Extended Abstract Track

Leveraging Symmetry to Accelerate Learning of Trajectory Tracking Controllers for Free-Flying Robotic Systems

Editors: List of editors' names

Abstract

Tracking controllers enable robotic systems to accurately follow planned reference trajectories. In particular, reinforcement learning (RL) has shown promise in the synthesis of controllers for systems with complex dynamics and modest online compute budgets. However, the poor sample efficiency of RL and the challenges of reward design make training slow and sometimes unstable, especially for high-dimensional systems. In this work, we leverage the inherent Lie group symmetries of robotic systems with a floating base to mitigate these challenges when learning tracking controllers. We model a general tracking problem as a Markov decision process (MDP) that captures the evolution of both the physical and reference states. Next, we show that symmetry in the underlying dynamics and running costs leads to an MDP homomorphism, a mapping that allows a policy trained on a lower-dimensional "quotient" MDP to be lifted to an optimal tracking controller for the original system. We compare this symmetry-informed approach to an unstructured baseline, using Proximal Policy Optimization (PPO) to learn tracking controllers for three systems: the `Particle` (a forced point mass), the `Astrobee` (a fully-actuated space robot), and the `Quadrotor` (an underactuated system). Results show that a symmetry-aware approach accelerates training and reduces tracking error after the same training duration.

Keywords: MDP Homomorphisms, Reinforcement Learning, Lie groups, Equivariance, Robotics, Trajectory tracking

1. Introduction

To achieve real-time operation, most robotic systems utilize a "tracking controller" to stabilize a pre-planned reference trajectory. However, tracking controllers designed analytically often assume properties not enjoyed by all robotic systems (*e.g.*, "full actuation" [3, 12, 27] or "differential flatness" [4]), while optimization-based methods frequently rely on linearization or simplified models to meet compute constraints [16]. In contrast, controllers trained via reinforcement learning (RL) have relaxed structural assumptions while enabling real-time operation with moderate resources [9]. In [14], the authors train a single hovering policy for deployment across a range of quadrotors, generalizing satisfactorily to moving references. Meanwhile, massively parallel training of quadrupedal walking policies from high-dimensional observations enabled startling robustness to uneven terrain [22], and learned controllers augmented with adaptive feedforward compensation have been shown to reject large disturbances [8]. Unfortunately, all these benefits come at a price: RL tends to scale poorly with the size of the given Markov decision process (MDP), making it challenging to perform the exploration needed to discover high-performance policies.

Exploiting Symmetry in Reinforcement Learning To mitigate this burden, an RL agent should share experience across all those states that can be considered "equivalent" with respect to the reward and dynamics. Indeed, robotic systems enjoy substantial symmetry [15, 18, 17], which has been thoroughly exploited in analytical control design [7, 28, 6] and optimization [24]. In fact, many learned controllers have leveraged symmetry in an ad hoc or approximate manner (*e.g.*, penalizing the *error* between actual and reference states [14] or working in the body frame [8]). More formally, the optimal policy of an MDP with symmetry is equivariant (and its value function is invariant) [26], and neural architectures can be designed accordingly to improve sample efficiency and generalization [25].

Homomorphisms of Markov Decision Processes Instead of incorporating symmetry into the network architecture directly, Ravindran [20] proposed "MDP homomorphisms", which establish a mapping from the given MDP to one of lower dimension. There, a policy may be trained more easily (using standard tools) and then lifted back to the original setting. Such methods were originally restricted to discrete state and action spaces, necessitating coarse discretization of robotic tasks (which are naturally described on smooth manifolds). [29] explored related ideas in continuous state and action spaces, but assumed deterministic dynamics (whereas stochasticity is fundamental to many tasks). However, Panangaden et al. [19] recently extended the theory of homomorphisms of stochastic MDPs to the continuous setting, recovering analogous value equivalence and policy lifting results.

Our Contributions In this work, we exploit the Lie group symmetries of free-flying robotic systems to learn tracking controllers efficiently. We cast a general tracking control problem as a continuous MDP, using a stochastic process to model the (*a priori* unknown) reference trajectory. We show that this MDP inherits the symmetry enjoyed by the underlying dynamics and running costs, and prove a theorem showing that such symmetries can be used to construct an MDP homomorphism, reducing the dimensionality. (Due to space constraints, these rigorous mathematical arguments are relegated to the appendices.) We use these theoretical tools to learn tracking controllers for three example systems, showing that our symmetry-informed approach accelerates training, improves tracking accuracy, and generalizes zero-shot to new trajectories. Ultimately, these insights will facilitate the efficient development of accurate tracking controllers for various robotic systems.

2. Symmetry-Aware Methods for Learning Tracking Controllers

This section summarizes our approach. A detailed treatment is found in the appendices.

Tracking Control as a Markov Decision Process We formalize the *tracking control problem* for a physical system with dynamics $x_{t+1} \sim f(\,\cdot\,|\,x_t, u_t)$ (where $x \in \mathcal{X}$ and $u \in \mathcal{U}$) as the continuous MDP $\mathcal{M}_\mathcal{T} = (\mathcal{S} = \mathcal{X} \times \mathcal{X} \times \mathcal{U}, \mathcal{A} = \mathcal{U}, R, \tau, \gamma)$, where the MDP state is the concatenation of the *actual state* x, the *reference state* x^d, and the *reference action* u^d. The reward $R\big((x, x^\mathrm{d}, u^\mathrm{d}), u\big) := -J_\mathcal{X}(x, x^\mathrm{d}) - J_\mathcal{U}(u, u^\mathrm{d})$ incorporates *tracking* and *effort costs* $J_\mathcal{X}$ and $J_\mathcal{U}$, and states evolves as $x_{t+1} \sim f(\,\cdot\,|\,x_t, u_t)$, $x^\mathrm{d}_{t+1} \sim f(\,\cdot\,|\,x^\mathrm{d}_t, u^\mathrm{d}_t)$, and $u^\mathrm{d}_{t+1} \sim \rho$, where $\rho \in \Delta(\mathcal{U})$ is a stationary distribution. (See Defs. 6 and 7 in App. C). This allows us to model a tracking control problem over a broad class of reference trajectories as a single stationary MDP. We could alternately formulate a (*non-stationary*) MDP corresponding to a *particular* reference trajectory (by making the tracking cost a function of time t and the actual state x), but an optimal policy for that MDP would be useless for tracking *other* references, and the reference trajectory may differ between train-time and test-time. We show empirically (see Fig. 2) that policies trained on the proposed MDP also effectively track pre-planned reference trajectories, for which the sequence of reference actions $\{u_0^{\mathrm{d}}, u_1^{\mathrm{d}}, u_2^{\mathrm{d}}, \cdots\}$ is chosen to induce a pre-selected state trajectory $\{x_0^{\mathrm{d}}, x_1^{\mathrm{d}}, x_2^{\mathrm{d}}, \cdots\}$.

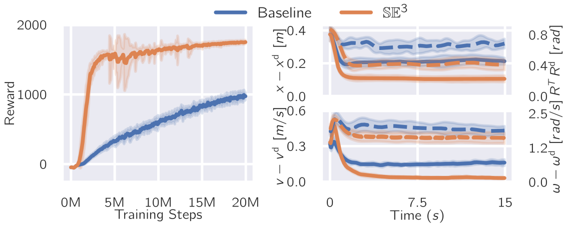
Symmetries of Tracking Control MDPs We prove (see Thm. 8 in App. C) that a tracking control MDP induced by dynamics and running costs with certain Lie group symmetries inherits a symmetry of its own, with certain nice properties. Moreover, we prove (see Thm. 5 in App. B) that a Lie group symmetry of this flavor can be used to construct a continuous MDP homomorphism in the sense of Panangaden et al. [19] (see App. A), who showed that such a homomorphism enables optimal policies for a "reduced" MDP of smaller dimension to be "lifted" to optimal policies for the original MDP. They also learned approximate homomorphisms from data, but did not give a sufficient condition to construct a nicely-behaved homomorphism (*i.e.*, for which the new state and action spaces are also smooth manifolds) from a continuous symmetry known *a priori* (as is the case for free-flying robotic systems). Since the random sampling of u^{d} makes $\mathcal{M}_{\mathcal{T}}$ stochastic (even when f is deterministic), we develop our method in the continuous, stochastic setting (although related results are known in the discrete [20] and deterministic [29] settings).

