SAMPLE COMPLEXITY OF EQUIVARIANT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) is a powerful framework for robotic control, yet its practical application is often hindered by high sample complexity. This is particularly restrictive in physical domains where interaction data is costly. While the world often exhibits geometric and physical symmetries, standard RL algorithms typically fail to exploit this structure. In this paper, we demonstrate that exploiting group symmetries significantly reduces the sample complexity of RL. Focusing on finite-horizon Markov decision processes, we find that leveraging homomorphisms induced by group symmetries significantly reduces the theoretical upper and lower bounds on the number of environment interactions required to reach an optimal return. We further extend these bounds to continuous state and action spaces, providing corresponding sample-complexity guarantees under appropriate regularity assumptions. Beyond theory, we validate our findings through controlled experiments and demonstrate the advantages of symmetry-aware policy learning on high-dimensional continuous robotic simulations. Our results show that integrating symmetry into the learning pipeline yields substantial gains in sample efficiency and performance, offering a principled path toward more data-efficient robotics.

1 INTRODUCTION

Reinforcement learning (RL) is a prominent framework for learning decision-making policies directly from interaction, with notable successes in robotics Levine et al. (2016); Kalashnikov et al. (2018); Mazouz et al. (2022). Recent advances in learning-enabled systems have shifted the paradigm from carefully engineered control and perception pipelines toward data-driven approaches that learn from high-dimensional sensory inputs. While these methods have demonstrated strong empirical performance, their deployment in real-world robotic systems remains constrained by poor sample efficiency and limited generalization Zhang et al. (2025); Ngo & Ravanbakhsh (2025). One source of this inefficiency is that many environments possess geometric structure that is not explicitly used by standard learning architectures Van der Pol et al. (2020). When symmetry-related configurations are represented as distinct inputs, an agent may need to relearn the same behavior across redundant representations of the environment, increasing the amount of data required for learning. This motivates the use of equivariance as a principled mechanism for exploiting latent structure in RL.

In parallel, a growing body of work in geometric deep learning has shown that encoding symmetry through equivariant architectures can improve generalization and sample efficiency Bronstein et al. (2017, 2021); Cohen & Welling (2016); Monti et al. (2017); Papillon et al. (2025); Shewmake et al. (2023); Vadgama et al. (2025); Islam et al. (2025). These advances have primarily focused on supervised and representation learning, where the role of symmetry can be studied without the additional complications of exploration, sequential dependence, and policy-induced distribution shift Zhang et al. (2016); Kumar et al. (2020). In RL, symmetry-aware methods and Markov Decision Process (MDP) homomorphisms have also been explored Van der Pol et al. (2020); Li et al. (2006); Ravindran (2003), but existing work does not quantify how symmetry-induced reductions translate into sample-complexity gains in discrete and continuous domains. In this work, we address this gap by studying the role of symmetry as an inductive bias. Our central premise is that symmetries of the MDP induce equivalence classes over states and actions. When these equivalences are exploited, the agent can share information across equivalent problems rather than learning over redundant indexed representations, reducing the effective size of the learning problem.

Fig. 1 illustrates this reduction through an orbit-collapse view. A homomorphism h maps the original MDP M to a reduced MDP $\bar{M}$ by collapsing symmetry-induced equivalence classes into

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

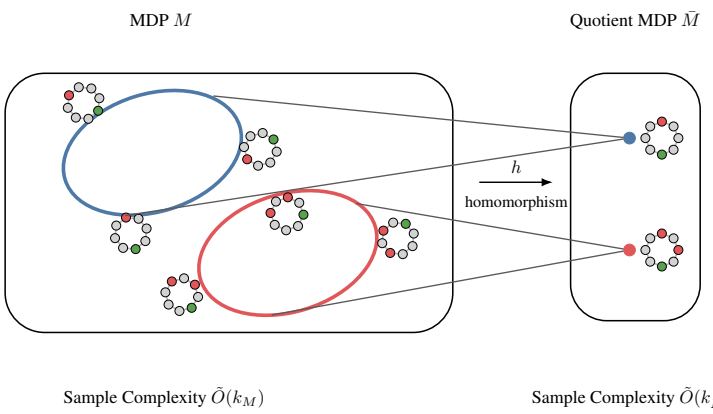


Figure 1: The homomorphism h maps equivalent states in M to representative states in the quotient MDP $\bar{M}$, reducing the effective sample complexity from $\tilde{\mathcal{O}}(k_M)$ to $\tilde{\mathcal{O}}(k_{\bar{M}})$ with $k_{\bar{M}} \ll k_M$.

representative states and actions. In the figure, each colored orbit in M contains several rotated versions of the same underlying ring configuration, while the corresponding state in $\bar{M}$ represents the entire orbit. The optimal policy found in the reduced MDP corresponds to an optimal policy in the original MDP Van der Pol et al. (2020). Thus, the agent may learn over the smaller quotient problem while still obtaining a policy that is optimal for the original space. This reduction is directly reflected in the sample complexity: the number of episodes required to reach the same performance level decreases from $\tilde{\mathcal{O}}(k_M)$ in the original MDP to $\tilde{\mathcal{O}}(k_{\bar{M}})$ in the quotient MDP.

In this work, to formally reason about sample complexity, we formalize MDPs that admit group symmetries via a *graph-theoretic* framework. This perspective exposes the symmetry-induced equivalence classes and provides the basis for constructing the corresponding quotient MDP (Fig. 1). We show that learning on the quotient MDP requires fewer episodes to obtain a near-optimal policy for the original MDP than learning directly over the unreduced state-action space. Building on the PAC framework for episodic RL Dann & Brunskill (2015), we derive *upper* and *lower* bounds that make the dependence of sample complexity on symmetry explicit. We further extend this analysis to *continuous state-action* spaces, deriving sample-complexity guarantees in terms of symmetry-reduced covering numbers under suitable regularity assumptions. These bounds show that, under suitable structural conditions, symmetry changes the effective scaling of the learning problem, yielding improvements that grow with the number of redundant representations identified by symmetry. Beyond theory, we demonstrate the practical implications through designed graph-structured experiments and broader RL benchmarks. By comparing standard and equivariant learning algorithms under identical settings, we show that exploiting latent structure improves learning efficiency and performance.

