Transfer Learning for a Class of Cascade Dynamical Systems

Shima Rabiei†, Sandipan Mishra⋆ and Santiago Paternain†

***Abstract*— This work considers the problem of transferring a policy in the context of reinforcement learning. Specifically, we consider training a policy in a reduced order system and deploying it in the full state system. The motivation for this training strategy is that simulating full-state systems with complex dynamics may take excessive time. While transfer learning alleviates this issue, the transfer guarantees depend on the discrepancy between the two systems. In this work, we consider a class of cascade dynamical systems, where a subset of the state variables influences the dynamics of the remaining states but not vice-versa. We refer to the former as internal states. The reinforcement learning agent learns a policy using a model that ignores the internal states, treating them instead as commanded inputs. In deploying the policy in the full-state system, classic controllers (e.g., a PID) handle the dynamics of the internal states so that they track the reference signal provided by the reinforcement learning policy. The cascade structure allows us to provide transfer guarantees that depend on the stability of the inner loop controller. Numerical experiments on a quadrotor support the theoretical findings.**

I. Introduction

Reinforcement Learning (RL) has successfully solved control problems with complex dynamics and uncertainty in robotics [1], [2], power systems [3], [4] and aerial vehicles [5], [6] among others. Their success notwithstanding, RL algorithms require large amounts of data and considerable time [7], [8]. Moreover, the complexity of these problems increases with the dimensionality of the state-action space, something known as the curse of dimensionality [9]. To alleviate this burden (and also due to safety considerations), it is common to train in simulations [10], [11]. High fidelity simulators consider the full state of the system and the time to run them could be orders of magnitude larger than the process that is being simulated [11], [12]. Although this reduces the amount of data needed from the real system, it does not solve the problem of taking excessive time to converge. Hence, to improve the running time, it is not uncommon to consider reduced-order models with simplified dynamics [11], [13]–[16]. However, these suffer from the sim2real gap (see e.g., [17]). Specifically, a policy trained for certain dynamics may not necessarily maintain its performance when deployed in the real world.

Several approaches have been proposed in the literature to bridge this gap. System identification, parameter estimation or model learning focus on calibrating simulation models to match real-world dynamics using data [18]. Another line of work consists in fine-tuning the policy learned in simulation with real-world data [19]–[21]. Meta-learning follows this idea, however, during the training stage agents are presented with several tasks. This enables agents to quickly adapt to new environments, thereby facilitating their adaptation to discrepancies between simulation and reality [22], [23]. A related approach is adversarial training, where disturbances are introduced to improve robustness to unexpected scenarios [24]. Curriculum learning trains agents progressively with increasing complexity to aid generalization [25].

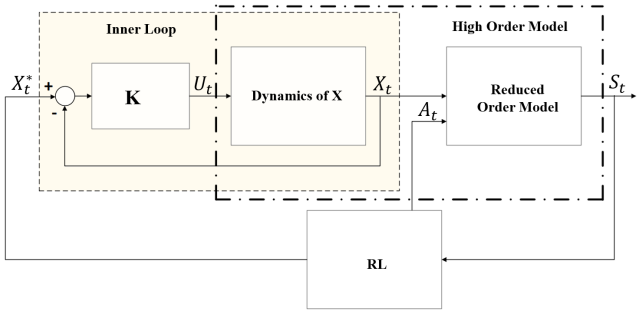

Fig. 1: Overview of the approach. The RL agent is trained on the reduced order model, where the inner loop is considered to have a unit transfer function. $X_t^\star$ and A_t are the actions of the RL agent. The learned policy is then transferred to the full system that includes the dynamics of the inner state X.

In this work, we exploit the structure of cascade dynamical system to guarantee transferability. In particular, we focus on systems where a subset of the state variables influences the rest but not vice-versa (see Figure 1). These problems are motivated by aerial vehicles, where the linear accelerations depend on their attitude [11], [26], but they find applications in other domains such as robotics [27] and process control [28]. In these scenarios, it is common to design controllers with nested loops. Often, the dynamics of the inner states are simple, and classic controllers, e.g., PID, can be designed [29]. Thus, our approach consists in training the RL policies in the reduced order model (where these inner states are treated as inputs), and then deploy the policies in the full state system (see Figure 1).

In the next section, we formalize the transfer learning problem and the hypotheses that define the class of dynamical systems considered. Section III provides the main results of this work, where we establish bounds in performance degradation under input-to-state stability of the inner loop dynamics. Other than concluding remarks, this work finishes with numerical experiments (Section IV) where we validate the theoretical findings in a quadrotor navigation problem.

†The authors are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute. Email: {rabiei, paters}@rpi.edu

⋆ The author is with the Department of Mechanical, Aerospace and Nuclear Engineering, Rensselaer Polytechnic Institute. Email: msihr2@rpi.edu

II. Problem Statement

In this paper, we consider the problem of training a policy in a reduced order model and transfer it to the full state dynamical system. To be formal, let us denote by $\mathcal{S} \times \mathcal{X} \in \mathbb{R}^n \times \mathbb{R}^m$ the state space of the latter dynamical system and let $\mathcal{A} \times \mathcal{U} \in \mathbb{R}^p \times \mathbb{R}^q$ be its input space. The states $X \in \mathcal{X}$ are the internal states in the cascade structure and the inputs $U \in \mathcal{U}$ are the inputs that directly affect the internal states (see Figure 1). As such, the state and action spaces of the reduced order model are $\mathcal{S}$ and $\mathcal{A} \times \mathcal{X}$, respectively (see Figure 1). The dynamics of the systems are characterized by the state transition probabilities. Let $\Delta(\cdot)$ represent the simplex over a set and denote by $\mathbb{P}_R : \mathcal{S} \times (\mathcal{A} \times \mathcal{X}) \to \Delta(\mathcal{S})$, and $\mathbb{P}_H : (\mathcal{S} \times \mathcal{X}) \times (\mathcal{A} \times \mathcal{U}) \to \Delta(\mathcal{S} \times \mathcal{X})$ the transition probabilities of the reduced order model and the full state system respectively. These represent the probabilities to transition to a state given the current state and the action selected. Since the RL policy is to be trained in the reduced order model, we defined it as a probability distribution $\pi : \mathcal{S} \to \Delta(\mathcal{A} \times \mathcal{X})$. Likewise, the reward is given by $r : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$. With these definitions, the optimal policy is

$$\pi_R^\star = \arg\max_\pi \mathbb{E}_R \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t, X_t) \right], \tag{1}$$

where the expectation is with respect to the reduced-order transition probabilities and the policy. To ensure finite expected returns, a requirement for convergence of many RL algorithms (see e.g., [30], [31]), we assume bounded rewards. We formalize this assumption next.