Application to Free-Flying Robotic Systems We apply the method to reduce the tracking MDP for three example systems, the `Particle`, the `Quadrotor`, and the `Astrobee`. Here, we briefly describe the `Astrobee` system (see App. D for a thorough presentation of all three systems). This space robot [2] has state $x = (q, \xi)$ in $\mathcal{X} = SE(3) \times \mathbb{R}^6$ (*i.e.*, the pose q and the twist $\xi = (\omega, v)$) and the action $u = (\mu, f)$ in $\mathcal{U} = \mathbb{R}^6$ (*i.e.*, the applied wrench). The system dynamics are defined by

$$q_{t+1} = q_t \exp(\hat{\xi}_t \,\mathrm{d}t), \qquad v_{t+1} = v_t + \tfrac{1}{m} f_t \,\mathrm{d}t, \qquad \omega_{t+1} = \omega_t + \mathbb{J}^{-1}(\mu_t - \omega_t \times \mathbb{J}\,\omega_t)\,\mathrm{d}t, \tag{1}$$

where $\hat{\cdot} : \mathbb{R}^6 \to \mathfrak{se}(3)$ and r and R are the $\mathbb{R}^3$ and $SO(3)$ components of q, and the running costs (given in App. D.3) depend only on $r - r^{\mathrm{d}}$, $R^{\mathrm{T}} R^{\mathrm{d}}$, and $\xi - \xi^{\mathrm{d}}$. It is well-known [18] and easily verified that this system has $SE(3)$ symmetry, which can be described by the group action $\Psi_k(q, \xi) := (kq, \xi)$. Using Theorem 8 to derive a symmetry of $\mathcal{M}_{\mathcal{T}}$, we apply Theorem 5 to obtain an MDP homomorphism (p, h), where $h_s = \mathrm{id}$ for all $s = (q, \xi, q^{\mathrm{d}}, \xi^{\mathrm{d}}, u^{\mathrm{d}}) \in \mathcal{S}$, and $p(s) := \big(q^{-1} q^{\mathrm{d}}, \xi, \xi^{\mathrm{d}}, u^{\mathrm{d}}\big)$. Hence, Theorem 3 in App. A (due to Panangaden et al. [19]) implies that there exists an optimal policy that observes only the *error* between the actual and reference poses, instead of observing these absolute poses directly.

Numerical Experiments Environments were implemented for the tracking control MDP of each system, written in `jax` [1] for performance (see App. D.1-D.3 and open-source code for details). To implement environments for the quotient MDP arising from reduction by a symmetry group, we modify each environment's observation to the reduced state given in (43), (53), and (59) (whereas the baseline sees the full-state observation $(x, x^{\mathrm{d}}, u^{\mathrm{d}})$). We also modify the actions according to the definition of h. For the `Particle` environment, we isolate the effects of reduction by different subgroups of the symmetry by also implementing environments reduced by translational symmetry alone (*i.e.*, $p(s) := (r - r^{\mathrm{d}}, v, v^{\mathrm{d}}, u^{\mathrm{d}})$) and by translational and velocity symmetry alone (*i.e.*, $p(s) := (r - r^{\mathrm{d}}, v - v^{\mathrm{d}}, u^{\mathrm{d}})$). We use a custom implementation of PPO [23] (see code for details), with the same hyperparameters across all variants of each environment. During training, the reference actions are sampled from a stationary distribution (as in Def. 7), but we evaluate zero-shot on pre-planned (dynamically feasible) reference trajectories. Fig. 2 and Table 1 report total reward (during training) and average tracking error (during evaluation) after randomizing the initial state.



Environment	$\mathcal{G}$	r [m]	v [m/s]	R [rad]	ω [rad/s]
`Particle`	Baseline	2.50±0.25	0.93±0.11	-	-
	$\mathbb{R}^3$	0.12±0.4	0.06±0.23	-	-
	$T\mathbb{R}^3$	0.11±0.41	**0.04±0.19**	-	-
	$T\mathbb{R}^3 \times \mathbb{R}^3$	**0.09±0.42**	**0.04±0.19**	-	-
`Astrobee`	Baseline	0.23±0.01	0.17±0.01	0.70±0.03	1.91±0.08
	$SE(3)$	**0.10±0.01**	**0.03±0.04**	**0.41±0.04**	**1.51±0.07**
`Quadrotor`	Baseline	0.91±0.62	0.97±0.78	0.51±0.19	0.27±0.21
	$SE(2) \times \mathbb{R}$	**0.25±0.30**	**0.14±0.26**	**0.29±0.02**	**0.04±0.07**

Figure 1: (a) Plots on the left depict the reward during training and tracking error components during evaluation for the `Astrobee` with translational errors as solid lines and rotational errors as dashed lines. (b) Table on the right depicts the comparison of RMS tracking errors on pre-planned trajectories. The mean and standard deviation are calculated over n = 20 training seeds of the policy's RMS tracking error, on a dataset of m = 20 trajectories.

3. Discussion

Benefits of Exploiting Symmetry Fig. 2 in App. E shows a clear trend across the board: greater symmetry exploitation leads to improved sample efficiency. The tracking error evaluation shown in Table 1 and Fig. 2 follows a similar trend. For the `Particle`, the vast majority of this benefit is achieved by reduction of the translational symmetry, although incorporating the velocity and force symmetries yields modest additional gains. This seems consistent with the large improvement we see for the `Astrobee` and `Quadrotor` after reduction by (a subgroup of) $SE(3)$. Careful reward engineering or hyperparameter tuning might improve performance (especially for the baselines that currently fail to learn effectively), but we instead focus on analyzing the benefit of exploiting symmetry for a fixed reward. Nonetheless, any reward depending only on the reduced state $\tilde{s} = p(s)$ would preserve the symmetry.

Limitations Our approach assumes that at deployment, an upstream planner provides dynamically feasible reference trajectories. For the (underactuated) `Quadrotor`, these trajectories were planned using differential flatness [13] from Lissajous curves in the flat space. However, in theory any other method (*e.g.*, direct collocation [10]) could be used to generate a suitable reference. We expect our policies to generalize well to a wide range of upstream planning methodologies, and future work should explore this hypothesis. Going forward, we also hope to apply these methods to new robot morphologies that are too complex for real-time numerical optimal control or for which no explicit analytical controllers are known.

4. Conclusion

In this work, we exploit the natural Lie group symmetries of free-flying robotic systems to mitigate the challenges of learning trajectory tracking controllers. We formulate the tracking problem as a single stationary MDP, proving that the underlying symmetries of the dynamics and running costs permit the reduction of this MDP to a lower-dimensional problem. When learning tracking controllers for space and aerial robots, training is accelerated and tracking error is reduced after the same number of training steps. We believe our theoretical framework provides valuable insight into more optimal use of RL for systems with symmetry in robotics applications.

References

[1] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

[2] M. Bualat, J. Barlow, T. Fong, C. Provencher, and T. Smith. Astrobee: Developing a free-flying robot for the international space station. In *AIAA SPACE 2015 conference and exposition*, page 4643, 2015.

[3] F. Bullo and R. M. Murray. Tracking for fully actuated mechanical systems: a geometric framework. *Automatica*, 35(1):17–34, 1999. ISSN 0005-1098.

[4] M. Fliess, J. Levine, P. Martin, F. Ollivier, and P. Rouchon. Controlling nonlinear systems by flatness. In C. I. Byrnes, B. N. Datta, C. F. Martin, and D. S. Gilliam, editors, *Systems and Control in the Twenty-First Century*, pages 137–154, Boston, MA, 1997. Birkhäuser Boston. ISBN 978-1-4612-4120-1.

[5] J. Gallier and J. Quaintance. *Differential Geometry and Lie Groups, A Second Course*, volume 12 of *Geometry and Computing*. Springer, 2020.

[6] M. Hampsey, P. van Goor, T. Hamel, and R. Mahony. Exploiting different symmetries for trajectory tracking control with application to quadrotors. *IFAC-PapersOnLine*, 56(1):132–137, 2023. ISSN 2405-8963. 12th IFAC Symposium on Nonlinear Control Systems NOLCOS 2022.

[7] R. L. Hatton, Z. Brock, S. Chen, H. Choset, H. Faraji, R. Fu, N. Justus, and S. Ramasamy. The geometry of optimal gaits for inertia-dominated kinematic systems. *IEEE Transactions on Robotics*, 38(5):3279–3299, 2022. doi: 10.1109/TRO.2022.3164595.

