RL-based tracking control strategy [18] developed under an augmented system (referred to as AS-RL for simplicity) would fail on the above task, as illustrated in Fig. 7. The considered task is to design a control input τ to enable the state $x = [q_1, q_2, \dot{q}_1, \dot{q}_2]^T$ to perfectly follow the desired trajectory $x_d = [k_{p1} \cos(t), k_{p2} \cos(t), -k_{p1} \sin(t), -k_{p2} \sin(t)]^T$, where $k_{p1} = 0.5$, $k_{p2} = 1$ for $t \in [0, \frac{61\pi}{2})$, and $k_{p1} = 0.25$, $k_{p2} = 0.5$ for $t \in [\frac{61\pi}{2}, 400]$. This adopted piecewise trajectory x_d violates Assumption 2 in [18] that the desired trajectory is generated from one assumed reference trajectory dynamics.

To achieve an accurate value function approximation, we implement the AS-RL method by using the following 23-D activation function provided in [18] but adopting our proposed

TABLE IV: The parameter settings for the joint-space task.

	AS-RL [18]	DR-RL
Initial value conditions	$x(0) = [0.5, 1.1, 0, 0]^\top$, $u(0) = [0, 0]^\top$, $\dot{W}(0) = 0_{23 \times 1}$, $x_d(0) = [0.5, 1, 0, 0]^\top$.	$x(0) = [0.5, 1.1, 0, 0]^\top$, $u(0) = [0, 0]^\top$, $\bar{g}_i = 7$, $\dot{W}_i(0) = 0_{4 \times 1}$, $k_i = \text{diag}[10, 10]$.
Cost function parameters	$Q = \text{diag}(600, 600, 1, 1)$, $R = \text{diag}(1, 1)$.	$Q_i = \text{diag}(600, 1)$, $\beta = 0.1$, $\bar{c}_i = 29$.
Weight learning parameters	$k_t = 80$, $k_e = 1$, $\Gamma = \text{diag}(I_{1 \times 23})$, $P = 25$,	$k_{t_i} = 80$, $k_{e_i} = 1$, $\Gamma_i = \text{diag}(I_{1 \times 4})$, $P_i = 12$, $i = 1, 2$.

Fig. 7: The evolution trajectories of the i th subsystem's the tracking error e_{i1} of AS-RL and DR-RL methods, $i = 1, 2$.

weight update law (27) to achieve a fair comparison ⁴.

$$\Phi(\rho) = \frac{1}{2}[\rho_1^2, \rho_2^2, 2\rho_1\rho_3, 2\rho_1\rho_4, 2\rho_2\rho_3, 2\rho_2\rho_4, \rho_1^2\rho_5^2, \rho_1^2\rho_6^2, \rho_1^2\rho_7^2, \rho_1^2\rho_8^2, \rho_2^2\rho_5^2, \rho_2^2\rho_6^2, \rho_2^2\rho_7^2, \rho_2^2\rho_8^2, \rho_3^2\rho_5^2, \rho_3^2\rho_6^2, \rho_3^2\rho_7^2, \rho_3^2\rho_8^2, \rho_4^2\rho_5^2, \rho_4^2\rho_6^2, \rho_4^2\rho_7^2, \rho_4^2\rho_8^2]^\top,$$

where $\rho = [e^\top, x_d^\top]^\top \in \mathbb{R}^8$, $e = x - x_d \in \mathbb{R}^4$. Our developed approach only needs the 4-D activation function same as the one used in Section D for each subsystem. The detailed parameter settings for AS-RL and DR-RL based tracking control schemes are provided in Table IV.

As displayed in Fig. 7, both AS-RL and DR-RL based tracking control policies achieve a satisfying performance in the first interval $t \in [0, \frac{61\pi}{2})$. However, when the desired trajectory changes at $\frac{61\pi}{2}$ s, the learning inefficiency of the AS-RL based tracking control policy (with fine-tuned hyperparameters) hinders us from getting a high-precision tracking performance but with steady-state errors. Our proposed tracking control policy efficiently drives the robot manipulator to track the desired trajectories with performance competitive with the model-based AS-RL method, while enjoying better scalability to value function approximation and flexibility to varying desired trajectories. Furthermore, as shown in the right part of Fig. 7, the error trajectory of DR-RL enjoys a faster convergence rate than the one of AS-RL, which is due to the guideline from the baseline policy. Note that the oscillation of the DR-RL method in Fig. 7 is the cost of keeping the parameters same as the AS-RL method. Even so, regarding the varying tasks, our proposed method under bad hyperparameters shows better performance than the well debugged AS-RL method. Note that our proposed DR-RL would behave better in other parameter settings.

⁴The weight update law proposed in [18] requires directly adding probing noise to control inputs to meet the required persistent excitation condition for the weight convergence, which causes undesirable oscillations.