Assumption 1: The rewards are bounded by a constant $B > 0$, i.e., $|r(s, a, x)| \leq B$ for all $(s, a, x) \in \mathcal{S} \times \mathcal{A} \times \mathcal{X}$.

The problem above can be solved through a myriad of methods, e.g., [30], [32]. Since the solution depends on the dynamics of the reduced-order model (which ignores the dynamics of X), its performance may degrade when transferred to the full-state system. Moreover, the policy cannot be directly applied to the high-order system as it does not provide an action $U_t \in \mathcal{U}$. Instead, the output of the RL policy can be considered a reference $X_t^\star$ for state X_t (see Figure 1). Then, a controller (depicted with the block K in Figure 1) is in charge of tracking the reference by setting

$$U_t = K(X_t^\star, X_t). \tag{2}$$

In what follows we use the subindex K to represent the Markov Decision Process (MDP) that results from incorporating the inner controller into the high-order transitions $\mathbb{P}_H$. Thus, we define the transition probability of transitioning from state S_t to state S_{t+1} given the selected action A_t, the inner-loop state X_t and the RL reference $X_t^\star$

$$\mathbb{P}_K(S_{t+1} \mid S_t, A_t, X_t^\star) = \sum_{X_t \in \mathcal{X}} \mathbb{P}_H(S_{t+1} \mid S_t, A_t, X_t, K(X_t^\star, X_t)). \tag{3}$$

We then define the expected cumulative return in the full-order system under the transferred policy $\pi_R^\star$ as

$$V_K^{\pi_R^\star} = \mathbb{E}_K \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t, X_t, X_t^\star) \mid \pi_R^\star \right]. \tag{4}$$

The above quantity is our primary metric for evaluating the performance degradation of transferring the optimal policy $\pi_R^\star$ in (1). In Section III we provide guarantees on this degradation. Before doing so, we formalize the cascade structure of the system.

Assumption 2: The dynamics of the state $X_t \in \mathcal{X}$ are independent of the state $S_t \in \mathcal{S}$ and action $A_t \in \mathcal{A}$, i.e.,

$$\mathbb{P}_H(X_{t+1} \mid X_t, S_t, A_t, U_t) = \mathbb{P}_H(X_{t+1} \mid X_t, U_t). \tag{5}$$

Assumption 3: For all $U_t \in \mathcal{U}$ it holds that:

$$\mathbb{P}_H(S_{t+1} \mid S_t, A_t, X_t, U_t) = \mathbb{P}_R(S_{t+1} \mid S_t, A_t, X_t). \tag{6}$$

Assumption 2 formalizes the cascade structure, where the state $X_{t+1} \in \mathcal{X}$ in the full-state system is independent of the state $S_t \in \mathcal{S}$ and the action $A_t \in \mathcal{A}$ (cf., Figure 1). Assumption 3 states that if the current state S_t, the action A_t and X_t (which is an inner state in the full-state system and an action in the reduced order model) are equal, then the probability of transitioning to any state S_{t+1} for both systems is the same. These assumptions hold in various fields where the dynamics have a cascade structure. A concrete example is a quadrotor, where the pitch or roll angles influence the linear accelerations, but the latter do not affect the former (see Section IV for details).

Note that the initial state distribution affects the trajectory and therefore, the expected returns of the two models (1) and (4). Our next assumption is that the initial state distribution in both systems is the same. Hence, it removes a source of discrepancy when evaluating the degradation. Furthermore, it guarantees that the latter depends only on the difference in the dynamics of the systems.

Assumption 4: Let $\mu_R(S_0)$ and $\mu_H(S_0)$ be the initial state distributions of the reduced-order and high-order models respectively. For all $S_0 \in \mathcal{S}$ $\mu_R(S_0) = \mu_H(S_0)$.

We also make the following technical assumption regarding the transition probabilities of the reduced order model.

Assumption 5: The transition probability of the reduced model is L-Lipschitz continuous in X in total variation norm. This is, for any $X_t, X_t' \in \mathcal{X}$ it follows that

$$\|\mathbb{P}_R(S_{t+1} \mid S_t, A_t, X_t') - \mathbb{P}_R(S_{t+1} \mid S_t, A_t, X_t)\|_{TV} \leq L\|X_t' - X_t\|. \tag{7}$$

The above assumption is mild. For instance, the total variation of dynamical systems with Gaussian noise is bounded by a function proportional to the Euclidean distance between the means [33]. Thus, Assumption 5 holds if the mean is Lipschitz, which is often the case, see e.g., Section IV.

III. Transfer guarantees

In this section, we focus on establishing theoretical guarantees on the performance of the policy $\pi_R^\star$ in (1) (learned in the reduced order model) when transferred to the full state dynamical system. The inner controller tracks the reference

$X_t^\star$ provided by the RL agent, and we assume the inner loop to be stable. This is the subject of the following assumption.

Assumption 6: Consider a closed loop system composed by the system with the transition $\mathbb{P}_H(X_{t+1} \mid X_t, U_t)$. The selected controller $U_t = K(X_t^\star, X_t)$ is such that

$$\mathbb{E}\Big[\|X_t - X_t^\star\|_P\Big] \leq \alpha\mathbb{E}\Big[\|X_{t-1} - X_{t-1}^\star\|_P\Big] + \beta\mathbb{E}\Big[\|X_t^\star - X_{t-1}^\star\|_P\Big], \quad (8)$$

where $P \in \mathbb{S}_{++}^{m\times m}$ is a positive definite matrix,$\|\cdot\|_P$ denotes its induced norm and $\alpha \in (0,1)$ and $\beta > 0$.