[8] K. Huang, R. Rana, A. Spitzer, G. Shi, and B. Boots. DATT: Deep Adaptive Trajectory Tracking for Quadrotor Control. In J. Tan, M. Toussaint, and K. Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 326–340. PMLR, 06–09 Nov 2023.

[9] J. Hwangbo, I. Sa, R. Y. Siegwart, and M. Hutter. Control of a Quadrotor With Reinforcement Learning. *IEEE Robotics and Automation Letters*, 2:2096–2103, 2017.

[10] M. Kelly. An introduction to trajectory optimization: How to do your own direct collocation. *SIAM Review*, 59(4):849–904, 2017.

[11] J. M. Lee. *Introduction to Smooth Manifolds*. Springer New York, second edition, 2013.

[12] D. Maithripala, J. Berg, and W. Dayawansa. Almost-global tracking of simple mechanical systems on a general class of lie groups. *IEEE Transactions on Automatic Control*, 51(2):216–225, 2006. doi: 10.1109/TAC.2005.862219.

[13] D. Mellinger and V. Kumar. Minimum snap trajectory generation and control for quadrotors. In *2011 IEEE International Conference on Robotics and Automation*, pages 2520–2525, 2011. doi: 10.1109/ICRA.2011.5980409.

[14] A. Molchanov, T. Chen, W. Hönig, J. A. Preiss, N. Ayanian, and G. S. Sukhatme. Sim-to-(Multi)-Real: Transfer of Low-Level Robust Control Policies to Multiple Quadrotors. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 59–66, 2019. doi: 10.1109/IROS40897.2019.8967695.

[15] R. M. Murray. Nonlinear control of mechanical systems: A Lagrangian perspective. *Annual Reviews in Control*, 21:31–42, 1997. ISSN 1367-5788.

[16] K. Nguyen, S. Schoedel, A. Alavilli, B. Plancher, and Z. Manchester. TinyMPC: Model-Predictive Control on Resource-Constrained Microcontrollers. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[17] D. F. Ordonez-Apraez, M. Martin, A. Agudo, and F. Moreno. On discrete symmetries of robotics systems: A group-theoretic and data-driven analysis. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.053.

[18] J. Ostrowski. Computing reduced equations for robotic systems with constraints and symmetries. *IEEE Transactions on Robotics and Automation*, 15(1):111–123, 1999. doi: 10.1109/70.744607.

[19] P. Panangaden, S. Rezaei-Shoshtari, R. Zhao, D. Meger, and D. Precup. Policy Gradient Methods in the Presence of Symmetries and State Abstractions. *Journal of Machine Learning Research*, 25(71):1–57, 2024.

[20] B. Ravindran. *An algebraic approach to abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2004.

[21] S. Rezaei-Shoshtari, R. Zhao, P. Panangaden, D. Meger, and D. Precup. Continuous MDP Homomorphisms and Homomorphic Policy Gradient. In *Advances in Neural Information Processing Systems*, volume 35, pages 20189–20204, 2022.

[22] N. Rudin, D. Hoeller, P. Reist, and M. Hutter. Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 91–100. PMLR, 08–11 Nov 2022.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[24] S. Teng, D. Chen, W. Clark, and M. Ghaffari. An Error-State Model Predictive Control on Connected Matrix Lie Groups for Legged Robot Control. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8850–8857, 2022. doi: 10.1109/IROS47612.2022.9981282.

[25] E. van der Pol, D. Worrall, H. van Hoof, F. Oliehoek, and M. Welling. MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4199–4210. Curran Associates, Inc., 2020.

[26] D. Wang, R. Walters, X. Zhu, and R. Platt. Equivariant Q Learning in Spatial Action Spaces. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 1713–1723. PMLR, 08–11 Nov 2022.

[27] J. Welde and V. Kumar. Almost Global Asymptotic Trajectory Tracking for Fully-Actuated Mechanical Systems on Homogeneous Riemannian Manifolds. *IEEE Control Systems Letters*, 8:724–729, 2024. doi: 10.1109/LCSYS.2024.3396565.

[28] J. Welde, M. D. Kvalheim, and V. Kumar. The Role of Symmetry in Constructing Geometric Flat Outputs for Free-Flying Robotic Systems. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12247–12253, 2023. doi: 10.1109/ICRA48891.2023.10160628.

[29] B. Yu and T. Lee. Equivariant Reinforcement Learning for Quadrotor UAV. In *2023 American Control Conference (ACC)*, pages 2842–2847, 2023. doi: 10.23919/ACC55779.2023.10156379.

[30] R. Y. Zhao. Continuous Homomorphisms and Leveraging Symmetries in Policy Gradient Algorithms for Markov Decision Processes. Master's thesis, McGill University, 2022.

Appendix A. Continuous Markov Decision Processes

In this appendix, we review background and preliminary notions regarding Markov decision processes with continuous state and action spaces and homomorphisms between them. We follow the treatment of Panangaden et al. [19], who (along with Rezaei-Shoshtari et al. [21]) extended the work of Ravindran [20] to study homomorphisms of Markov decision processes with continuous (*i.e.*, not discrete) state and action spaces.

Definition 1 (see [19]) *A* continuous Markov decision process[1] *(briefly, an MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, \tau, \gamma)$, where:*

- *the* state space $\mathcal{S}$ *is a smooth manifold,*
- *the* action space $\mathcal{A}$ *is a smooth manifold,*
- *the* instantaneous reward *is* $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$*,*
- *the* transition dynamics *are* $\tau : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$*, and*
- *the* discount factor γ *is a value in the interval* $[0, 1)$*.*

After taking action a_t from state s_t, the probability that s_{t+1} is contained in a set $B \in \mathsf{B}(\mathcal{S})$ is given by $\tau(B \,|\, s_t, a_t)$. A *policy* for $\mathcal{M}$ is a map $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. The *action-value function* $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a given policy π is defined by

$$Q^\pi(s, a) := \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a \right], \tag{2}$$

[1] The more general definition of Panangaden et al. [19] does *not* assume $\mathcal{S}$ and $\mathcal{A}$ are smooth manifolds, nor that $\tau(\,\cdot \,|\, s, a)$ is a Borel measure, but the present level of generality is all we need.

where $\tau \sim \pi$ denotes the expectation over both the transitions and the policy (*i.e.*, $s_{t+1} \sim \tau(\,\cdot \,|\, s_t, a_t)$ and $a_t \sim \pi(\,\cdot\,|\, s_t)$ for all $t \in \mathbb{N}$). A policy π^* is *optimal* if, for all $s \in \mathcal{S}$,

$$\pi^* = \arg\max_{\pi} \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s \right]. \tag{3}$$

A.1. Homomorphisms of Markov Decision Processes

The following notion describes a powerful relationship between two continuous MDPs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ of (perhaps) different dimensions.

Definition 2 (see Panangaden et al. [19, Defs. 11 and 14]) *A pair of maps* $p : \mathcal{S} \to \widetilde{\mathcal{S}}$ *and* $h : \mathcal{S} \times \mathcal{A} \to \widetilde{\mathcal{A}}$ *is called a* continuous MDP homomorphism *from* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, \tau, \gamma)$ *to* $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}, \widetilde{R}, \widetilde{\tau}, \gamma)$ *if* p *and, for each* $s \in \mathcal{S}$*, the map* $h_s : a \mapsto h(s, a)$ *are measurable, surjective maps, such that*

$$R\big(s, a\big) = \widetilde{R}\big(p(s), h(s, a)\big), \tag{4a}$$

$$\tau\big(p^{-1}(\widetilde{B}) \,|\, s, a\big) = \widetilde{\tau}\big(\widetilde{B} \,|\, p(s), h(s, a)\big) \tag{4b}$$

for all $s \in \mathcal{S}$*,* $a \in \mathcal{A}$*, and* $\widetilde{B} \in \mathsf{B}(\widetilde{S})$*. Given a continuous MDP homomorphism* (p, h)*, a policy* π *for* $\widetilde{\mathcal{M}}$*, and a policy* $\widetilde{\pi}$ *for* $\mathcal{M}$*,* π *is called a lift of* $\widetilde{\pi}$ *if for all* $s \in \mathcal{S}$ *and* $A \in \mathsf{B}(\mathcal{A})$*,*

$$\pi\big(h_s^{-1}(\widetilde{A}) \,|\, s\big) = \widetilde{\pi}\big(\widetilde{A} \,|\, p(s)\big). \tag{5}$$

Subsequently, we often omit the word "continuous" for brevity. MDP homomorphisms facilitate the synthesis of an optimal policy for the original MDP $\mathcal{M}$ from an optimal policy for the "quotient" MDP $\widetilde{\mathcal{M}}$, via the following theorem.