Contributions. In summary, the main contributions of this paper are as follows:

- A formal framework for analyzing the sample complexity of reinforcement learning under symmetry through the lens of graph-based MDPs.
- Sample-complexity guarantees for continuous state-action spaces in terms of symmetry-reduced covering numbers.
- Upper and lower sample-complexity bounds that make the role of symmetry explicit, showing that exploiting symmetry changes the effective scaling of the learning problem.
- Empirical results and benchmarks demonstrating that these theoretical predictions translate into consistent improvements in learning efficiency and performance.

Related Work. *Sample complexity* in RL has been extensively studied in multi-armed bandits, infinite-horizon MDPs, and episodic fixed-horizon settings Lai & Robbins (1985); Auer et al. (2002); Kakade (2003); Strehl et al. (2006). Early PAC analyses established finite-sample guarantees for exploration in MDPs, often with strong dependencies on the horizon, state space, and action space, with subsequent work improving these dependencies through refined optimistic and upper confidence bound methods Auer & Ortner (2008); Jaksch et al. (2010); Fiechter (1994). Sample complexity

has also been studied in continuous state-action spaces through regularity assumptions and covering arguments Maran et al. (2024), while recent work has considered scaling laws relating data and compute Rybkin et al. (2025). However, these analyses do not quantify how group symmetries reduce the effective complexity of the learning problem. A complementary line of work studies structural abstractions, including MDP homomorphisms Li et al. (2006); Ravindran (2003) and symmetry-aware representations such as homomorphic networks and group-equivariant RL Van der Pol et al. (2020); Finzi et al. (2021). Similarly, graph-based representations have been widely used to model structured decision-making problems Battaglia et al. (2018); Jiang et al. (2018); Albrecht et al. (2024). While these approaches exploit symmetry and structure algorithmically, they do not characterize the resulting reduction in sample complexity. This paper closes this gap by deriving symmetry-dependent sample-complexity guarantees for discrete and continuous state-action spaces.

2 Problem Formulation

In this work, we are interested in the analysis of the sample complexity of RL algorithms when the underlying environment contains symmetry-induced structure. We begin by reviewing the standard finite-horizon Markov Decision Process (MDP) setting, in accordance with the following definition.

Definition 1 (Markov Decision Process). *A Markov Decision Process (MDP) is a tuple $M = (\mathcal{S}, \mathcal{A}, p, p_0, r, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces. Here, $\mathcal{A}$ denotes the ambient action set; in settings with state-dependent admissibility, the transition and reward functions are evaluated only on admissible pairs (s, a) with $a \in \mathcal{A}(s)$. The function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a stationary transition kernel satisfying $\sum_{s' \in \mathcal{S}} p(s' \mid s, a) = 1$ for all admissible pairs (s, a). The initial state distribution is $p_0 : \mathcal{S} \to [0, 1]$ with $\sum_{s \in \mathcal{S}} p_0(s) = 1$. The reward function r is possibly time-dependent and takes values in $\mathbb{R}$. The horizon $H \in \mathbb{N}_{\geq 1}$ denotes the finite decision horizon.*

The agent interacts with this MDP over episodes of fixed horizon H. At time $t \in \{1, \dots, H\}$, the agent observes a state $s_t \in \mathcal{S}$ and selects an action $a_t = \pi_t(s_t)$ according to a (possibly time-dependent) policy $\pi = \{\pi_t\}_{t=1}^{H}$, where $\pi_t : \mathcal{S} \to \mathcal{A}$. The next state is sampled as $s_{t+1} \sim p(\cdot \mid s_t, a_t)$, and the initial state is drawn as $s_1 \sim p_0$.

At each time t, the agent receives a reward drawn from a distribution with mean $r_t(s_t, a_t)$, determined by the reward function. It is assumed that the reward function r is known to the agent, whereas the transition kernel p is unknown. The performance of a policy π on M is defined as the *total expected reward* of an episode $R_M^\pi = \mathbb{E}\left[\sum_{t=1}^{H} r_t(s_t, a_t)\right]$. For a policy π and transition kernel p, the sequence $\{s_t\}_{t=1}^{H}$ is a Markov process with a well-defined probability measure Bertsekas & Shreve (1996). The expectation in R_M^π is taken with respect to this measure. For this setup, sample complexity can be studied through *probably approximately correct* (PAC) bounds.

Definition 2 (Sample Complexity). *Let M be an MDP as in Definition 1, and let $\pi^\star$ denote an optimal policy with reward value $R_M^{\pi^\star}$. For accuracy $\epsilon > 0$ and confidence $\delta \in (0, 1)$, the* sample complexity *of a learning algorithm is the number of episodes in which the policy executed by the algorithm is ϵ-suboptimal, i.e., $R_M^{\pi_t} < R_M^{\pi^\star} - \epsilon$, with probability at least $1 - \delta$.*

The goal of this work is to understand how this sample complexity changes when the MDP contains symmetry-induced redundancy in its representation. In what follows, we formalize the problem.

Problem 1. *Let M be a finite-horizon MDP. For given $\epsilon > 0$ and $\delta \in (0, 1)$, we seek to quantify how exploiting symmetry-induced redundancy changes the sample complexity.*

Approach Our approach to Problem 1 is to make explicit the transition structure underlying the MDP. This perspective views the MDP as an action-labeled transition graph: states form the nodes, and actions label the possible transitions between states. We then study settings in which this transition graph admits group symmetry. These symmetries identify redundant state-action representations, reducing the effective size of the learning problem. This leads to upper and lower sample-complexity bounds that depend explicitly on the available group symmetry, and provides the principled basis for comparing structured learning with learning over the unreduced representation.