Assumption 6 encodes that the inner-loop dynamics are *input-to-state stable (ISS)* in expectation [34], [35] with respect to changes in the reference $X_t^\star$. Indeed, notice that applying the previous expression recursively we have that

$$\mathbb{E}\Big[\|X_t - X_t^\star\|_P\Big] \leq \alpha^{t-1}\mathbb{E}\Big[\|X_0 - X_0^\star\|_P\Big] + \beta\sum_{l=1}^{t}\alpha^{t-l}\mathbb{E}\Big[\|X_l^\star - X_{l-1}^\star\|_P\Big]. \quad (9)$$

Defining $e_t = \mathbb{E}\left[\|X_t - X_t^\star\|\right]$ and $z_t = \mathbb{E}\left[\left\|X_t^\star - X_{t-1}^\star\right\|\right]$, the first term on the right hand side of the above equation is a class $\mathcal{KL}$ function of e_0 and the second term is a class $\mathcal{K}$ function of $\|z\|_\infty$ (see e.g., [36]). While this assumption simplifies the theoretical analysis of our transfer guarantees, it is typically satisfied by any stabilizing controller K. That is, if the closed-loop system exponentially (or contractively) stabilizes around $X_t^\star$, then (8) naturally holds.

Assumption 6 is, in general, not sufficient to provide guarantees in the performance degradation when transferring the policy. Indeed, the error between the state X_t and the commanded action by the policy $X_t^\star$ depends on the variation of the commanded action. Since the $\mathcal{K}$ function is such that it goes to infinity when its argument goes to infinity, without guarantees on the boundedness of the variation $\mathbb{E}\Big[\|X_{t+1}^\star - X_t^\star\|_P\Big]$, the tracking error (9) could be arbitrarily large. Thus, we introduce the following assumption which bounds the infinity norm of the variation.

Assumption 7: There exists a constant $C > 0$ such that

$$\mathbb{E}\Big[\|X_t^\star - X_{t-1}^\star\|_P\Big] \;\leq\; C, \quad (10)$$

where P is the matrix defined in Assumption 6.

Assumption 7 imposes that the difference between consecutive reference states $X_t^\star$ and $X_{t-1}^\star$ remains bounded in expectation. Although the original motivation to introduce this assumption stems from the analysis, this requirement is conceptually aligned with standard control-design practices. Indeed, abrupt shifts in the reference are undesirable and potentially infeasible. Furthermore, when training a reinforcement learning agent to provide the reference signal (see Figure 1), one can promote bounded variations by penalizing large changes in the policy's actions or reference trajectories. Moreover, such bounds can be explicitly enforced by incorporating constraints on the policy itself, see e.g., [37].

Under these assumptions, we are now in conditions to establish a bound in the total variation distance between the transition probabilities of the reduced and high-order systems with the inner loop controller. We define for simplicity

$$TV(t+1) := \sup_{S_t\in\mathcal{S}, A_t\in\mathcal{A}, X_t^\star\in\mathcal{X}} \frac{1}{2}\sum_{S_{t+1}\in\mathcal{S}} \left|\mathbb{P}_K(S_{t+1} \mid S_t, A_t, X_t^\star) - \mathbb{P}_R(S_{t+1} \mid S_t, A_t, X_t^\star)\right|, \quad (11)$$

where $\mathbb{P}_K$ is the transition defined in (3). It is important to point out that $TV(t)$ could be time-dependent. Indeed, a stable inner controller tracks the commanded $X_t^\star$, resulting in dynamics that approach those of the reduced order model. The next proposition formalizes this idea.

Proposition 1: Under Assumption 2–7, the total variation defined in (11) satisfies

$$TV(1) \leq \frac{L}{2}\mathbb{E}\left[\|X_0 - X_0^\star\|\right], \quad (12)$$

and for all time $t \geq 1$ it holds that

$$TV(t+1) \leq \frac{L\rho}{2}\left(\alpha^t\mathbb{E}\left[\|X_0 - X_0^\star\|\right] + \beta C\frac{1-\alpha^t}{1-\alpha}\right), \quad (13)$$

where ρ is the square root of the condition number of the matrix P defined in Assumption 6.

Proof: See Appendix A. ∎

Proposition 1 shows that the total variation bound consists of two parts: one part decays exponentially (depending on α) due to the stability of the inner-loop controller, and the other part depends on how much the reference $X_t^\star$ changes between consecutive steps. Specifically, the second term can be made zero only if (i) the reference signal is constant (so $C = 0$), or (ii) $\beta = 0$.

The condition $\beta = 0$ requires that the controller anticipates changes in the upcoming reference $X_{t+1}^\star$. In other words, the controller must already know (or perfectly predict) $X_{t+1}^\star$ at time t. In practice, when $X_{t+1}^\star$ is generated by an RL policy, it is only revealed at time $t+1$. Hence, $\beta = 0$ is generally not achievable without an explicit predictive mechanism. That said, one may partially approximate this ideal scenario by introducing additional complexity into the controller design (e.g., a model predictive approach) so that future references are estimated or anticipated more accurately.

The key implication of Proposition 1 is that it allows us to bound the performance degradation of the optimal policy when transferred to the full state system. Note that the faster the controller is (α closer to zero), the smaller the error between the dynamics; hence, we should expect better transfer. This is the subject of the next theorem.

Theorem 1: Under Assumption 1–7, the performance degradation by transfer for all t can be bounded as

$$(1-\gamma)\left|V_K^{\pi_R^\star} - V_R^\star\right| \leq \frac{BL\gamma}{1-\alpha\gamma}\left(\frac{\rho\gamma\beta C}{1-\gamma} + (1+\alpha\gamma(\rho-1))\mathbb{E}\left[\|X_0 - X_0^\star\|\right]\right). \quad (14)$$

Proof: See Appendix B. ∎

The previous theorem confirms the intuition derived earlier. In particular, if the initial inner loop tracking error is zero ($X_0 = X_0^\star$) and the inner loop is exponentially stable ($\beta = 0$), then there is no loss by transfer. This is not surprising since an exponentially stable controller will track the reference without error if the initial error is zero. Thus, the dynamics of the reduced and the high-order systems are the same. When the initial state does not correspond to the initial commanded action the effect of α can be observed. Indeed note that the bounds are monotonically increasing with α. This confirms the intuition that a good controller (small α) should achieve a better transfer than a poor controller ($\alpha \approx 1$). This is also the case under the ISS assumption. The term involving C and β highlights the impact of the rate of change of the reference trajectory. A larger C (indicating greater changes between consecutive references) or β (larger gain) leads to a larger bound. Lastly, it is important to notice that, as it is usual in problems with discounted infinite horizon, the more importance is given to the future, i.e., larger γ , the larger the bound.