Theorem 3 (see Panangaden et al. [19, Thms. 12 and 16]) *Suppose* (p, h) *is an MDP homomorphism from* $\mathcal{M}$ *to* $\widetilde{\mathcal{M}}$ *and* π *is a lift of* any *policy* $\widetilde{\pi}$ *for* $\widetilde{\mathcal{M}}$*. Then,* $Q^{\pi}(s, a) = \widetilde{Q}^{\widetilde{\pi}}\big(p(s), h(s, a)\big)$*. Moreover, if* $\widetilde{\pi}$ *is optimal for* $\widetilde{\mathcal{M}}$*, then* π *is optimal for* $\mathcal{M}$*.*

Appendix B. Continuous MDPs with Lie Group Symmetries

In this appendix, we present a theorem which uses a certain kind of Lie group symmetry of a continuous MDP to construct a continuous MDP homomorphism, reducing the dimension of the state space by that of the symmetry group.

We first recall the basics of group actions. A *(left) group action* of a Lie group $\mathcal{G}$ on a smooth manifold $\mathcal{X}$ is a smooth map $\Phi : \mathcal{G} \times \mathcal{X} \to \mathcal{X}$ (often written $\Phi_g(x) := \Phi(g, x)$ for brevity) such that for all $x \in \mathcal{X}$ and $g, h \in \mathcal{G}$, $\Phi(1_{\mathcal{G}}, x) = x$ (where $1_{\mathcal{G}} \in \mathcal{G}$ is the identity) and $\Phi\big(g, \Phi(h, x)\big) = \Phi\big(gh, x\big)$. The Φ-*orbit* of x is the set $\Phi_G(x) := \{\Phi_g(x) : g \in \mathcal{G}\}$, while $\mathcal{X}/\mathcal{G}$ is a set whose elements are all the orbits of Φ. An action Φ is *proper* if the map $(g, x) \mapsto \big(\Phi_g(x), x\big)$ is proper (*i.e.*, the preimage of any compact set is compact), and *free* if $\Phi_g(x) = x$ implies $g = 1_{\mathcal{G}}$. A group $\mathcal{G}$ acts on itself via $L : (h, g) \mapsto hg$.

B.1. Lie Group Symmetries of Markov Decision Processes

We now formulate the following definition of a Lie group symmetry of a continuous MDP, which differs from some prior literature.

Definition 4 *Given an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, \tau, \gamma)$, a pair of actions (Φ, Ψ) of the Lie group $\mathcal{G}$ on $\mathcal{S}$ and $\mathcal{A}$ respectively is called a* Lie group symmetry *of $\mathcal{M}$ if, for all Φ-invariant sets $B \in \mathsf{B}(\mathcal{S})$ and all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $g \in \mathcal{G}$, we have*

$$R\big(s, a\big) = R\big(\Phi_g(s), \Psi_g(a)\big), \tag{6a}$$

$$\tau\big(B \,|\, s, a\big) = \tau\big(B \,|\, \Phi_g(s), \Psi_g(a)\big). \tag{6b}$$

The qualifier "Φ-invariant" on B broadens the class of symmetries considered, and this is more general than [26] and [25], as noted in [30, Def. 35]. The deterministic case (*i.e.*, when $\tau(\,\cdot\,|\, s_t, a_t)$ is the Dirac measure corresponding to $\{s_{t+1}\} \subseteq \mathcal{S}$) gives the intuition, since then (6b) requires the image of any orbit in $\mathcal{S} \times \mathcal{A}$ to lie within some orbit in $\mathcal{S}$, *without* enforcing equivariance *within* each orbit.

B.2. Continuous MDP Homomorphisms Induced by Lie Group Symmetries

The following theorem shows that Lie group symmetries with certain properties can be used to construct a continuous MDP homomorphism and to lift "downstairs" policies. Related results are known in the discrete [20] and deterministic [29] settings, and [19] learned data-driven *approximate* homomorphisms for systems with similar properties, we do not know of a prior result for Lie group symmetries known *a priori* for a continuous MDP.

Theorem 5 *Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, \tau, \gamma)$ with a Lie group symmetry (Φ, Ψ). Suppose that Φ is free and proper and $\lambda : \mathcal{S} \to \mathcal{G}$ is any[2] equivariant map. Define*

$$p : S \to \mathcal{S}/\mathcal{G}, \;\; s \mapsto \Phi_G(s), \tag{7a}$$

$$h : \mathcal{S} \times \mathcal{A} \to \mathcal{A}, \;\; (s, a) \mapsto \Psi_{\lambda(s)^{-1}}(a). \tag{7b}$$

Then, (p, h) is an MDP homomorphism from $\mathcal{M}$ to $\widetilde{\mathcal{M}} = \big(\widetilde{\mathcal{S}} = \mathcal{S}/\mathcal{G}, \widetilde{\mathcal{A}} = \mathcal{A}, \widetilde{R}, \widetilde{\tau}, \gamma\big)$, where

$$\widetilde{R}(\tilde{s}, \tilde{a}) := R\big(s, \Psi_{\lambda(s)}(\tilde{a})\big)\,\big|_{\,s \in p^{-1}(\tilde{s})}, \tag{8a}$$

$$\widetilde{\tau}(\widetilde{B} \,|\, \tilde{s}, \tilde{a}) := \tau\big(p^{-1}(\widetilde{B}) \,|\, s, \Psi_{\lambda(s)}(\tilde{a})\big)\,\big|_{\,s \in p^{-1}(\tilde{s})} \tag{8b}$$

independent of the particular choice of s. Also, for any policy $\widetilde{\pi}$ for $\widetilde{\mathcal{M}}$, a policy for $\mathcal{M}$ that is a lift of $\widetilde{\pi}$ is given by

$$(\widetilde{\pi})^{\uparrow}(A \,|\, s) := \widetilde{\pi}\big(\Psi_{\lambda(s)^{-1}}(A) \,|\, p(s)\big). \tag{9}$$

Proof Because Φ is free and proper, $\mathcal{S}/\mathcal{G}$ is a smooth manifold of dimension $\dim \mathcal{S} - \dim \mathcal{G}$ [11, Thm. 21.10]. We first verify that $\widetilde{R}$ and $\widetilde{\tau}$ are well-defined (*i.e.*, their values do not

[2] Since λ need not be continuous, it can be constructed from a collection of *local* trivializations of the principal $\mathcal{G}$-bundle $p : \mathcal{S} \to \mathcal{S}/\mathcal{G}$ [5, §9.9].

depend on the particular choice of $s \in p^{-1}(\tilde{s})$). Since p maps states to Φ-orbits, for any $s_1, s_2 \in p^{-1}(\tilde{s})$, there exists some $g \in \mathcal{G}$ such that $s_1 = \Phi_g(s_2)$. Thus, following (8a),

$$\begin{align}
R(\tilde{s}, \tilde{a}) &= R\big(s_1, \Psi_{\lambda(s_1)}(\tilde{a})\big) \tag{10}\\
&= R\big(\Phi_g(s_2), \Psi_{g\lambda(s_2)}(\tilde{a})\big) \tag{11}\\
&= R\big(s_2, \Psi_{\lambda(s_2)}(\tilde{a})\big) = R(\tilde{s}, \tilde{a}), \tag{12}
\end{align}$$

where (11) follows from the equivariance of λ and the invariance of the reward. Similarly, from (8b), we compute

$$\begin{align}
\widetilde{\tau}(\widetilde{B} \,|\, \tilde{s}, \tilde{a}) &= \tau\big(p^{-1}(\widetilde{B}) \,|\, s_1, \Psi_{\lambda(s_1)}(\tilde{a})\big) \tag{13}\\
&= \tau\big(p^{-1}(\widetilde{B}) \,|\, \Phi_g(s_2), \Psi_{g\lambda(s_2)}(\tilde{a})\big) \tag{14}\\
&= \tau\big(p^{-1}(\widetilde{B}) \,|\, s_2, \Psi_{\lambda(s_2)}(\tilde{a})\big) = \widetilde{\tau}(\widetilde{B} \,|\, \tilde{s}, \tilde{a}), \tag{15}
\end{align}$$