3 PRELIMINARIES

In this section, we show that for symmetric environments, standard MDP formulations suffer from an explosion of the state space by naively treating structurally equivalent configurations as distinct entities. To formalize this redundancy, we use graph representations of the environment, which naturally expose its connectivity and symmetries. We then relate these symmetries to MDP homomorphisms.

3.1 GRAPH-BASED MDP FORMULATION

While a discrete MDP can be represented by its transition graph, we formulate the MDP directly over a directed graph $\mathbb{G} = (V, E)$ representing the environment's connectivity. In this setting, an MDP state is a configuration of features across the vertices of $\mathbb{G}$, together with the agent's spatial location.

Definition 3 (Graph-Valued State Space)**.** *Let X be a finite alphabet of node features, represented as one-hot vectors. The graph-valued state space is $\mathcal{S} := X^V \times V$. Each state $s \in \mathcal{S}$ is written as $s = (x, q)$, where $x : V \to X$ assigns a value $x(v) \in X$ to each node $v \in V$, and $q \in V$ denotes the current position of the agent.*

Intuitively, picture $\mathbb{G}$ as a game board (e.g., a Sokoban room or a chessboard): x describes the current arrangement of the pieces (the state of the board), q marks where the agent is standing. The topology $\mathbb{G}$ is fixed across all states; the labelling x and position q are what change from one state to the next. With the graph-valued state space specified, we now define the admissible action set for each state.

Definition 4 (Graph Action Space)**.** *For any state $s = (x, q)$, an action corresponds to selecting a neighboring node reachable from the agent's node q. Thus, the admissible action set is $\mathcal{A}(s) := \{v \in V : (q, v) \in E\}$. Equivalently, each action may be identified with an admissible edge $(q, v) \in E$.*

We now specify the stochastic dynamics and objective of the graph-based MDP.

Definition 5 (Graph Transition Kernel)**.** *For any state $s = (x, q) \in \mathcal{S}$ and action $a \in \mathcal{A}(s)$, the kernel $p(s' \mid s, a)$ denotes the probability of transitioning to a successor state $s' = (x', q') \in \mathcal{S}$. This kernel characterizes the joint evolution of the graph valuation x and the agent location q.*

There is a directed edge $s \xrightarrow{a} s'$ in the state-transition graph whenever $p(s' \mid s, a) > 0$. Similarly, we define the reward structure associated with the graph-based MDP.

Definition 6 (Graph-Based Reward Function)**.** *The graph-based reward function is a possibly time-dependent mean reward map $r_t : \{(s, a) : s \in \mathcal{S},\ a \in \mathcal{A}(s)\} \to \mathbb{R}$. For a state $s = (x, q) \in \mathcal{S}$ and an action $a \in \mathcal{A}(s)$, the scalar $r_t(s, a)$ denotes the expected reward obtained at time t.*

Definitions 3–6 define the full graph-based MDP tuple. With a slight abuse of notation, we use the same representation $M = (\mathcal{S}, \mathcal{A}, p, p_0, r, H)$ for this graph-based MDP, since the underlying MDP components remain unchanged, i.e., the representation is now interpreted through the graph structure. Next, we focus on group symmetries as a special type of MDP symmetry.

Definition 7 (Group Symmetry of a Graph-Based MDP)**.** *Let M be a graph-based MDP whose state and action spaces are induced by graph $\mathbb{G} = (V, E)$. A group G is a symmetry group of M if each $g \in G$ acts on V in a way that preserves the graph connectivity and induces maps on states and actions such that, for every state $s = (x, q)$ and admissible action $a \in \mathcal{A}(s)$, $g \cdot s = (g \cdot x, g \cdot q) \in \mathcal{S}$ and $g \cdot a \in \mathcal{A}(g \cdot s)$. Moreover, the action preserves the MDP rewards and transition probabilities: for all successor states $c \in \mathcal{S}$, $r_t(g \cdot s, g \cdot a) = r_t(s, a)$ and $p(g \cdot c \mid g \cdot s, g \cdot a) = p(c \mid s, a)$.*

This is the standard notion of a group-invariant MDP Wang et al. (2022), specialized here to graph-based MDPs. For further background, we refer the reader to Appendix A. We finally define the orbit and stabilizer associated with a group action, which we will use to construct the quotient MDP.

Definition 8 (Orbit and Stabilizer)**.** *Let a finite group G act on a set Ω. For $\omega \in \Omega$, the* orbit *and* stabilizer *of ω are $\mathcal{G}_\omega := \{g \cdot \omega : g \in G\}$ and $G_\omega := \{g \in G : g \cdot \omega = \omega\} \leq G$, respectively. The orbit–stabilizer theorem (Dummit & Foote, 2004, Thm. 5.1) gives $|\mathcal{G}_\omega| = |G|/|G_\omega|$.*

Proposition 1 (Quotient MDP)**.** *Let a finite group G act on the state space $\mathcal{S}$ and action space $\mathcal{A}$ of M. For $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $\mathrm{Succ}(s, a) = \{s' \in \mathcal{S} : p(s' \mid s, a) > 0\}$ denote the set of possible*

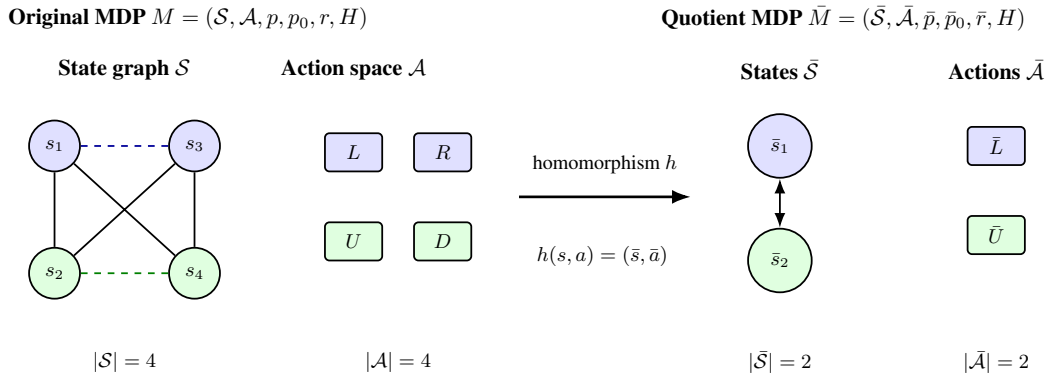


Figure 2: Example where homomorphism h maps the original MDP M to the quotient MDP $\bar{M}$.