In the next section we focus on demonstrating the practical implications of the above theory in the problem of a quadrotor navigating to a desired destination.

IV. Experimental Results

In this section, we present numerical results evaluating the performance degradation when transferring a policy from a reduced-order model to a full state state system. We consider a simulated quadrotor navigation task. The horizontal and vertical positions of the quadrotor are given by $[y, z]^\top \in [0, 10] \times [0, 10]$. The objective of the agent is to reach a fixed target at $[y_{\text{target}}, z_{\text{target}}]^\top = [9, 9]$. In the reduced-order model, the attitude of the quadrotor (pitch or roll) is considered an input, while in the high-order system, it is a state. In what follows we provide details of the two systems.

A. Reduced-Order Model of a Quadrotor

We describe the motion of the quadrotor with the state

$$S_t \;=\; [\,y_t,\; v_t^y,\; z_t,\; v_t^z\,]^\top,$$

where y_t, z_t are the horizontal and vertical positions at time t, and v_t^y, v_t^z are the corresponding velocities. The control input (action) at each time step are the thrust T_t and the pitch θ_t

$$A_t \;=\; [T_t,\; \theta_t]^\top \;\in\; [-1, 1] \;\times\; \left[-\frac{\pi}{8}, \frac{\pi}{8}\right].$$

We consider a discretized version of the dynamics [26] via a forward-Euler scheme with sampling time Δt. Define

$$a_{t+1}^y \;=\; \frac{mg + T_t}{m}\,\sin\theta_t, \quad a_{t+1}^z \;=\; \frac{mg + T_t}{m}\,\cos\theta_t \;-\; g.$$

TABLE I: Training Hyperparameters for PPO

Hyperparameter	Value	Description
Learning rate	3×10^{-4}	Step size for the optimizer
Optimizer	Adam	Optimization algorithm
Number of episodes	$1,000,000$	Total training episodes
Batch size	64	Number of samples per batch
Clip range ϵ	0.2	PPO clipping parameter

Then our discrete-time state updates become

$$v_{t+1}^y \;=\; v_t^y \;+\; a_{t+1}^y\,\Delta t, \quad y_{t+1} \;=\; y_t \;+\; v_{t+1}^y\,\Delta t,$$
$$v_{t+1}^z \;=\; v_t^z \;+\; a_{t+1}^z\,\Delta t, \quad z_{t+1} \;=\; z_t \;+\; v_{t+1}^z\,\Delta t.$$

Note that to hover, the quadrotor must compensate the gravitational force mg, then the total thrust is $mg + T_t$. Throughout our experiments, we use the parameter values $m = 1\,\text{kg}$, $g = 9.81\,\text{m/s}^2$, and $\Delta t = 0.05\,\text{s}$.

B. High-Order Model of a quadrotor

In addition to the state of the reduced order model, the high order model also includes the angle as part of the state $[S^\top, \theta]^\top$. We consider first-order dynamics $\dot\theta = u$, where the input is given by a proportional controller $u = -K_p(\theta - \theta^\star)$ so that the state tracks the commanded angle. The exact discretization of the above dynamics yields

$$\theta_{t+1} = e^{-K_p\Delta t}\theta_t + \left(1 - e^{-K_p\Delta t}\right)\,\theta_t^\star. \quad (15)$$

Note that the gain of the controller determine the discrepancy between the reduced and higher order models

$$\theta_{t+1} - \theta_{t+1}^\star = e^{-K_p\Delta t}\left(\theta_t - \theta_t^\star\right) + \left(\theta_t^\star - \theta_{t+1}^\star\right). \quad (16)$$

Along the lines of Assumption 6, we have that $\beta = 1$ and $\alpha = e^{-K_p\Delta t}$, which implies that the larger the gain the smaller α. Our theoretical results predict that larger gains K_p achieve better transfer guarantees as measured by the discrepancy in the value functions.

C. Training the agent in the reduced order system.

We start the section by defining the reward function. The agent's objective is to attain the target while avoiding collisions with the boundaries of the space. As such, we design the following reward to inform these goals

$$r(s_t, a_t, x_t) = \mathbf{1}_{\text{target}}(s_t, a_t, x_t) - \frac{\|p_t - p_{\text{target}}\|}{d_{\max}} - C\mathbf{1}_{\text{boundary}}(s_t, a_t, x_t), \quad (17)$$

where $\mathbf{1}_{\text{target}}$ denotes the indicator function taking the value if $\|p_t - p_{\text{target}}\|_\infty \leq 0.05$ and zero otherwise and $\mathbf{1}_{\text{boundary}}$ is the indicator function taking the value one if $p_t \notin [0, 10] \times [0, 10]$ and zero otherwise. The second term on the right-hand side of (17) represents the normalized distance between the agent and the target, with $d_{\max} = 10\sqrt{2}$ being the maximum possible distance within the environment. We set the value of the penalty for violating the environement's boundaries to $C = 5 \times 10^3$. The discount factor γ is set to 0.995.

We train the agent in the reduced order model using Proximal Policy Optimization (PPO) [38]. The training hyperparameters are listed in Table I.

D. Policy Transfer and Evaluation

We assess the performance degradation of the transferred policy by examining the differences in expected discounted returns between the full state system (4) and the reduced-order model (1) for different values of K_p. Figure 2 shows the mean differences in expected discounted rewards, calculated by averaging the returns over 100 iterations with random initialization of the state S. Consistent with Theorem 1, we observe that a larger K_p leads to a smaller degradation of the performance. The reason for the difference in performance is due to the distance to the target that each system attains. Indeed, the reward function (17) penalizes the distance to the target. As it can be observed in Figure 3, the larger the gain the faster the convergence to the target.

Figure 4 shows the mean and standard deviation of the orientation error, of 100 trajectories with random initial states S. The plot shows that as K_p increases, the orientation error is reduced. Thus reducing the discrepancy between the transition probabilities of the full-state and reduced-order models. Thus improving the transfer to the full-state system.This outcome aligns with Proposition 1.

