Formalizing Task-Space Complexity for Zero-Shot Generalization

Jung-Hoon Cho JHOONCHO@MIT.EDU
Massachusetts Institute of Technology

Heling Zhang HZHNG120@ILLINOIS.EDU
Siqi Du SIQIDU3@ILLINOIS.EDU
Roy Dong ROYDONG@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Cathy Wu CATHYWU@MIT.EDU
Massachusetts Institute of Technology

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Policies must operate across diverse conditions, yet a single policy is often conservative while fully adaptive schemes can be complex. We study zero-shot generalization in contextual dynamical systems and introduce a performance-centric, directional task dissimilarity—the signed divergence—that upper bounds the generalization gap from a source context to a target context. The signed divergence induces ε-tolerance sets that certify when a source policy class generalizes, and it yields a concrete notion of task-space complexity: the minimum number of source contexts needed so that every target context incurs at most ε generalization gap. Under a mild local smoothness assumption on performance, the induced tolerance sets admit certified inner/outer balls and instance-dependent volume bounds on task-space complexity. In the finite-oracle setting, source selection reduces to set cover; a greedy strategy inherits the standard $H(n)$ approximation guarantee. Using a Mass-Spring-Damper system with linear–quadratic regulator (LQR) controllers and a nonlinear CartPole system with deep reinforcement learning controllers, we show that greedy selection achieves the same ε-coverage with fewer policies than uniform or random baselines. Our approach delivers a performance-based task similarity measure and practical certificates for building generalizable control with simple policies.

Keywords: Generalization, Reinforcement Learning, Task Space Complexity, Zero-shot Transfer.

1. Introduction

Modern control systems increasingly operate across diverse conditions—robots handle objects of varying mass, autonomous vehicles drive on surfaces with different friction, and physical systems vary in stiffness, damping, and load. In such contextual dynamical systems, each task is parameterized by an observable context vector θ in a context space Θ. Traditional control theory, and its machine learning extension, reinforcement learning (RL), have long focused on designing a single feedback law that works for a class of dynamical systems. However, as system diversity grows, this one-size-fits-all paradigm is reaching its limits.

For instance, classical robust control seeks a simple context-independent controller that stabilizes all possible parameter realizations, often yielding conservative performance (Zhou and Doyle, 1998). Adaptive control continuously tunes its parameters online, but can become unnecessarily complex when environments change slowly or when online identification is difficult (Ioannou and Sun, 1996). Similarly, multi-task deep RL aims to train a single context-conditioned controller

across tasks, but such policies often underperform specialized single-task policies or require excessive capacity and data (Kang et al., 2011; Parisotto et al., 2016).

Recent empirical work, including our own, suggests a complementary alternative: rather than searching for one universal controller, train multiple "simple" context-independent controllers, each specialized for a narrow region of the context space, and switch among them as contexts change (Cho et al., 2023, 2024). This approach, akin to gain scheduling, has demonstrated strong empirical performance in several settings (Rugh and Shamma, 2000). However, to the best of our knowledge, despite these successes, there remains no theoretical framework explaining when and why such strategies succeed. To address this gap, we aim to formalize the *task-space complexity* of a class of dynamical systems according to the difficulty of achieving a uniform bound of ε-optimal performance, which we characterize in terms of the minimal number of narrowly-trained controllers. Owing to the empirical motivation, this work focuses on *zero-shot generalization* in contextual dynamical systems, although we expect that the framework will generalize to other training and generalization paradigms. In this setting, a controller trained at one context is directly applied to another without retraining or adaptation.

To make this notion operational, we require a principled performance-based measure of task dissimilarity between contexts that reflects the generalization gap rather than raw parameter distance. A common heuristic is to use Euclidean or Wasserstein distance in context space, implicitly assuming that nearby contexts are similar (Slivkins, 2011; Modi et al., 2018; Cho et al., 2024; Dick et al., 2025). However, this assumption can be misleading: performance may be insensitive to some coordinates and highly sensitive to others. We therefore introduce a *performance-based* and *directional* dissimilarity, called the signed divergence, that upper-bounds source-relative transfer loss when reusing policies from one context at another. The signed divergence naturally leads to a formal definition of task-space complexity—the minimal number of simple policies needed to achieve ε-tolerant performance across all contexts. Intuitively, systems with smooth performance landscapes (where the signed divergence changes slowly across contexts) exhibit low task-space complexity, while systems with sharp transitions require many specialized controllers. Operationally, calculating task-space complexity reduces to selecting a minimal set of sources whose ε-tolerance sets cover the context space Θ. This is exactly an instance of a *set cover* instance over the context space: each trained source induces a tolerance set, and the goal is to cover all contexts with as few such sets as possible. We design a greedy selection rule that iteratively chooses the source providing the largest marginal coverage, enjoying the classical $H(n)$-approximation guarantee for set cover.