successor states, and define $C = \max_{s,a} |\mathrm{Succ}(s, a)| \leq |\mathcal{S}|$. *The quotient sizes are determined by the corresponding orbit sizes:* $|\bar{\mathcal{S}}| = |\mathcal{S}|/|\mathcal{G}_{\mathcal{S}}|$, $|\bar{\mathcal{A}}| = |\mathcal{A}|/|\mathcal{G}_{\mathcal{A}}|$, *and* $\bar{C} = C/|\mathcal{G}_C|$, *where* $\mathcal{G}_{\mathcal{S}}$, $\mathcal{G}_{\mathcal{A}}$, *and* $\mathcal{G}_C$ *denote the orbits of representative states, actions, and successor states, respectively.*

Proof. By Definition 8, quotienting identifies elements in the same orbit. Thus, the numbers of distinct state, action, and successor representatives are reduced by the corresponding orbit sizes. □

3.2 MDP Homomorphisms

The graph-based formulation above makes explicit the structure underlying the MDP. In particular, the topology of $\mathbb{G} = (V, E)$ determines which actions are admissible from each state, while the graph valuation $x : V \to X$ determines the configuration of local features across the environment. This structure allows us to compare states not only as distinct elements of $\mathcal{S}$, but also according to whether they represent the same configuration up to a symmetry of the underlying graph.

For symmetric environments, this distinction is crucial. Two states $s = (x, q)$ and $s' = (x', q')$ may be different elements of $\mathcal{S} = X^V \times V$, while still being structurally equivalent under a symmetry of $\mathbb{G}$. A standard tabular MDP treats such states as unrelated, even when they induce identical rewards and transition behavior after relabeling the graph. This redundancy is best captured by MDP homomorphisms, which are structure-preserving maps from an original MDP to an abstract MDP Van der Pol et al. (2020), as conceptually visualized in Fig. 2.

Definition 9 (MDP Homomorphism)**.** *Let* $M = (\mathcal{S}, \mathcal{A}, p, p_0, r, H)$ *be an MDP as in Definition 1, and let* $\bar{M} = (\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{p}, \bar{p}_0, \bar{r}, H)$ *be an abstract MDP. An MDP homomorphism from* M *to* $\bar{M}$ *is a tuple* $h = (\sigma, \{\alpha_s\}_{s \in \mathcal{S}})$, *where* $\sigma : \mathcal{S} \to \bar{\mathcal{S}}$ *is a surjective state map and, for each* $s \in \mathcal{S}$, $\alpha_s : \mathcal{A}(s) \to \bar{\mathcal{A}}(\sigma(s))$ *is a surjective action map. These maps satisfy, for all* $s, s' \in \mathcal{S}$ *and* $a \in \mathcal{A}(s)$,

$$\bar{r}\big(\sigma(s), \alpha_s(a)\big) = r(s, a), \qquad \bar{p}\big(\sigma(s') \mid \sigma(s), \alpha_s(a)\big) = \sum_{s'' \in \sigma^{-1}(\sigma(s'))} p(s'' \mid s, a).$$

An exact MDP homomorphism induces a model-equivalent abstraction Li et al. (2006): state-action pairs that are mapped to the same abstract state-action pair have identical reward and aggregated transition behavior. Given a homomorphism $h = (\sigma, \{\alpha_s\}_{s \in \mathcal{S}})$, two state-action pairs (s, a) and (s', a') are h-equivalent if $\sigma(s) = \sigma(s')$ and $\alpha_s(a) = \alpha_{s'}(a')$. In this work, we use the group-symmetry setting introduced in Definition 7 as the relevant special case: the equivalence classes are the group orbits, and the resulting abstract model is the quotient MDP characterized in Proposition 1.

4 Equivariant Sample Complexity

We consider the episodic fixed-horizon graph-based quotient MDP $\bar{M}$. The agent interacts with the MDP over episodes of length H, and we study the number of episodes in which the executed policy is ϵ-suboptimal with probability at least $1 - \delta$. We quantify the reduction in sample complexity bounds relative to the original MDP M. For details on the episodic PAC framework, we refer to Appendix B.

4.1 Equivariant Sample Complexity Bounds

We are now ready to incorporate the group symmetry into the upper and lower bounds of the sample-complexity. We note that $\tilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors in H, $|\mathcal{S}|$, $|\mathcal{A}|$, $|G|$, $1/\epsilon$, and $1/\delta$. This corresponds to a Uniform-PAC guarantee, i.e., the bound is polynomial in all of its arguments Dann et al. (2017). Our strategy is to apply the standard PAC bound to the quotient MDP. By the homomorphism property, an optimal policy in $\bar{M}$ lifts to an optimal policy in M.