Furthermore, Figure 5 compares the orientation trajectories

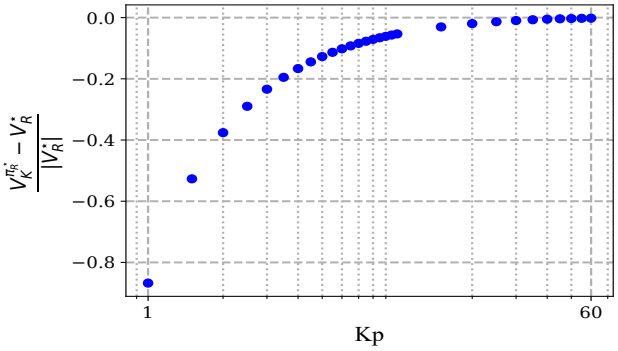

Fig. 2: Illustration of the relative average differences in expected discounted returns between the full-state system (4) and the reduced-order model (1) for various values of K_p. These are averaged over 100 experiments with randomly initialized states S. In accordance with Theorem 1, we observe that increasing K_p results in less performance degradation.

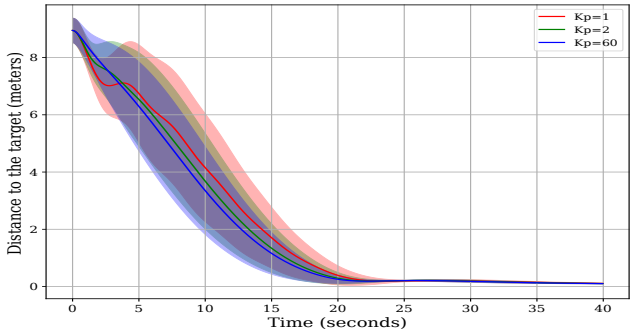

Fig. 3: Mean and standard deviation of the distance to the target over 100 iterations for proportional gains K_p set to 1, 2 and 60.

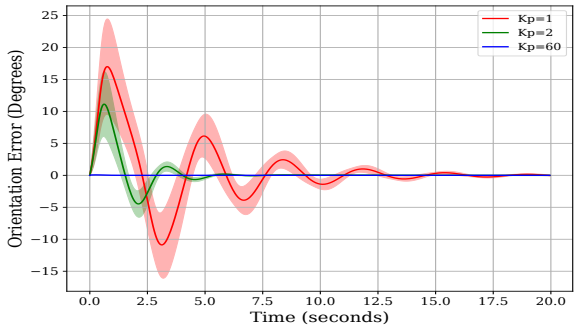

Fig. 4: Mean and standard deviation of the orientation error computed over 100 iterations with random initial states S. The plot illustrates that as K_p increases, the orientation error approaches zero. Thus reducing the total variation between the two transitions. This result aligns with Proposition 1.

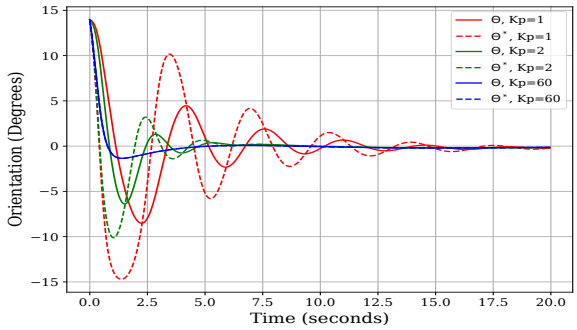

Fig. 5: Comparison of the orientation trajectories of the full-state and reduced-order models for different values of K_p. Similar to Figure 4 we observe that the larger K_p the more similar the trajectories between θ and the reference $\theta^\star$. Furthermore, we observe that the larger K_p the less variations there are on $\theta^\star$. According to Theorem 1 a large variation on the reference results in worse transfer which is consistent with this experiment.

of the full-state and reduced-order models. Note that with a small value of K_p, the variation in the orientation of the reduced-order model (dashed lines) is larger. This is, there is more commanded effort. Since the bound in Theorem 1 depends on this variation, the larger the bound in this variation the more degradation in the transfer one can expect. This effect added to the larger value of α for smaller gains K_p explains the degradation curve in Figure 2.

V. Conclusion

In this work we considered the transfer of an optimal policy learned in a reduced order model into a full state system for a class of cascade systems. Our theoretical guarantees establish that the performance degradation depends on the stability properties of the inner loop controller. In particular, the better the commanded signal by the RL policy can be tracked by the state of the high-order system, the smaller the

loss in performance. We verified our theoretical findings with the example of a quadrotor, where its attitude corresponds to the inner state that is an action for the reduced order model.

ACKNOWLEDGEMENT

This work was sponsored by the Office of Naval Research (ONR), under contract number N00014-23-1-2377.

APPENDIX

A. Proof of Proposition 1

Marginalizing the joint probability of S_{t+1} and X_t and using Bayes rule it follows that

$$\begin{aligned}\mathbb{P}_K(S_{t+1}\,|S_t,A_t,X_t^\star) &= \sum_{X_t\in\mathcal{X}}\mathbb{P}_K(S_{t+1}\mid S_t,A_t,X_t,X_t^\star)\mathbb{P}_K(X_t)\\ &= \sum_{X_t\in\mathcal{X}}\mathbb{P}_H(S_{t+1}\mid S_t,A_t,X_t,X_t^\star)\mathbb{P}_K(X_t).\end{aligned} \tag{18}$$

The second equality holds since the impact of the controller is only on defining X_t and not on the transitions of S_t given X_t. Furthermore, using that the transitions of the reduced and high-order systems are related by Assumption 3, (18) can be written as

$$\begin{aligned}\mathbb{P}_K(S_{t+1}\,|S_t,A_t,X_t^\star) &= \sum_{X_t\in\mathcal{X}}\mathbb{P}_R(S_{t+1}\mid S_t,A_t,X_t)\mathbb{P}_K(X_t).\end{aligned} \tag{19}$$

On the other hand, since the transitions of the reduce order model are independent of the variable X_t, it follows that

$$\begin{aligned}\mathbb{P}_R(S_{t+1}\,|S_t,A_t,X_t^\star) &= \sum_{X_t\in\mathcal{X}}\mathbb{P}_R(S_{t+1}\mid S_t,A_t,X_t^\star)\mathbb{P}_K(X_t).\end{aligned} \tag{20}$$