where (15) follows from (6b), since for any $\widetilde{B} \in \mathsf{B}(\widetilde{\mathcal{S}})$, the Borel set $p^{-1}(\widetilde{B}) \in \mathsf{B}(\mathcal{S})$ is Φ-invariant. We now verify the MDP homomorphism. Since for each $g \in \mathcal{G}$, the map Ψ_g is a diffeomorphism, h_s is measurable and surjective for each $s \in \mathcal{S}$. On the other hand, p is surjective by construction and measurable because orbits of proper actions are closed [11, Cor. 21.8]. Since $s \in p^{-1}\big(p(s)\big)$, we have

$$\widetilde{R}\big(p(s), h(s,a)\big) = R\big(s, \Psi_{\lambda(s)} \circ \Psi_{\lambda(s)^{-1}}(a)\big) = R(s,a),$$

hence (4a) holds. We verify (4b) similarly, since by (8b),

$$\begin{align}
\widetilde{\tau}\big(\widetilde{B} \,|\, p(s), h(s,a)\big) &= \tau\big(p^{-1}(\widetilde{B}) \,|\, s, \Psi_{\lambda(s)} \circ \Psi_{\lambda(s)^{-1}}(a)\big) \tag{16}\\
&= \tau\big(p^{-1}(\widetilde{B}) \,|\, s, a\big). \tag{17}
\end{align}$$

Thus, (p,h) is an MDP homomorphism from $\mathcal{M}$ to $\widetilde{\mathcal{M}}$. Finally, to see that $(\widetilde{\pi})^{\uparrow}$ is a lift of $\widetilde{\pi}$, we compute

$$\begin{align}
(\widetilde{\pi})^{\uparrow}\big(h_s^{-1}(\widetilde{A}) \,|\, s\big) &= (\widetilde{\pi})^{\uparrow}\big(\Psi_{\lambda(s)}(\widetilde{A}) \,|\, s\big) \tag{18}\\
&= \widetilde{\pi}\big(\Psi_{\lambda(s)^{-1}} \circ \Psi_{\lambda(s)}(\widetilde{A}) \,|\, p(s)\big) \tag{19}\\
&= \widetilde{\pi}\big(\widetilde{A} \,|\, p(s)\big), \tag{20}
\end{align}$$

where (18) and (19) follow directly from (7b) and (9), and the fact that Ψ_g is a diffeomorphism for all $g \in \mathcal{G}$. ■

Appendix C. Tracking Control Problems with Lie Group Symmetries

In this appendix, we formulate a general trajectory tracking problem as an MDP that models the evolution of both the physical and reference systems. We give a sufficient condition for this MDP to have a Lie group symmetry that can be used to reduce the problem size.

Definition 6 *A* tracking control problem *is a tuple* $\mathcal{T} = (\mathcal{X}, \mathcal{U}, f, J_{\mathcal{X}}, J_{\mathcal{U}}, \rho, \gamma)$*, where:*

- $\mathcal{X}$ *is the* physical state space *(a smooth manifold),*
- $\mathcal{U}$ *is the* physical action space *(a smooth manifold),*
- $f : \mathcal{X} \times \mathcal{U} \to \Delta(\mathcal{X})$ *is the the* physical dynamics*,* i.e.*,* $x_{t+1} \sim f(\,\cdot\,|\, x_t, u_t)$*,*
- $J_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is the* tracking cost*,*
- $J_{\mathcal{U}} : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ *is the* effort cost*,*
- $\rho \in \Delta(\mathcal{U})$ *is the* reference action distribution*, and*
- $\gamma \in [0, 1)$ *is the* discount factor*.*

The distribution ρ is not usually included in the definition of a tracking problem, but it will play an essential role in our approach.

C.1. Modeling a Tracking Control Problem as an MDP

We model the tracking task for *a priori* unknown references in the following manner.

Definition 7 *A given tracking control problem* $\mathcal{T} = \big(\mathcal{X}, \mathcal{U}, f, J_{\mathcal{X}}, J_{\mathcal{U}}, \rho, \gamma\big)$ *induces a* tracking control MDP *given by* $\mathcal{M}_{\mathcal{T}} = (\mathcal{S} = \mathcal{X} \times \mathcal{X} \times \mathcal{U}, \mathcal{A} = \mathcal{U}, R, \tau, \gamma)$*, where:*

- *the state is* $(x, x^{\mathrm{d}}, u^{\mathrm{d}})$*, where* x*,* $x^{\mathrm{d}} \in \mathcal{X}$ *are the actual and reference states and* $u^{\mathrm{d}} \in \mathcal{U}$ *is the reference action,*
- *the actions are* $a = u \in \mathcal{U}$ *(i.e., the actual action),*
- *the instantaneous reward* $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is given by*
$$R\big((x, x^{\mathrm{d}}, u^{\mathrm{d}}), u\big) := -J_{\mathcal{X}}(x, x^{\mathrm{d}}) - J_{\mathcal{U}}(u, u^{\mathrm{d}}), \tag{21}$$
- *and the transitions* $\tau : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ *are defined by*
$$x_{t+1} \sim f(\,\cdot\,|\, x_t, u_t), \quad x^{\mathrm{d}}_{t+1} \sim f(\,\cdot\,|\, x^{\mathrm{d}}_t, u^{\mathrm{d}}_t), \quad u^{\mathrm{d}}_{t+1} \sim \rho. \tag{22}$$

C.2. Symmetries of Tracking Control MDPs

In this section, we will show that the MDP induced by a tracking control problem with certain symmetries will inherit a related symmetry with certain convenient properties.

We now prove the primary result.

Theorem 8 *Consider a tracking control problem* $\mathcal{T} = (\mathcal{X}, \mathcal{U}, f, J_{\mathcal{X}}, J_{\mathcal{U}}, \rho, \gamma)$ *as well as Lie group actions* $\Upsilon : \mathcal{K} \times \mathcal{X} \to \mathcal{X}$ *and* $\Theta : \mathcal{H} \times \mathcal{U} \to \mathcal{U}$*. Suppose that* $J_{\mathcal{X}}$ *is* Υ*-invariant and* $J_{\mathcal{U}}$ *is* Θ*-invariant,* i.e.*, for all* $x, x^{\mathrm{d}} \in \mathcal{X}$*,* $u, u^{\mathrm{d}} \in \mathcal{U}$*,* $k \in \mathcal{K}$*, and* $h \in \mathcal{H}$*, we have*
$$J_{\mathcal{X}}(x, x^{\mathrm{d}}) = J_{\mathcal{X}}\big(\Upsilon_k(x), \Upsilon_k(x^{\mathrm{d}})\big), \quad J_{\mathcal{U}}(u, u^{\mathrm{d}}) = J_{\mathcal{U}}\big(\Theta_h(u), \Theta_h(u^{\mathrm{d}})\big). \tag{23}$$
Suppose also that for each $(k, h) \in \mathcal{K} \times \mathcal{H}$*, there exists* $k' \in \mathcal{K}$ *such that for all* $(x, u) \in \mathcal{X} \times \mathcal{U}$ *and* $B \in \mathsf{B}(\mathcal{X})$*, we have*
$$f\big(\Upsilon_{k'}(B) \,|\, x, u\big) = f\big(B \,|\, \Upsilon_k(x), \Psi_h(u)\big). \tag{24}$$
Define actions of the direct product $\mathcal{G} = \mathcal{K} \times \mathcal{H}$ *on* $\mathcal{S} = \mathcal{X} \times \mathcal{X} \times \mathcal{U}$ *and* $\mathcal{A} = \mathcal{U}$ *given by*
$$\Phi_{(k,h)}(x, x^{\mathrm{d}}, u^{\mathrm{d}}) := \big(\Upsilon_k(x), \Upsilon_k(x^{\mathrm{d}}), \Theta_h(u^{\mathrm{d}})\big), \quad \Psi_{(k,h)}(u) := \Theta_h(u). \tag{25}$$
Then, (Φ, Ψ) *is a Lie group symmetry of* $\mathcal{M}_{\mathcal{T}}$*. Moreover, if* Υ *and* Θ *are free and proper, then* Φ *is also free and proper.*