APPENDIX H

THE INCREMENTAL SUBSYSTEMS OF QUADROTOR

This section presents the detailed procedures to decouple the 6-DoF quadrotor into 6 incremental subsystems. By introducing pseudo control inputs, we transform the original underactuated quadrotor model into a fully-actuated model. Thereby, our developed tracking control scheme can be applied to a quadrotor.

Let $\zeta = [x, y, z]^\top \in \mathbb{R}^3$, and $\eta = [\phi, \theta, \psi]^\top \in \mathbb{R}^3$ represent the absolute linear position and Euler angles defined in the inertial frame, respectively. The E-L equation of a quadrotor follows

$$m\ddot{\zeta} + mgI_z = RT_B \quad (68a)$$

$$J(\eta)\ddot{\eta} + C(\eta, \dot{\eta})\dot{\eta} = \tau_B, \quad (68b)$$

where $m \in \mathbb{R}^+$ denotes the mass of the quadrotor; $g \in \mathbb{R}^+$ is the gravity constant; $I_z = [0, 0, 1]^\top$ represents a column vector; $T_B = [0, 0, T]^\top \in \mathbb{R}^3$, where $T \in \mathbb{R}$ is the thrust in the direction of the body z -axis; $\tau_B = [\tau_\phi, \tau_\theta, \tau_\psi]^\top \in \mathbb{R}^3$ denotes the torques in the direction of the corresponding body frame angles; $R, J(\eta), C(\eta, \dot{\eta}) \in \mathbb{R}^{3 \times 3}$ represent the rotation matrix, Jacobian matrix, and Coriolis term, respectively.

Expanding the translational dynamics (68a) yields

$$\begin{aligned} \ddot{x} &= \frac{1}{m}T(C_\psi S_\theta C_\phi + S_\psi S_\phi) \\ \ddot{y} &= \frac{1}{m}T(S_\psi S_\theta C_\phi - C_\psi S_\phi) \\ \ddot{z} &= -g + \frac{1}{m}TC_\theta C_\phi \end{aligned} \quad (69)$$

where $C(\cdot)$ and $S(\cdot)$ denote $\cos(\cdot)$ and $\sin(\cdot)$, respectively.

By introducing pseudo controls $u_1 = T(C_\psi S_\theta C_\phi + S_\psi S_\phi)$, $u_2 = T(S_\psi S_\theta C_\phi - C_\psi S_\phi)$, and $u_3 = TC_\theta C_\phi$, and denoting $x_{11} = x$, $x_{12} = \dot{x}$, $x_{21} = y$, $x_{22} = \dot{y}$, $x_{31} = z$, $x_{32} = \dot{z}$, we finally decouple the translational dynamics (68a) into the following three subsystems

$$\dot{x}_{11} = x_{12}, \quad \dot{x}_{12} = \frac{1}{m}u_1 \quad (70a)$$

$$\dot{x}_{21} = x_{22}, \quad \dot{x}_{22} = \frac{1}{m}u_2 \quad (70b)$$

$$\dot{x}_{31} = x_{32}, \quad \dot{x}_{32} = -g + \frac{1}{m}u_3. \quad (70c)$$

By following the same procedures (2)–(8) clarified in Section III, we get three subsystems for the rotational dynamics (68b):

$$\dot{x}_{41} = x_{42}, \quad \dot{x}_{42} = -\frac{h_1}{J_{11}} + \frac{1}{J_{11}}u_4 \quad (71a)$$

$$\dot{x}_{51} = x_{52}, \quad \dot{x}_{52} = -\frac{h_2}{J_{22}} + \frac{1}{J_{22}}u_5 \quad (71b)$$

$$\dot{x}_{61} = x_{62}, \quad \dot{x}_{62} = -\frac{h_3}{J_{33}} + \frac{1}{J_{33}}u_6, \quad (71c)$$

where $h_i = \sum_{j=1, j \neq i}^3 J_{ij}\ddot{\eta}_j + C_i\dot{\eta}_j \in \mathbb{R}$, $i = 1, 2, 3$; $u_4 = \tau_\phi$, $u_5 = \tau_\theta$, and $u_6 = \tau_\psi$.

The aforementioned procedures (69)–(71) allow us to get 6 subsystems in the same form as (2). Then, we could adopt our developed DR-RL based tracking control policy, as clarified in

Section IV, to drive the quadrotor (68) to track the predefined reference trajectory $x_r = [x_d, y_d, z_d, \psi_d]^\top$. Note that after the explicit values of pseudo controls u_1 , u_2 , and u_3 are gotten, the trust T , and reference angles ϕ_d , θ_d are obtained as

$$T = \sqrt{u_1^2 + u_2^2 + u_3^2} \quad (72)$$

$$\phi_d = \arctan\left(\frac{u_1 S_\psi - u_2 C_\psi}{\sqrt{(u_1 C_\psi + u_2 S_\psi)^2 + u_3^2}}\right), \quad \phi_d \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \quad (73)$$

$$\theta_d = \arctan\left(\frac{u_1 C_\psi + u_2 S_\psi}{u_3}\right), \quad \theta_d \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \quad (74)$$