Lemma 1. *Let $M = (\mathcal{S}, \mathcal{A}, p, p_0, r, H)$ be an MDP and let h be the mapping that induces an MDP homomorphism from M to $\bar{M}$ as in Definition 9, where $\bar{M} = (\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{p}, \bar{p}_0, \bar{r}, H)$. Applying the result of (Dann & Brunskill, 2015, Thm. 1) to $\bar{M}$ yields the following sample-complexity upper bound*

$$\tilde{\mathcal{O}}\left(\frac{H^2 \bar{C} |\bar{\mathcal{S}} \times \bar{\mathcal{A}}|}{\epsilon^2} \ln \frac{1}{\delta}\right).$$

Proof. By Definition 9, the mapping h induces the quotient MDP $\bar{M}$, whose state and action spaces are $\bar{\mathcal{S}}$ and $\bar{\mathcal{A}}$ and whose successor-state parameter is $\bar{C}$. Substituting these quantities into the bound of Dann & Brunskill (2015), namely replacing $\mathcal{S}, \mathcal{A}, C$ by $\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{C}$, directly yields the result. □

Lemma 1 applies the standard PAC sample-complexity bound directly to the quotient MDP. Theorem 1 uses this substitution to quantify the resulting reduction relative to the original MDP.

Theorem 1 (Upper Bound)**.** *Let $M = (\mathcal{S}, \mathcal{A}, p, p_0, r, H)$ be an episodic fixed-horizon MDP and let $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, the number of episodes satisfying $R^* - R^{\pi^k} = V^*_{1:H}(s_0) - V^{\pi^k}_{1:H}(s_0) > \epsilon$ is reduced by a factor of $|\mathcal{G}_\mathcal{S}|\,|\mathcal{G}_\mathcal{A}|\,|\mathcal{G}_C|$ relative to the corresponding bound for the original MDP, yielding the upper bound*

$$\tilde{\mathcal{O}}\left(\frac{H^2 C |\mathcal{S} \times \mathcal{A}|}{|\mathcal{G}_\mathcal{S}|\,|\mathcal{G}_\mathcal{A}|\,|\mathcal{G}_C|\,\epsilon^2} \ln \frac{1}{\delta}\right).$$

Proof. By Lemma 1, the sample complexity is bounded in terms of $\bar{\mathcal{S}}$, $\bar{\mathcal{A}}$, and $\bar{C}$. Per Proposition 1, the equivalence classes on the quotient spaces satisfy $|\bar{\mathcal{S}}| = |\mathcal{S}|/|\mathcal{G}_\mathcal{S}|$ and $|\bar{\mathcal{A}}| = |\mathcal{A}|/|\mathcal{G}_\mathcal{A}|$. Similarly, the successor-state complexity satisfies $\bar{C} = C/|\mathcal{G}_C|$. Substituting these quantities yields the complexity reduction $|\mathcal{G}_\mathcal{S}|\,|\mathcal{G}_\mathcal{A}|\,|\mathcal{G}_C|$. □

The utility of Theorem 1 is best illustrated by the structural compression observed in the quotient MDP (Fig. 2). While the original problem is governed by a complexity factor of $C|S \times A| = 64$, the symmetry-aware mapping reduces this to $\bar{C}|\bar{S} \times \bar{A}| = 8$. This 2^3 reduction is not merely a numerical artifact of this example, but a direct manifestation of the abstract bounds established in Theorem 1. The following corollary makes this reduction explicit for a cyclic ring (Fig. 1).

Corollary 1 (Ring Sample Complexity)**.** *For a ring domain with $|G|$ cyclic rotations, the symmetry group G acts freely on states, while actions and the local successor branching factor are preserved. Hence, Theorem 1 yields $\tilde{\mathcal{O}}\left(\frac{H^2 C |\mathcal{S} \times \mathcal{A}|}{|G| \epsilon^2} \ln \frac{1}{\delta}\right)$.*

Proof. Quotienting reduces the number of states by $|G|$, while the action space $\mathcal{A}$ and the successor-state complexity C is preserved. □

We next establish a matching lower bound. Here, $\tilde{\Omega}(\cdot)$ hides logarithmic factors, analogously to $\tilde{\mathcal{O}}(\cdot)$.

Theorem 2 (Lower Bound)**.** *Let $M = (\mathcal{S}, \mathcal{A}, p, p_0, r, H)$ be an episodic fixed-horizon MDP and let $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, the number of episodes satisfying $R^* - R^{\pi^k} = V^*_{1:H}(s_0) - V^{\pi^k}_{1:H}(s_0) > \epsilon$ has a lower bound reduced by a factor of $|\mathcal{G}_\mathcal{S}|\,|\mathcal{G}_\mathcal{A}|$ relative to the corresponding bound for the original MDP, yielding the lower bound*

$$\tilde{\Omega}\left(\frac{H^2 |\mathcal{S} \times \mathcal{A}|}{|\mathcal{G}_\mathcal{S}||\mathcal{G}_\mathcal{A}|\epsilon^2} \ln \frac{1}{\tilde{\delta}}\right),$$

where $\tilde{\delta} = \delta + c$ for an absolute constant $c > 0$.

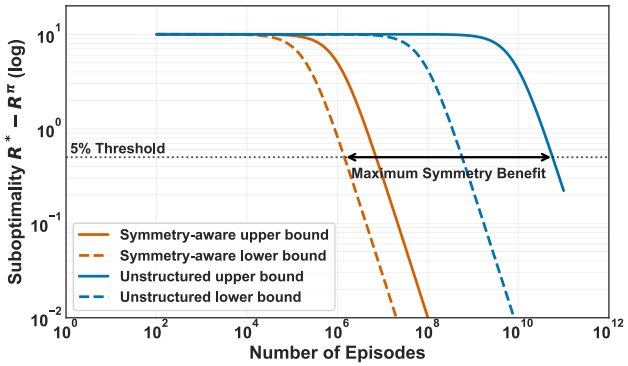


Figure 3: Theoretical comparison between unstructured and symmetry-aware sample complexity for a graph-based MDP with $|\mathcal{S}| = 10^5$ states before quotienting by symmetry. For the ring graph, the symmetry group has full state-size symmetry, $|G| = 10^5$, yielding a potential five-order-of-magnitude reduction in the number of episodes required to reach 95% optimality with high probability.