Our contributions are as follows:

1. A directional, performance-based signed divergence that upper-bounds zero-shot transfer loss and is weaker than standard model smoothness assumptions.
2. A formal definition of task-space complexity via ε-tolerance sets, together with inner/outer geometric certificates and volume-based bounds.
3. A set-cover formulation of source selection with a greedy $H(n)$ guarantee in the finite-oracle setting, validated on linear and nonlinear systems using oracle tolerance sets.

2. Related Work

Generalization in Control and RL. The challenge of designing controllers for a family of systems is a long-standing problem in control theory. Gain scheduling is a classical engineering approach in which controllers are designed for a set of operating points and interpolated for intermediate points

(Rugh and Shamma, 2000). While practical, this method often lacks formal performance guarantees for the interpolated controllers. Our work provides a discrete alternative based on certifying transfer regions. In RL, generalization remains a central challenge, with efforts focused on domain randomization (Tobin et al., 2017), meta-learning (Finn et al., 2017), and multi-task learning (Caruana, 1997; Wilson et al., 2007; Zhang and Yang, 2021; Hendawy et al., 2024) to train policies that are robust to environmental variations. Recent work has studied when generalizable RL is tractable (Malik et al., 2021) and how to obtain provable zero-shot transfer in offline settings (Wang et al., 2025; Zhang et al., 2025). Multi-policy training with zero-shot transfer has also been explored empirically to mitigate instability in both optimization and generalization (Cho et al., 2023, 2024).

Contextual Dynamical Systems. We consider contextual dynamical systems, which are a collection of related dynamical systems parameterized by the context vector. For example, consider a set of linear systems parameterized by a context vector θ, or consider Contextual Markov Decision Processes (CMDPs). The CMDP framework, which extends MDPs by introducing a context vector that parameterizes the system dynamics and rewards, is the formal model for the systems we consider (Hallak et al., 2015; Modi et al., 2018). Much of the theory for CMDPs derives generalization bounds from smoothness assumptions on the underlying model parameters. A canonical example is Cover-Rmax algorithm proposed by (Modi et al., 2018), where the ball radii are determined by known global Lipschitz constants of the dynamics. Our work departs from this by working directly with the performance function. This can be substantially less conservative because policies may never visit the regions where model mismatch is large.

Task Similarity. Quantifying similarity between tasks is a prerequisite for transfer (Ammar et al., 2014; Zamir et al., 2018; Standley et al., 2020). Prior RL approaches include model-based comparisons of MDP parameters (Ammar et al., 2014) and bisimulation-style behavioral metrics (Ferns et al., 2004; Castro, 2020). Our signed divergence differs in being directional, performance-based, and explicitly tied to zero-shot policy reuse. Closest in spirit to our work is Preiss and Sukhatme (2021), who study suboptimal coverings for continuous spaces of control tasks. Their framework also asks how a task space can be covered efficiently, but it starts from an externally specified notion of task similarity and focuses on constructing coverings for continuous task spaces. Our setting differs in three ways: (i) the coverage relation is induced by a directional performance divergence rather than a symmetric task metric; (ii) our main object is the minimum cover size itself, interpreted as an intrinsic task-space complexity; and (iii) our geometric certificates are derived from local behavior of the performance function. Related algorithmic ideas also arise in facility-location formulations for acquiring policy libraries in human–robot settings (Vats et al., 2022). These methods reason about generalized or probabilistic coverage and query costs, whereas we study deterministic tolerance sets and use them to define a complexity notion for contextual control families.

Task-Space Complexity. The idea that tasks possess an intrinsic complexity beyond the size of the search space has appeared in several fields (Gill and Murphy, 2011). In RL, task complexity influences controller design (Kim et al., 2019) and exploration (Li et al., 2025). Building on this, our framework provides a direct, computable answer to how many policies are sufficient to solve a task space, thereby quantifying the complexity of the task space itself.