Proof From (21), we compute the transformed reward as

$$R\big(\Phi_{(k,h)}(s), \Psi_{(k,h)}(a)\big) = -J_{\mathcal{X}}\big(\Upsilon_k(x), \Upsilon_k(x^{\mathrm{d}})\big) - J_{\mathcal{U}}\big(\Theta_h(u), \Theta_h(u^{\mathrm{d}})\big) \tag{26}$$

$$= -J_{\mathcal{X}}(x, x^{\mathrm{d}}) - J_{\mathcal{U}}(u, u^{\mathrm{d}}) = R(s, a), \tag{27}$$

where we have substituted in (25) and simplified using (23). Thus, (6a) holds. Considering now the transitions, we note that (22) can also be written using the "product measure" as

$$\tau\big(\,\cdot\,|\,(x, x^{\mathrm{d}}, u^{\mathrm{d}}), u\big) := f(\,\cdot\,|\,x, u) \times f(\,\cdot\,|\,x^{\mathrm{d}}, u^{\mathrm{d}}) \times \rho. \tag{28}$$

We then apply (25) to (28) to compute

$$\tau\big(\,\cdot\,|\,\Phi_{(k,h)}(s), \Psi_{(k,h)}(a)\big) = f\big(\,\cdot\,|\,\Upsilon_k(x), \Theta_h(u)\big) \times f\big(\,\cdot\,|\,\Upsilon_k(x^{\mathrm{d}}), \Theta_h(u^{\mathrm{d}})\big) \times \rho \tag{29}$$

$$= f\big(\Upsilon_{k'}(\,\cdot\,)|\,x, u\big) \times f\big(\Upsilon_{k'}(\,\cdot\,)|\,x^{\mathrm{d}}, u^{\mathrm{d}}\big) \times \rho, \tag{30}$$

where (30) follows from (24). Considering any Φ-invariant $B \in \mathsf{B}(\mathcal{S})$, we note that $B = \Phi_{(k'^{-1}, 1_{\mathcal{H}})}(B)$, and compute

$$\tau\big(B\,|\,\Phi_{(k,h)}(s), \Psi_{(k,h)}(a)\big) = \tau\big(\Phi_{(k'^{-1}, 1_{\mathcal{H}})}(B)\,|\,\Phi_{(k,h)}(s), \Psi_{(k,h)}(a)\big) \tag{31}$$

$$= \big(f(\,\cdot\,|\,x, u) \times f(\,\cdot\,|\,x^{\mathrm{d}}, u^{\mathrm{d}}) \times \rho\big)(B) = \tau\big(B\,|\,s, a\big), \tag{32}$$

where (32) follows directly from (30) and (25). Thus, (6b) holds as well, and (Φ, Ψ) is a Lie group symmetry of $\mathcal{M}_{\mathcal{T}}$. Assuming that Υ and Θ are free and proper, it is readily verified that Φ is free and proper after noting that Φ is the product action of Γ and Θ (*i.e.*, $\Phi_{(k,h)} = \Gamma_k \times \Theta_h$), where Γ is the diagonal action of Υ (*i.e.*, $\Gamma_k = \Upsilon_k \times \Upsilon_k$). ■

Note that because we do *not* assume that $k' = k$, (24) is more general than equivariance of the transitions. However, k' must depend only on k and h, and not on x and u.

Appendix D. Application to Free-Flying Robotic Systems

In this appendix, we apply the proposed method to three example systems, showing a detailed worked example for a simple, pedagogical system and summarizing the method as applied to two realistic free-flying robotic systems.

D.1. The `Particle` System

Consider a particle in $\mathbb{R}^3$ with mass m subject to a controlled external force (sometimes used as a reduced-order model for a quadrotor or rocket as in, *e.g.*, Huang et al. [8]). The state $x = (r, v) \in \mathcal{X} = T\mathbb{R}^3 \simeq \mathbb{R}^3 \times \mathbb{R}^3$ consists of the particle's position and velocity, and the control input is the applied force $u \in \mathcal{U} = \mathbb{R}^3$. The (deterministic) equations of motion, when discretized with timestep $\mathrm{d}t$, are given by

$$r_{t+1} = r_t + v_t\,\mathrm{d}t, \quad v_{t+1} = v_t + \tfrac{1}{m} u_t\,\mathrm{d}t, \tag{33}$$

so the transition probabilities $f : T\mathbb{R}^3 \times \mathbb{R}^3 \to \Delta(T\mathbb{R}^3)$ are

$$f(B\,|\,x, u) := \begin{cases} 1, \ (r + v\,\mathrm{d}t, v + \tfrac{1}{m} u\,\mathrm{d}t) \in B, \\ 0, \ \text{otherwise}. \end{cases} \tag{34}$$

For some $c_r, c_v, c_u \geq 0$, we define the running costs

$$J_{T\mathbb{R}^3}\big((r,v),(r^{\mathrm{d}},v^{\mathrm{d}})\big) := \alpha(r - r^{\mathrm{d}}) + c_v\|v - v^{\mathrm{d}}\|, \tag{35a}$$

$$J_{\mathbb{R}^3}(u, u^{\mathrm{d}}) := c_u\|u - u^{\mathrm{d}}\|, \tag{35b}$$

where $\alpha(y) := c_r\|y\| + \tanh(a_r\|y\|) - 1$. Selecting a covariance Σ and a discount factor $0 \leq \gamma < 1$, we define the tracking problem $\mathcal{T} = \big(T\mathbb{R}^3, \mathbb{R}^3, f, J_{T\mathbb{R}^3}, J_{\mathbb{R}^3}, \mathcal{N}(0,\Sigma), \gamma\big)$.

Following (28), the dynamics of $\mathcal{M}_{\mathcal{T}}$ for `Particle` can be expressed as

$$r_{t+1} = r_t + v_t\,\mathrm{dt}, \quad v_{t+1} = v_t + \tfrac{1}{m}u_t\,\mathrm{dt}, \tag{36a}$$

$$r^{\mathrm{d}}_{t+1} = r^{\mathrm{d}}_t + v^{\mathrm{d}}_t\,\mathrm{dt}, \quad v^{\mathrm{d}}_{t+1} = v^{\mathrm{d}}_t + \tfrac{1}{m}u^{\mathrm{d}}_t\,\mathrm{dt}, \tag{36b}$$

$$u^{\mathrm{d}}_{t+1} \sim \mathcal{N}(0,\Sigma), \tag{36c}$$

where $\big((r,v),(r^{\mathrm{d}},v^{\mathrm{d}}),u^{\mathrm{d}}\big) \in \mathcal{S} = T\mathbb{R}^3 \times T\mathbb{R}^3 \times \mathbb{R}^3$ and $u \in \mathcal{A} = \mathbb{R}^3$. From (21), the reward is given by

$$R(s,a) = -\alpha(r - r^{\mathrm{d}}) - c_v\|v - v^{\mathrm{d}}\| - c_u\|u - u^{\mathrm{d}}\|. \tag{37}$$

Considering the Lie groups $\mathcal{K} := T\mathbb{R}^3$ (with the group operation inherited from its identification with $\mathbb{R}^3 \times \mathbb{R}^3$) and $\mathcal{H} := \mathbb{R}^3$, we let an $\mathcal{K}$-action on $\mathcal{S} = T\mathbb{R}^3$ and an $\mathcal{H}$-action on $\mathcal{A} = \mathbb{R}^3$ be given by the left action of the groups on themselves, *i.e.*,

$$\Upsilon_k(r,v) := L_{(k_1,k_2)}(r,v) = (r + k_1, v + k_2), \tag{38a}$$

$$\Theta_h(u) := L_h(u) = u + h, \tag{38b}$$

which are free and proper. It is clear that the tracking and effort costs (35) are invariant to these actions, *i.e.*, (23) holds. Moreover, for any $B \in \mathsf{B}(T\mathbb{R}^3)$,

$$\begin{aligned} f\big(B \,|\, \Upsilon_k(x), \Theta_h(u)\big) &= f\big(B \,|\, (r + k_1, v + k_2), u + h\big) \\ &= \begin{cases} 1, & \begin{pmatrix} r + k_1 + (v + k_2)\,\mathrm{dt} \\ v + k_2 + \frac{1}{m}(f + h)\,\mathrm{dt} \end{pmatrix} \in B, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{39}$$

$$= f\big(\Upsilon_{k'}(B) \,|\, x, u\big), \quad k' = -(k_1,k_2) - (k_2, \tfrac{1}{m}h)\,\mathrm{dt}. \tag{40}$$