Combining (19) and (20) one can write the difference of transition probabilities in both MDPs as

$$\begin{aligned}&\mathbb{P}_K(S_{t+1}\mid S_t,A_t,X_t^\star)-\mathbb{P}_R(S_{t+1}\mid S_t,A_t,X_t^\star)\\ &=\sum_{X_t\in\mathcal{X}}\Big(\mathbb{P}_R(S_{t+1}\,|S_t,A_t,X_t)\\ &\qquad-\mathbb{P}_R(S_{t+1}\mid S_t,A_t,X_t^\star)\Big)\mathbb{P}_K(X_t).\end{aligned} \tag{21}$$

From the triangle inequality and the definition of total variation (see (11)) it follows that

$$\begin{aligned}2TV(t+1)\leq\sum_{S_{t+1}\in\mathcal{S}}\sum_{X_t\in\mathcal{X}}&\Big|\mathbb{P}_R(S_{t+1}\mid S_t,A_t,X_t)\\ &-\mathbb{P}_R(S_{t+1}\mid S_t,A_t,X_t^\star)\Big|\mathbb{P}_K(X_t).\end{aligned} \tag{22}$$

In the above expression, we have also used that $\mathbb{P}_K(X_t)$ is non-negative. We can further upper bound the above expression by using that the transition probabilities of the reduced order model are Lipschitz with respect to X (Assumption 5)

$$\begin{aligned}&2TV(t+1)\\ &\leq L\sum_{X_t\in\mathcal{X}}\|X_t-X_t^\star\|\mathbb{P}_K(X_t)=L\mathbb{E}\left[\|X_t-X_t^\star\|\mid X_t^\star\right],\end{aligned} \tag{23}$$

where the equality follows directly from the definition of conditional expectation. Observe that for the above expression to hold, X_t needs to be independent of $X_t^\star$.

For any positive definite matrix P with minimum and maximum eigenvalues denoted by $\lambda_{\min}$ $\lambda_{\max}$, the following inequality holds (see e.g., [39])

$$\begin{aligned}\sqrt{\lambda_{\min}}\|X_t-X_t^\star\|&\leq\|X_t-X_t^\star\|_P\\ &\leq\sqrt{\lambda_{\max}}\|X_t-X_t^\star\|.\end{aligned} \tag{24}$$

Using (24) and applying (8) recursively, it follows that

$$\begin{aligned}&\mathbb{E}\Big[\|X_t-X_t^\star\|\Big]\leq\frac{1}{\sqrt{\lambda_{\min}}}\mathbb{E}\Big[\|X_t-X_t^\star\|_P\Big] \qquad (25)\\ &\leq\frac{1}{\sqrt{\lambda_{\min}}}\alpha^t\mathbb{E}\Big[\|X_0-X_0^\star\|_P\Big]+\sum_{k=0}^{t-1}\alpha^k\beta\mathbb{E}\Big[\|X_t^\star-X_{t-1}^\star\|_P\Big].\end{aligned}$$

Substituting the bound given in Assumption (7) into (25) and writing the geometric sum compactly yields

$$\begin{aligned}&\mathbb{E}\Big[\|X_t-X_t^\star\|\Big]\leq\\ &\frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}}\left(\alpha^t\mathbb{E}\left[\|X_0-X_0^\star\|\right]+\beta C\frac{1-\alpha^t)}{1-\alpha}\right).\end{aligned} \tag{26}$$

Substituting the bound (26) into (23) and recognizing that the condition number of P is defined as $\lambda_{\max}/\lambda_{\min}$ completes the proof of the result. ■

B. Proof of Theorem 1.

We define $\mathbb{P}_K((s,a,x)_{0:t})$ and $\mathbb{P}_R((s,a,x)_{0:t})$ as the joint probability distributions over the trajectories of the full-state system (including the controller) and the reduced-order model, respectively. To be precise, applying Bayes' rule, and using that the policy depends only on the state S_t and the Markov property of the transitions we can recursively define

$$\begin{aligned}&\mathbb{P}_R^\star((s,a,x)_{0:t})=\\ &\mathbb{P}_R^\star((s,a,x)_{0:t-1})\mathbb{P}_R(s_t\mid s_{t-1},a_{t-1},x_{t-1})\pi_R^\star(a_t,x_t\mid s_t),\end{aligned} \tag{27}$$

where $\mathbb{P}_R(s_t\mid s_{t-1},a_{t-1},x_{t-1})$ is the transition of the reduced order model and $\pi_R^\star$ is the solution to (1). Analogously

$$\begin{aligned}&\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t})=\\ &\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})\mathbb{P}_K(s_t\mid s_{t-1},a_{t-1},x_{t-1})\pi_R^\star(a_t,x_t\mid s_t),\end{aligned} \tag{28}$$

where the transition $\mathbb{P}_K(s_t\mid s_{t-1},a_{t-1},x_{t-1})$ is the one defined in (3). To support the proofs of Theorem 1, it is essential to bound the discrepancy between the trajectory distributions of the high-order system with the inner-loop controller and the reduced-order MDP. This discrepancy is measured by $\Delta P(t)$, defined as

$$\Delta P(t):=\sum_{(s,a,x)_{0:t}}\left|\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t})-\mathbb{P}^\star{}_R((s,a,x)_{0:t})\right|. \tag{29}$$

Lemma 1: Under Assumptions 4–7, the total variation distance between the trajectory distributions over the time horizon t satisfies $\Delta P(0) = 0$ and for $t \geq 1$ it holds that

$$\Delta P(t) \leq L\mathbb{E}\left[\|X_0 - X_0^\star\|\right] + L\rho\beta C\frac{(t-1) - t\alpha + \alpha^t}{(1-\alpha)^2} + L\rho\alpha\frac{1-\alpha^{t-1}}{1-\alpha}\mathbb{E}\left[\|X_0 - X_0^\star\|\right]. \quad (30)$$

Proof: See Appendix C. ∎

We begin by computing the difference between the expected cumulative returns associated with each MDP. Using the definition of expected cumulative return (see (1) and (4)) it follows that

$$V_K^{\pi_R^\star} - V_R^\star = \mathbb{E}_K^\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,x_t)\right] - \mathbb{E}_R^\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,x_t)\right]. \quad (31)$$