Proof. By the MDP homomorphism, the reduced MDP has state space $\bar{\mathcal{S}}$ and action space $\bar{\mathcal{A}}$. The equivalence classes on the quotient spaces satisfy $|\bar{\mathcal{S}}| = |\mathcal{S}|/|\mathcal{G}_\mathcal{S}|$ and $|\bar{\mathcal{A}}| = |\mathcal{A}|/|\mathcal{G}_\mathcal{A}|$. Substituting yields the complexity reduction $|\mathcal{G}_\mathcal{S}||\mathcal{G}_\mathcal{A}|$. □

The theoretical gap illustrated in Fig. 3 underscores the improvement in efficiency afforded by exploiting problem symmetries. For this graph-based MDP, where $\mathcal{S}$ denotes the original state space before quotienting and $|\mathcal{S}| = 10^5$, the symmetry-aware bounds reflect a substantial (potentially five orders of magnitude) reduction in the number of episodes required to achieve convergence to 95% optimality. This performance gap becomes more pronounced as the problem scale increases.

4.2 Continuous-Space Bounds

We now extend the sample-complexity analysis to continuous domains by replacing finite cardinalities with covering numbers. In particular, under suitable regularity assumptions, the continuous MDP can be discretized by an equivariant α-net that preserves the symmetry structure. The resulting quotient MDP then inherits the sample-complexity reduction of the discrete analysis, with the cardinalities of the quotient state and action spaces replaced by their corresponding covering numbers.

Theorem 3 (Continuous-Space Upper Bound)**.** *Let $(\mathcal{S}, \rho_\mathcal{S})$ and $(\mathcal{A}, \rho_\mathcal{A})$ be compact metric spaces for a finite-horizon MDP with L_r-Lipschitz rewards and L_p-Lipschitz transitions in the Wasserstein metric. Assume a group G acts freely and isometrically on $\mathcal{S}$ and $\mathcal{A}$, and that the MDP is G-invariant. Then, for $\epsilon, \delta \in (0, 1)$ and an equivariant discretization with resolution $\alpha = \alpha(\epsilon) \propto \epsilon/(HL_V)$, with probability at least $1 - \delta$, the number of ϵ-suboptimal episodes is upper bounded by*

$$\tilde{\mathcal{O}}\left(\frac{H^2 \bar{C}_\alpha N(\mathcal{S}/G, \alpha) N(\mathcal{A}/G, \alpha)}{\epsilon^2} \ln \frac{1}{\delta}\right),$$

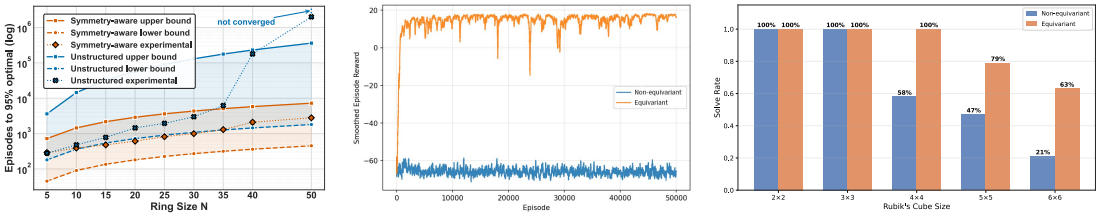
where $\bar{C}_\alpha$ is the successor-state complexity of the discretized quotient MDP and L_V is determined by the reward and transition Lipschitz constants.

Proof. Let $\alpha > 0$ and construct equivariant α-nets of $\mathcal{S}$ and $\mathcal{A}$. Since the group action is isometric, the discretization preserves the G-symmetry and induces a quotient discretized MDP. Under the Lipschitz assumptions, the value function is Lipschitz with constant L_V. Choosing

$$\alpha(\epsilon) = \frac{\epsilon}{2c_0 H L_V},$$

for an absolute constant c_0, bounds the discretization error by $\epsilon/2$. The resulting quotient MDP has state and action cardinalities bounded by $N(\mathcal{S}/G, \alpha(\epsilon))$ and $N(\mathcal{A}/G, \alpha(\epsilon))$, respectively. Applying Theorem 1 to this quotient MDP gives

$$\tilde{\mathcal{O}}\left(\frac{H^2 \bar{C}_{\alpha(\epsilon)} N(\mathcal{S}/G, \alpha(\epsilon)) N(\mathcal{A}/G, \alpha(\epsilon))}{\epsilon^2} \ln \frac{1}{\delta}\right).$$



(a) Cyclic ring experiment (1D). (b) Cyclic ring experiment (3D). (c) Rubik's cube experiment.

Figure 4: Performance across the 1D cyclic ring, 3D cyclic ring, and Rubik's cube experiments. Symmetry-aware learning reduces the number of episodes needed for better performance.

Finally, lifting the resulting policy to the original continuous MDP through the homomorphism introduces at most $\epsilon/2$ discretization error, yielding an ϵ-optimal policy and the stated bound. □

In the next section, we empirically evaluate the theoretical results and their implications across discrete and continuous domains.

5 Experiments

We evaluate the practical implications of the sample complexity bounds on a set of representative environments with symmetry. Across all experiments, the baseline and symmetry-aware variants are compared under identical training settings, including network architecture, optimizer, and hyperparameter choices. We consider five tasks of increasing structural complexity: 1) a cyclic environment illustrating basic group symmetry, 2) a toroidal environment with three coupled rings illustrating product cyclic structure, 3) a structured combinatorial domain based on a Rubik's game, 4) a block-stacking robotics task and 5) a nut-assembly robotics task, both from Mandlekar et al. (2023). Additional setup and implementation details are provided in Appendix C.