Thus, the transitions satisfy (24). In the manner of (25), the group actions (38) induce actions of $\mathcal{G} = \mathcal{K} \times \mathcal{H} = T\mathbb{R}^3 \times \mathbb{R}^3$ on $\mathcal{S}$ and $\mathcal{A}$, given by

$$\Phi_{(k,h)}(s) := (r, v, r^{\mathrm{d}}, v^{\mathrm{d}}, u^{\mathrm{d}}) + (k_1, k_2, k_1, k_2, h), \tag{41a}$$

$$\Psi_{(k,h)}(a) := u^{\mathrm{d}} + h. \tag{41b}$$

Thus, by Theorem 8, (Φ, Ψ) is a Lie group symmetry of $\mathcal{M}_{\mathcal{T}}$ for the `Particle`, and moreover, Φ is free and proper. Using the symmetry (41), we will construct an MDP homomorphism using Theorem 5. We first define

$$\lambda\big((r,v),(r^{\mathrm{d}},v^{\mathrm{d}}),u^{\mathrm{d}}\big) := \big((r^{\mathrm{d}},v^{\mathrm{d}}),u^{\mathrm{d}}\big), \tag{42}$$

$$p\big((r,v),(r^{\mathrm{d}},v^{\mathrm{d}}),u^{\mathrm{d}}\big) := (r - r^{\mathrm{d}}, v - v^{\mathrm{d}}). \tag{43}$$

It is easily verified that λ is equivariant and p maps each state s to its Φ-orbit. We now define a quotient MDP $\widetilde{\mathcal{M}_\mathcal{T}}$ as described in Theorem 5. The state of $\widetilde{\mathcal{M}_\mathcal{T}}$ is $\tilde{s} = (r^{\mathrm{e}}, v^{\mathrm{e}}) \in \widetilde{\mathcal{S}} = \mathcal{S}/\mathcal{G} \simeq T\mathbb{R}^3$ and the actions are $\tilde{a} = u^{\mathrm{e}} \in \widetilde{A} = \mathbb{R}^3$. From (7b) and (42),

$$h(s, a) = \Psi_{(-r^{\mathrm{d}}, -v^{\mathrm{d}}, -u^{\mathrm{d}})}(u) = u - u^{\mathrm{d}}. \tag{44}$$

Since clearly $\big((r^{\mathrm{e}}, v^{\mathrm{e}}), (0, 0), 0\big) \in p^{-1}(r^{\mathrm{e}}, v^{\mathrm{e}})$, from (8a), (41b), and (42) we may construct the reduced reward as

$$\widetilde{R}(\tilde{s}, \tilde{a}) = R\big(s, \Psi_{\lambda(s)}(\tilde{a})\big)\Big|_{s=\left((r^{\mathrm{e}}, v^{\mathrm{e}}),(0,0),0\right)} = -\alpha(r^{\mathrm{e}}) - c_v\|v^{\mathrm{e}}\| - c_u\|u^{\mathrm{e}}\|. \tag{45}$$

Likewise, a straightforward calculation using (36), (41b), and (8b) will show that the reduced transitions are given by

$$\begin{aligned}\widetilde{\tau}(\widetilde{B} \mid \tilde{s}, \tilde{a}) &= \tau\big(p^{-1}(\widetilde{B}) \mid s, \Psi_{\lambda(s)}(\tilde{a})\big)\Big|_{s=\left((r^{\mathrm{e}}, v^{\mathrm{e}}),(0,0),0\right)} \\ &= \begin{cases} 1, \ (r^{\mathrm{e}} + v^{\mathrm{e}}\,\mathrm{d}t, v^{\mathrm{e}} + \frac{1}{m}u^{\mathrm{e}}\,\mathrm{d}t) \in \widetilde{B}, \\ 0, \text{ otherwise,}\end{cases}\end{aligned} \tag{46}$$

which is nothing but the usual "error dynamics" [12], *i.e.*

$$r^{\mathrm{e}}_{t+1} = r^{\mathrm{e}}_t + v^{\mathrm{e}}_t\,\mathrm{d}t, \quad v^{\mathrm{e}}_{t+1} = v^{\mathrm{e}}_t + \tfrac{1}{m}u^{\mathrm{e}}_t\,\mathrm{d}t. \tag{47}$$

Finally, by Theorem 5, (p, h) is an MDP homomorphism from $\mathcal{M}_\mathcal{T}$ to $\widetilde{\mathcal{M}_\mathcal{T}} = \big(T\mathbb{R}^3, \mathbb{R}^3, \widetilde{R}, \widetilde{\tau}, \gamma\big)$, and moreover we may lift any policy $\widetilde{\pi}$ for $\widetilde{\mathcal{M}_\mathcal{T}}$ to $\widetilde{\mathcal{M}}$ using (9), obtaining

$$(\widetilde{\pi})^{\uparrow}(A \mid s) = \widetilde{\pi}(A - u^{\mathrm{d}} \mid r - r^{\mathrm{d}}, v - v^{\mathrm{d}}). \tag{48}$$

By Theorem 3, the action-value function for $(\widetilde{\pi})^{\uparrow}$ satisfies

$$Q^{(\widetilde{\pi})^{\uparrow}}(s, a) = \widetilde{Q}^{\widetilde{\pi}}\big((r - r^{\mathrm{d}}, v - v^{\mathrm{d}}), u - u^{\mathrm{d}}\big). \tag{49}$$

Thus, an optimal policy can observe only the position and velocity error and augment the result with the reference force (*i.e.*, $u = \widetilde{\pi}(r - r^{\mathrm{d}}, v - v^{\mathrm{d}}) + u^{\mathrm{d}}$ for deterministic policies).

D.2. The `Astrobee` System

This space robot (described in Bualat et al. [2]) has state $x = (q, \xi)$ in $\mathcal{X} = SE(3) \times \mathbb{R}^6$ (*i.e.*, the pose q as a homogeneous transform and twist $\xi = (\omega, v)$) and action $u = (\mu, f)$ in $\mathcal{U} = \mathbb{R}^6$ (*i.e.*, the applied wrench). The dynamics are

$$q_{t+1} = q_t \exp(\hat{\xi}_t\,\mathrm{d}t), \tag{50a}$$

$$v_{t+1} = v_t + \tfrac{1}{m}f_t\,\mathrm{d}t, \quad \omega_{t+1} = \omega_t + \mathbb{J}^{-1}(\mu_t - \omega_t \times \mathbb{J}\,\omega_t)\,\mathrm{d}t, \tag{50b}$$

where $\hat{\cdot} : \mathbb{R}^6 \to \mathfrak{se}(3)$. The running costs are defined by

$$J_\mathcal{X}(x, x^{\mathrm{d}}) := \alpha(r - r^{\mathrm{d}}) + c_R\|\log(R^{\mathrm{T}}R^{\mathrm{d}})\| \ + c_\xi\|\xi - \xi^{\mathrm{d}}\|, \tag{51a}$$

$$J_\mathcal{U}(u, u^{\mathrm{d}}) := c_u\|u - u^{\mathrm{d}}\|, \tag{51b}$$

where r and R are the $\mathbb{R}^3$ and $SO(3)$ components of q. Letting $\rho = \mathcal{N}(0, \Sigma)$, we may construct $\mathcal{M}_\mathcal{T}$ as in Def. 7. Next, let $\mathcal{K} = SE(3)$ act on $\mathcal{X}$ and $\mathcal{H} = \{1\}$ act on $\mathcal{U}$ via

$$\Psi_k(q, \xi) := (kq, \xi), \quad \Theta_h(w) := w, \tag{52}$$

where (23) and (24) hold for these free and proper actions. Using Theorem 8 to derive a symmetry of $\mathcal{M}_\mathcal{T}$ as in (25), we apply Theorem 5 with $\lambda : s \mapsto q$ to ultimately obtain an MDP homomorphism (p, h), where $h_s = \mathrm{id}$ for all $s = (q, \xi, q^\mathrm{d}, \xi^\mathrm{d}, u^\mathrm{d}) \in \mathcal{S}$, and

$$p(s) := \big(q^{-1} q^\mathrm{d}, \xi, \xi^\mathrm{d}, u^\mathrm{d}\big). \tag{53}$$

Thus, an optimal policy and its Q function can be written

$$(\widetilde{\pi})^\uparrow(A \,|\, s) = \widetilde{\pi}\big(A \,|\, (q^{-1} q^\mathrm{d}, \xi, \xi^\mathrm{d}, u^\mathrm{d})\big), \tag{54}$$

$$Q^{(\widetilde{\pi})^\uparrow}(s, a) = \widetilde{Q}^{\widetilde{\pi}}\big((q^{-1} q^\mathrm{d}, \xi, \xi^\mathrm{d}, u^\mathrm{d}), u\big). \tag{55}$$

Hence, an optimal policy can be learned using an observation that sees only the *error* between the actual and reference poses, instead of observing these poses separately.