Writing the expectation as a sum and exchanging the order of the sums, we have that

$$V_K^{\pi_R^\star} = \sum_{t=0}^{\infty}\sum_{(s,a,x)_{0:\infty}}\gamma^t r(s_t,a_t,x_t)\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:\infty}), \quad (32)$$

where in the above notation we use $(s,a,x)_{0:\infty}$ to denote all states an actions. Since the rewards are bounded (Assumption 1), the sums in (31) are bounded. Thus, the Tonelli-Fubini Theorem (see e.g., [40]) justifies the exchange of the sum over time and expectation. We further claim that

$$V_K^{\pi_R^\star} = \sum_{t=0}^{\infty}\sum_{(s,a,x)_{0:t}}\gamma^t r(s_t,a_t,x_t)\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t}) \quad (33)$$

and analogously that

$$V_R^\star = \sum_{t=0}^{\infty}\sum_{(s,a,x)_{0:t}}\gamma^t r(s_t,a_t,x_t)\mathbb{P}_R^\star((s,a,x)_{0:t}). \quad (34)$$

We defer the proof of these claims to the end of the proof.

Subtracting (34) to (33) taking the absolute value on both sides of the above equation, using the triangle inequality and the fact that $|r(s,a)| \leq B$ (Assumption 1) it follows that

$$\left|V_K^{\pi_R^\star} - V_R^\star\right| \leq B\sum_{t=0}^{\infty}\gamma^t\Delta P(t). \quad (35)$$

Where we have also used the definition of $\Delta P(t)$ (see (29)). Then, by virtue of Lemma 1 it follows that

$$\left|V_K^{\pi_R^\star} - V_R^\star\right| \leq BL\mathbb{E}\left[\|X_0 - X_0^\star\|\right]\sum_{t=1}^{\infty}\gamma^t + BL\rho\mathbb{E}\left[\|X_0 - X_0^\star\|\right]\frac{\alpha}{1-\alpha}\sum_{t=2}^{\infty}\gamma^t(1-\alpha^{t-1}) + \frac{B\beta CL\rho}{(1-\alpha)^2}\left(\sum_{t=2}^{\infty}\gamma^t(t-1) - \alpha\sum_{t=2}^{\infty}\gamma^t t + \sum_{t=2}^{\infty}\gamma^t\alpha^t)\right). \quad (36)$$

From the convergence of $\sum_{t=1}^{\infty}\gamma^t = \gamma/(1-\gamma)$ and the following geometric series (see, e.g., [41, Sec. 8, Thm. 8.22])

$$\sum_{t=2}^{\infty}\gamma^t(1-\alpha^{t-1}) = \frac{\gamma^2(1-\alpha)}{(1-\gamma)(1-\alpha\gamma)},\ \sum_{t=2}^{\infty}\gamma^t t = \frac{\gamma^2(2-\gamma)}{(1-\gamma)2},$$

$$\sum_{t=2}^{\infty}\gamma^t\alpha^t = \frac{(\alpha\gamma)^2}{1-\alpha\gamma},\ \sum_{t=2}^{\infty}\gamma^t(t-1) = \frac{\gamma^2}{(1-\gamma)^2},$$

and replacing them in (36), it follows that

$$\left|V_K^{\pi_R^\star} - V_R^\star\right| \leq \frac{\gamma BL}{1-\gamma}\mathbb{E}\left[\|X_0 - X_0^\star\|\right]\left(\frac{1-\alpha\gamma+\rho\alpha\gamma}{1-\alpha\gamma}\right) + BL\rho\beta C\frac{\gamma^2}{(1-\gamma)^2(1-\alpha\gamma)}. \quad (37)$$

Re-arranging the above expressions yields (14). To complete the proof we are left to prove (33) and (34). We will prove the latter. The former follows the same steps. Split the sum into the sum until time t and another term after $t+1$

$$\sum_{(s,a)_{0:\infty}}\gamma^t r(s_t,a_t,x_t)\mathbb{P}_R^\star((s,a,x)_{0:\infty} = \quad (38)$$

$$\sum_{(s,a,x)_{0:t}}\gamma^t r(s_t,a_t)\mathbb{P}_R^\star(s,a,x)_{0:t} \quad (39)$$

$$\sum_{(s,a,x)_{t+1:\infty}}\mathbb{P}_R^\star\left((s,a,x)_{t+1:\infty} \mid (s,a,x)_{0:t}\right).$$

Since $\mathbb{P}\left((s,a)_{t+1:\infty} \mid (s,a)_{0:t}\right)$ is a probability the rightmost sum in the above equation is equals to one. Therefore, the above equation reduces to (33). ∎

C. Proof of Lemma 1

From (28) and (27) it follows that $\Delta P(t)$ yields

$$\Delta P(t) = \sum_{a_t,x_t}\pi(a_t,x_t \mid s_t)\sum_{(s,a,x)_{0:t-1},s_t}\left|\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})\mathbb{P}_K(s_t \mid s_{t-1},a_{t-1},x_{t-1}) - \mathbb{P}_R^\star((s,a,x)_{0:t-1})\mathbb{P}_R(s_t \mid s_{t-1},a_{t-1},x_{t-1})\right| \quad (40)$$

Since $\pi(a_t,x_t \mid s_t)$ is a probability distribution, the leftmost sum in the above equation equals one. Therefore, by adding and subtracting the term $\mathbb{P}_R(s_t \mid s_{t-1},a_{t-1},x_{t-1})\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})$ to each term in the above sum yields

$$\Delta P(t) = \sum_{(s,a,x)_{0:t-1},s_t}\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})\left|\mathbb{P}_K - \mathbb{P}_R\right| + \left|\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1}) - \mathbb{P}_R^\star((s,a,x)_{0:t-1})\right|\mathbb{P}_R, \quad (41)$$

where in the above expression we have omitted the variables of the transition probabilities for simplicity in the notation. Using the definition of the total variation in (11) we can upper bound the first term of the above sum as

$$\sum_{(s,a,x)_{0:t-1},s_t}\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})\left|\mathbb{P}_K - \mathbb{P}_R\right| \leq 2\sum_{(s,a,x)_{0:t-1}}\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})TV(t-1) = 2TV(t-1), \quad (42)$$

where the equality follows from the fact that $\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1})$ is a probability distribution. Using the fact that $\mathbb{P}_R$ is also a probability distribution, we can write the second term in (41) as

$$\sum_{(s,a,x)_{0:t-1},s_t} \left|\mathbb{P}_K^{\pi_R^\star}((s,a,x)_{0:t-1}) - \mathbb{P}_R^\star((s,a,x)_{0:t-1})\right| \mathbb{P}_R = \Delta P(t-1). \tag{43}$$