5.1 Results

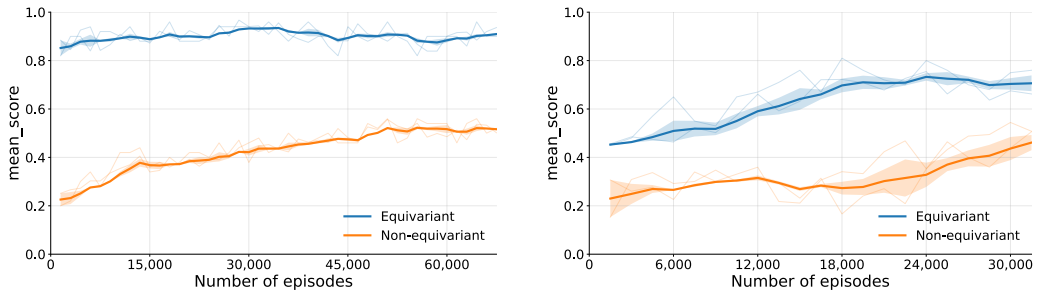
Structured Domains. We first consider ring-structured graphs with N nodes. In the 1D ring, the symmetry group is the cyclic group $G = \mathbb{Z}/N\mathbb{Z}$ of rotations, so $|G| = N$. In the 3D ring, the graph is the product of three cyclic rings with symmetry group $G = (\mathbb{Z}/N\mathbb{Z})^3$, so $|G| = N^3$. Figs. 4a and 4b show the theoretical bounds and empirical learning curves, respectively. The equivariant method reaches higher reward within fewer episodes and exhibits substantially better sample efficiency than the non-equivariant baseline. This effect is more pronounced in 3D, where the non-equivariant baseline incurs high loss and fails to converge within the training budget. For the unstructured baseline at $N = 50$ (Fig. 4a), the empirical sample complexity exceeds the plotted theoretical range and the method fails to reach the required performance threshold within the training budget. A similar trend appears in the Rubik's cube experiments in Fig. 4c. Canonicalization maps symmetry-equivalent cube states to a common representative, reducing the number of configurations the learner must distinguish. The symmetry-aware method maintains higher solve rates as the problem size increases, whereas the non-equivariant baseline degrades more rapidly, demonstrating improved information sharing across equivalent configurations.

Table 1: Convergence for the 1D cyclic ring over 10 independent random seeds (mean ± SE).

N	Non-equiv.	Equiv.
5	$\mathbf{280.5 \pm 20.9}$	290.1 ± 13.0
10	484.5 ± 22.9	$\mathbf{405.1 \pm 18.0}$
15	788.6 ± 33.0	$\mathbf{493.1 \pm 15.6}$
20	$1{,}444.1 \pm 137.0$	$\mathbf{617.1 \pm 28.8}$
25	$1{,}966.1 \pm 129.4$	$\mathbf{829.1 \pm 63.9}$
30	$3{,}016.3 \pm 158.0$	$\mathbf{997.0 \pm 71.0}$
35	$6{,}206.6 \pm 635.4$	$\mathbf{1{,}350.9 \pm 144.2}$
40	$177{,}334.3 \pm 21{,}201.9$	$\mathbf{2{,}157.4 \pm 208.2}$

Table 2: Rubik's cube success rates over 10 independent random seeds (mean ± SE).

Cube size	Non-equiv.	Canonical	O-equiv.
2×2	$\mathbf{100.0 \pm 0.0\%}$	$\mathbf{100.0 \pm 0.0\%}$	$\mathbf{100.0 \pm 0.0\%}$
3×3	$98.7 \pm 1.5\%$	$\mathbf{99.3 \pm 1.2\%}$	$99.0 \pm 1.7\%$
4×4	$61.7 \pm 4.0\%$	$\mathbf{97.0 \pm 3.0\%}$	$95.0 \pm 2.6\%$
5×5	$42.7 \pm 4.5\%$	$\mathbf{83.7 \pm 4.5\%}$	$81.0 \pm 2.6\%$



(a) Block stacking benchmark.

(b) Nut assembly benchmark.

Figure 5: Continuous-control manipulation benchmarks comparing symmetry-aware and non-equivariant learning across multiple random seeds.

Ablation Study. For the structured domains, we further evaluate the robustness of the observed gains over 10 independent random seeds. In the 1D cyclic ring, the results in Table 1 show that the advantage of equivariant learning grows with N: while both methods perform similarly for small rings, the gap increases substantially with problem size, with the non-equivariant method requiring $177{,}334.3 \pm 21{,}201.9$ steps at $N = 40$ compared with $2{,}157.4 \pm 208.2$ for the equivariant method. The same behavior is observed in the 3D cyclic ring, where the final reward is -67.38 ± 1.12 SE for the non-equivariant method and 16.96 ± 0.32 SE for the equivariant method. Finally, Table 2 compares the non-equivariant baseline, canonicalization, and an O-equivariant MDP-homomorphic network on the Rubik's cube task. The O-equivariant method uses an MDP-homomorphic network Van der Pol et al. (2020). Both symmetry-aware approaches remain effective as the cube size increases and substantially outperform the non-equivariant baseline, while canonicalization performs comparably to or slightly better than the network-level equivariant approach. Together, these results show that the gains from exploiting symmetry persist across random seeds, become more pronounced as problem complexity grows, and are not specific to a particular realization of equivariance.

Continuous-Control Tasks. We additionally evaluate the framework on *continuous-control* robotic manipulation benchmarks from Mandlekar et al. (2023). These tasks involve high-dimensional continuous observations and actions and contact-rich dynamics. In this setting, we model the relevant symmetry group as $\mathrm{SE}(3)$, the special Euclidean group in 3D. These experiments complement Theorem 3, which extends the symmetry-induced sample-complexity reduction to continuous state-action spaces through quotient covering numbers. In the block stacking benchmark (Fig. 5a), the agent must learn manipulation behavior that generalizes across symmetry-related object configurations. The non-equivariant baseline exhibits worse learning behavior, indicating that it must relearn similar behaviors across equivalent configurations. By contrast, the equivariant agent reaches useful performance levels earlier and achieves higher return, showing that the symmetry-aware representation improves learning efficiency. In the nut assembly benchmark (Fig. 5b), the task requires precise contact-rich control and alignment, making learning more sensitive to exploration and initialization. The same trend persists: the equivariant method consistently outperforms the non-equivariant baseline across training, with higher mean return. These results are consistent with Theorem 3, demonstrating that the benefits of symmetry-aware learning extend empirically to continuous-control domains.