D.3. The `Quadrotor` System

This aerial robot is described in, *e.g.*, Mellinger and Kumar [13]. Like the `Astrobee`, it has state $x = (q, \xi)$ in $\mathcal{X} = SE(3) \times \mathbb{R}^6$, but the actions are the "single-rotor thrusts" $u \in \mathcal{U} = \mathbb{R}^4$. The dynamics are given by

$$q_{t+1} = q_t \exp(\hat{\xi}_t \,\mathrm{dt}), \tag{56a}$$

$$v_{t+1} = v_t + \big(\tfrac{1}{m} f_t - {R_t}^\mathrm{T}(\mathrm{g}\,\mathrm{e}_3)\big)\,\mathrm{dt}, \quad \omega_{t+1} = \omega_t + \mathbb{J}^{-1}(\mu_t - \omega_t \times \mathbb{J}\,\omega_t)\,\mathrm{dt}, \tag{56b}$$

where g is the magnitude of gravitational acceleration, $R_t \in SO(3)$ is the rotation component of q_t, $\mathrm{e}_3 = (0, 0, 1)$, and f_t and torque μ_t are given in terms of the actions u_t by

$$f_t = u_t^1 + u_t^2 + u_t^3 + u_t^4, \quad \mu_t = \big(\ell(u_t^1 - u_t^3), \ell(u_t^2 - u_t^4), c(u_t^1 - u_t^2 + u_t^3 - u_t^4)\big). \tag{57}$$

The running costs (51) are the same as for the `Astrobee`. Gravity "breaks" the $SE(3)$ symmetry of the dynamics, but preserves the $SE(3)$ subgroup

$$\mathcal{K}' = \left\{ \begin{pmatrix} \mathrm{rot}_z(\theta) & r \\ 0 & 1 \end{pmatrix} : (r, \theta) \in \mathbb{R}^3 \times \mathbb{S}^1 \right\} \tag{58}$$

which is isomorphic (as a Lie group) to $SE(2) \times \mathbb{R}$ and acts on $SE(3) \times \mathbb{R}^6$ via the restriction of (52). Using Theorems 8 and 5, we may derive an MDP homomorphism (p, h) for which $h_s = \mathrm{id}$ for all $s = (q, \xi, q^\mathrm{d}, \xi^\mathrm{d}, u^\mathrm{d}) \in \mathcal{S}$ and

$$p(s) := (q^{-1} q^\mathrm{d}, R^\mathrm{T} \mathrm{e}_3, \xi, \xi^\mathrm{d}, u^\mathrm{d}), \tag{59}$$

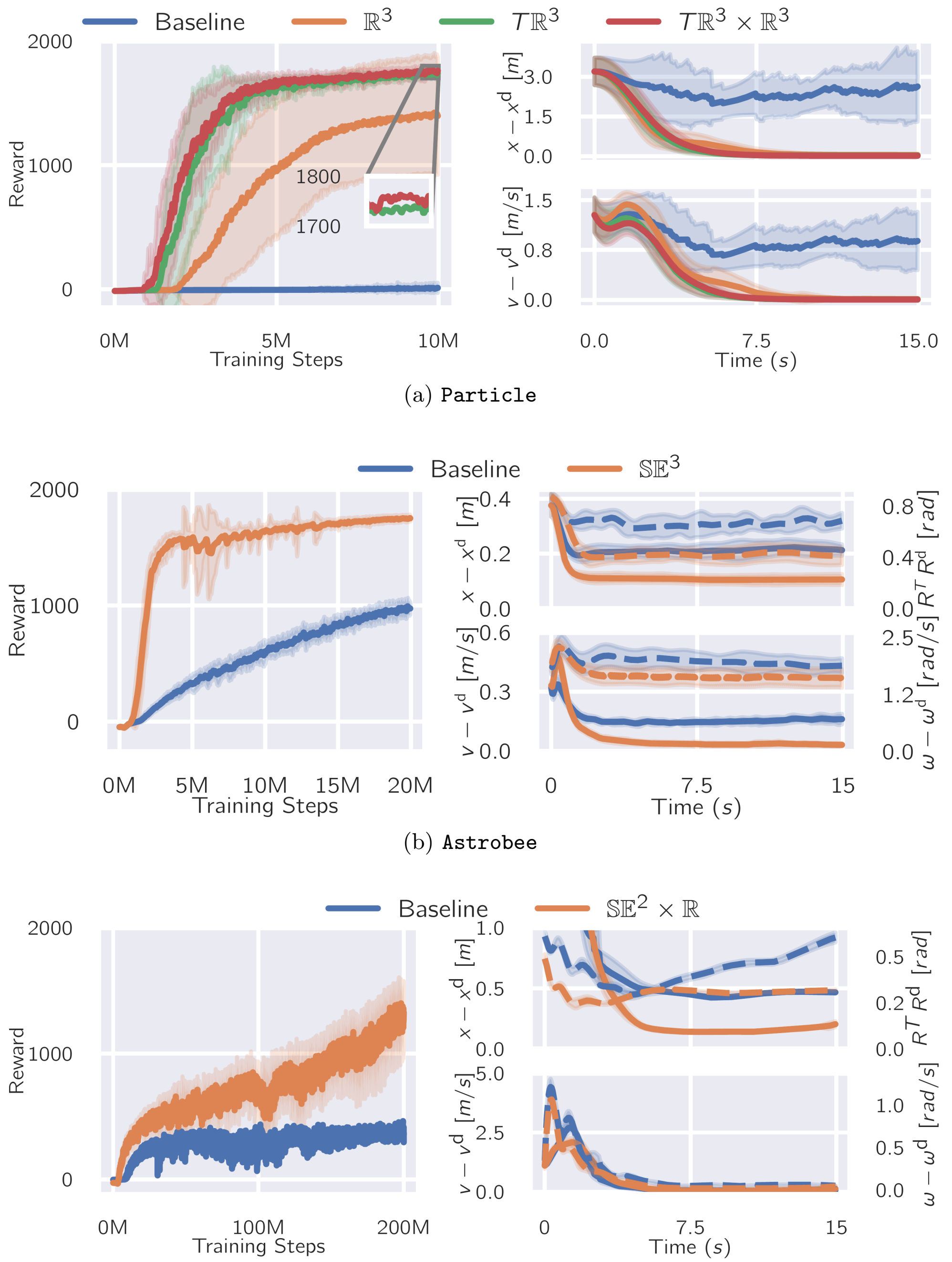
noting that $R^\mathrm{T}\mathrm{e}_3$ is the gravity direction in body coordinates. Thus, an optimal policy (and its Q function) can be written

$$(\widetilde{\pi})^\uparrow(A \,|\, s) = \widetilde{\pi}\big(A \,|\, (q^{-1} q^\mathrm{d}, R^\mathrm{T}\mathrm{e}_3, \xi, \xi^\mathrm{d}, u^\mathrm{d})\big), \tag{60}$$

$$Q^{(\widetilde{\pi})^\uparrow}(s, a) = \widetilde{Q}^{\widetilde{\pi}}\big((q^{-1} q^\mathrm{d}, R^\mathrm{T}\mathrm{e}_3, \xi, \xi^\mathrm{d}, u^\mathrm{d}), u\big). \tag{61}$$

Hence, our theory demonstrates that for quadrotors, the state space of the tracking problem can be reduced by replacing the reference and actual poses with the pose error and the body-frame gravity vector, without degrading the best-case learned policy. Consider how this differs from heuristic approximations in prior work such as [14], whose state included the entire orientation R (incompletely reducing the symmetry) and replaced the actual and reference angular velocities with the velocity error, which corresponds to an *approximate* symmetry due to the “cross terms” in the second half of (50b).

Appendix E. Complete Results of Numerical Experiments



(a) `Particle`

(b) `Astrobee`

(c) `Quadrotor`

Figure 2: Reward during training and tracking error components during evaluation for the `Particle`, `Astrobee`, and `Quadrotor`, with translational errors as solid lines and rotational errors (when applicable) as dashed lines.