Substituting (42) and (43) into (41) reduces to

$$\Delta P(t) \leq 2TV(t) + \Delta P(t-1) \leq 2\sum_{l=1}^{t} TV(l). \tag{44}$$

Where the right-most inequality follows from applying the left-most inequality recursively and using that $\Delta P(0) = 0$. The latter holds since by Assumption 4 the initial distribution of the state is the same for both systems. The proof is then completed by replacing the total variation by the bounds in Proposition 1.The proof of the result is completed by rearranging the terms and simplifying the sums. ■

REFERENCES

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[2] L. C. Garaffa, M. Basso, A. A. Konzen, and E. P. de Freitas, "Reinforcement learning for mobile robotics exploration: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 3796–3810, 2021.

[3] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: reinforcement learning framework," *IEEE transactions on power systems*, vol. 19, no. 1, pp. 427–435, 2004.

[4] A. Dwivedi, S. Paternain, and A. Tajer, "Blackout mitigation via physics-guided rl," *arXiv preprint arXiv:2401.09640*, 2024.

[5] P. Abbeel, A. Coates, M. Quigley, and A. Ng, "An application of reinforcement learning to aerobatic helicopter flight," *Advances in neural information processing systems*, vol. 19, 2006.

[6] Y. T. Liu, E. Price, M. J. Black, and A. Ahmad, "Deep residual reinforcement learning based autonomous blimp control," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12566–12573, IEEE, 2022.

[7] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[8] S. Kamthe and M. Deisenroth, "Data-efficient reinforcement learning with probabilistic model predictive control," in *International conference on artificial intelligence and statistics*, pp. 1701–1710, PMLR, 2018.

[9] R. Bellman, "Dynamic programming," *science*, vol. 153, no. 3731, pp. 34–37, 1966.

[10] B. Osiński, A. Jakubowski, P. Zięcina, P. Miłoś, C. Galias, S. Homoceanu, and H. Michalewski, "Simulation-based reinforcement learning for real-world autonomous driving," in *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 6411–6418, IEEE, 2020.

[11] D. Jayarathne, S. Paternain, and S. Mishra, "Safe residual reinforcement learning for helicopter aerial refueling," in *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 263–269, IEEE, 2023.

[12] F. Muratore, M. Gienger, and J. Peters, "Data-driven domain randomization with bayesian optimization," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2957–2964, 2021.

[13] Y.-M. Chen, H. Bui, and M. Posa, "Reinforcement learning for reduced-order models of legged robots," 2023.

[14] Y.-M. Chen, J. Hu, and M. Posa, "Beyond inverted pendulums: Task-optimal simple models of legged locomotion," 2024.

[15] S. Kajita and K. Tanie, "Study of dynamic biped locomotion on rugged terrain-derivation and application of the linear inverted pendulum mode," *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, pp. 1405–1411 vol.2, 1991.

[16] J. Furusho and M. Masubuchi, "A theoretically motivated reduced order model for the control of dynamic biped locomotion," *Journal of Dynamic Systems Measurement and Control-transactions of The Asme*, vol. 109, pp. 155–163, 1987.

[17] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744, IEEE, 2020.

[18] S. Koos, J.-B. Mouret, and S. Doncieux, "The transferability approach: Crossing the reality gap in evolutionary robotics," *IEEE Transactions on Evolutionary Computation*, vol. 17, no. 1, pp. 122–145, 2013.

[19] F. Zhang, J. Leitner, B. Upcroft, and P. Corke, "Vision-based reaching using modular deep networks: From simulation to the real world," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2197–2203, 2015.

[20] M. P. 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)*, pp. 465–472, 2011.

[21] A. A. Rusu, M. Vecerik, T. E. Rothörl, N. Heess, R. Pascanu, and R. Hadsell, "Sim-to-real robot learning from pixels with progressive nets," *arXiv preprint arXiv:1610.04286*, 2016.

[22] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[23] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," *arXiv preprint arXiv:1703.06907*, 2017.

[24] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," *arXiv preprint arXiv:1703.02702*, 2017.

[25] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48, 2009.

[26] R. W. Beard and T. W. McLain, *Quadrotor Dynamics and Control*. Brigham Young University, 2012.

[27] M. W. Spong, "Model-based control of robot manipulators," *IEEE Control Systems Magazine*, vol. 7, no. 1, pp. 26–32, 1987.

[28] I. Kaya, N. Tan, and D. P. Atherton, "Improved cascade control structure for enhanced performance," *Journal of Process Control*, vol. 17, no. 1, pp. 3–16, 2007.

[29] K. Ogata, *Modern Control Engineering*. Prentice Hall, 5th ed., 2010.

[30] J. N. Tsitsiklis, "Asynchronous stochastic approximation and q-learning," *Machine learning*, vol. 16, pp. 185–202, 1994.

[31] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[32] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[33] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional gaussians with the same mean," *arXiv preprint arXiv:1810.08693*, 2018.

[34] P. Culbertson, R. K. Cosner, M. Tucker, and A. D. Ames, "Input-to-state stability in probability," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 5796–5803, 2023.

[35] Y. Kawano and Y. Hosoe, "Contraction analysis of discrete-time stochastic systems," *IEEE Transactions on Automatic Control*, vol. 69, no. 2, pp. 982–997, 2024.

[36] H. K. Khalil, *Nonlinear systems; 3rd ed.* Upper Saddle River, NJ: Prentice-Hall, 2002.

[37] W. Chen, D. Subramanian, and S. Paternain, "Probabilistic constraint for safety-critical reinforcement learning," *IEEE Transactions on Automatic Control*, 2024.

[38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *ArXiv*, vol. abs/1707.06347, 2017.

[39] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2nd ed., 2013.

[40] W. Rudin, *Real and Complex Analysis*. McGraw-Hill Education, 1987.

[41] T. M. Apostol, *Mathematical Analysis*. Addison-Wesley, 2nd ed., 1974.