6 Conclusion

This work studies how symmetry-induced structure changes the sample complexity of reinforcement learning. For finite MDPs, we formalize how graph-induced symmetries partition redundant state-action representations into orbits and show how the resulting quotient changes the effective size of the learning problem. We further extend this analysis to continuous state-action spaces, where symmetry reduces the effective complexity through the covering numbers of the corresponding quotient spaces. The resulting bounds demonstrate that symmetry-aware learning can improve sample efficiency in a way that scales with the amount of redundancy present in the environment. Our results show that integrating symmetry into the learning pipeline yields substantial gains in sample efficiency and performance, offering a principled path toward more data-efficient general-purpose robotics. Future work will extend this framework toward equivariant world models for and on-robot demonstrations.

AI USE STATEMENT

Generative AI tools were used solely to assist with writing and editing the manuscript. They were not used in the development of the methodology, theoretical results, experimental design, data analysis, or scientific conclusions. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This work does not involve human subjects, private or sensitive data, or other ethical concerns requiring specific consideration.

REPRODUCIBILITY STATEMENT

Complete experimental settings, implementation details, architectures, hyperparameters, compute requirements, and environment specifications are provided in Appendix C. The theoretical assumptions and derivations are provided in the main text and appendices. The codebase is currently subject to release restrictions and will be made publicly available upon paper acceptance.

REFERENCES

Forest Agostinelli, Stephen McAleer, Alexander Shmakov, and Pierre Baldi. Solving the rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, 1(8):356–363, 2019.

Stefano V Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press, 2024.

Peter Auer and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2008.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Kang Chen, Zhihao Liu, Tonghe Zhang, Zhen Guo, Si Xu, Hao Lin, Hongzhi Zang, Xiang Li, Quanlu Zhang, Zhaofei Yu, Guoliang Fan, Tiejun Huang, Yu Wang, and Chao Yu. π_{RL}: Online RL fine-tuning for flow-based vision-language-action models. *arXiv preprint arXiv:2510.25889*, 2025.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

David S. Dummit and Richard M. Foote. *Abstract Algebra*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2004. ISBN 9780471433347.

Claude-Nicolas Fiechter. Efficient reinforcement learning in markov decision processes. *Proceedings of the Conference on Learning Theory*, 1994.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant neural networks for arbitrary matrix groups. In *International Conference on Machine Learning*, 2021.

Mohammad Mohaiminul Islam, Rishabh Anand, David R Wessels, Friso de Kruiff, Thijs P Kuipers, Rex Ying, Clara I Sánchez, Sharvaree Vadgama, Georg Bökman, and Erik J Bekkers. Platonic transformers: A solid choice for equivariance. *arXiv preprint arXiv:2510.03511*, 2025.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. *arXiv preprint arXiv:1810.09202*, 2018.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. In *Journal of Machine Learning Research*, 2016.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. Projection by convolution: Optimal sample complexity for reinforcement learning in continuous-space mdps. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 3743–3774. PMLR, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Rayan Mazouz, Karan Muvvala, Akash Ratheesh Babu, Luca Laurenti, and Morteza Lahijanian. Safety guarantees for neural network dynamic systems via stochastic barrier functions. *Advances in Neural Information Processing Systems*, 35:9672–9686, 2022.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5115–5124, 2017.

Khang Ngo and Siamak Ravanbakhsh. Scaling laws and symmetry, evidence from neural force fields. *arXiv preprint arXiv:2510.09768*, 2025.

Mathilde Papillon, Sophia Sanborn, Johan Mathe, Louisa Cornelis, Abby Bertics, Domas Buracas, Hansen J Lillemark, Christian Shewmake, Fatih Dinc, Xavier Pennec, et al. Beyond euclid: An illustrated guide to modern machine learning with geometric, topological, and algebraic structures. *Machine Learning: Science and Technology*, 6(3):031002, 2025.

Balaraman Ravindran. Smdp homomorphisms: An algebraic approach to abstraction in semi-markov decision processes. In *IJCAI*, 2003.

Oleh Rybkin, Michal Nauman, Preston Fu, Charlie Snell, Pieter Abbeel, Sergey Levine, and Aviral Kumar. Value-based deep rl scales predictably. *arXiv preprint arXiv:2502.04327*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, and et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Christian Shewmake, Nina Miolane, and Bruno Olshausen. Group equivariant sparse coding. In *International Conference on Geometric Science of Information*, pp. 91–101. Springer, 2023.

Alexander L Strehl, Lihong Li, and Michael L Littman. Pac model-free reinforcement learning. *Proceedings of the 23rd international conference on Machine learning*, 2006.

Sharvaree Vadgama, Mohammad Mohaiminul Islam, Domas Buracas, Christian Shewmake, Artem Moskalev, and Erik Bekkers. Probing equivariance and symmetry breaking in convolutional networks. *arXiv preprint arXiv:2501.01999*, 2025.

Elise Van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. Mdp homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4199–4210, 2020.

Dian Wang, Robin Walters, and Robert Platt. So(2)-equivariant reinforcement learning. *arXiv preprint arXiv:2203.04439*, 2022.

Ge Yan, Jiyue Zhu, Yuquan Deng, Shiqi Yang, Ri-Zhao Qiu, Xuxin Cheng, Marius Memmel, Ranjay Krishna, Ankit Goyal, Xiaolong Wang, and Dieter Fox. ManiFlow: A general robot manipulation policy via consistency flow training. In *Conference on Robot Learning (CoRL)*, 2025.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Fuxiang Zhang, Junyou Li, Yi-Chen Li, Zongzhang Zhang, Yang Yu, and Deheng Ye. Improving sample efficiency of reinforcement learning with background knowledge from large language models. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.