3. Preliminaries

Given a context vector $\theta \in \Theta \subset \mathbb{R}^d$, we consider a family of control systems that may differ in dynamics, reward structure, and temporal formulation. A policy π maps states to actions, and the

performance function $J(\theta, \pi)$ quantifies how well π performs on the system indexed by θ. This formulation is intentionally abstract and covers both continuous- and discrete-time systems. For instance, in a continuous-time linear system, the dynamics are $\dot{x}(t) = A(\theta)x(t) + B(\theta)u(t)$, where the matrices $A(\theta)$ and $B(\theta)$ depend on the context vector θ, while in discrete time the analogous dynamics are $x_{t+1} = A(\theta)x_t + B(\theta)u_t$. We use the convention that larger J is better. For a continuous-time linear-quadratic regulator (LQR) problem, for example, one may take $J(\theta, \pi) = -\int_0^\infty \left(x^\top Q x + u^\top R u\right)\, dt$. For a finite-horizon CMDP, one may instead use $J(\theta, \pi) = \mathbb{E}\left[\sum_{t=0}^{T} r(x_t, u_t; \theta)\right]$.

We denote by $\Pi(\theta)$ the set of source policies under consideration associated with context θ. It is the exogenously chosen source family returned by a specified training or synthesis pipeline at context θ—for example, multiple random seeds, checkpoints, or perturbed LQR gains. Throughout the algorithmic sections we focus on the practically relevant finite case. We refer to a simple (or context-independent) policy as a fixed feedback law $\pi(x)$ that chooses actions based only on the current state (e.g., a single gain matrix for linear systems, or a state-conditioned action distribution in RL). Such a policy is memoryless and static, i.e., it does not adapt to changes in environmental parameters. In contrast, a context-dependent policy chooses actions based on both state and context, either (i) explicitly, when the context θ is observable, $\pi(x, \theta)$, or (ii) implicitly, when the policy infers context from the trajectory history, $\pi(x, \text{history})$. Training a generalizable, context-dependent policy across diverse contexts is typically challenging; hence, we focus on composing a finite set of simple policies that, together, achieve ε-level performance across contexts.

This work focuses on quantifying the performance loss when a policy optimized for one context θ is applied to a different context $\tilde{\theta} \neq \theta$ (Kirk et al., 2023). This performance loss, often termed the generalization gap, is critical for understanding the transferability of policies and determining when new policies must be trained. We denote $\pi^*(\theta)$ as the best policy in the considered policy set, i.e., $\pi^*(\theta) = \arg\max_{\pi \in \Pi(\theta)} J(\theta, \pi)$. We represent the performance loss (or generalization gap) $\Delta J(\theta, \tilde{\theta})$ as the difference in performance between applying the optimal policy $\pi^*(\theta)$ in its original context θ and applying it in a different context $\tilde{\theta}$. Formally:

$$\Delta J(\theta, \tilde{\theta}) = J(\theta, \pi^*(\theta)) - J(\tilde{\theta}, \pi^*(\theta)). \tag{1}$$

This definition of the generalization gap is of particular interest because it quantifies the suboptimality incurred by zero-shot generalization. We retain ΔJ because it is the motivating one-policy quantity: it measures the loss incurred by transferring a single source-optimal controller. We say a policy π achieves ε-level performance at $\tilde{\theta}$ relative to θ if $\Delta J(\theta, \tilde{\theta}) \leq \varepsilon$.

4. Task Dissimilarity

4.1. Signed Divergence as Dissimilarity

For zero-shot transfer, what matters is not how far two contexts are in parameter space but how much performance degrades when a source policy is reused at the target. This motivates the following directional dissimilarity.

Definition 1 (Signed divergence) *Given a set of policies $\Pi(\theta_1)$ associated with context θ_1, define the signed divergence from θ_1 to θ_2 by*

$$D_{\Pi(\theta_1)}(\theta_1; \theta_2) := \sup_{\pi \in \Pi(\theta_1)} \left(J(\theta_1, \pi) - J(\theta_2, \pi)\right). \tag{2}$$

When the source family is clear from context, we write $D(\theta_1; \theta_2)$.

Compared to existing task similarity metrics, which often rely on state or policy embeddings (Agarwal et al., 2021) or heuristic definitions (Ferns et al., 2004), the proposed divergence measure has several key advantages in analyzing generalization. Its primary advantage is that it is performance-centric, directly tied to the generalization performance, providing a more relevant measure of dissimilarity than standard heuristics. By taking the supremum over a class of policies, the signed divergence provides a robust, worst-case guarantee, making it resilient to variations in training outcomes that might produce different (but near-optimal) policies. The signed divergence is also monotone in the exogenously given policy set (e.g., the limit points of a specified training pipeline at θ). As a special case, choosing $\Pi(\theta) = \{\pi^*(\theta)\}$ reduces the divergence to the loss of an optimal policy when transferred. If $\Pi_1(\theta_1) \subseteq \Pi_2(\theta_1)$ then $D_{\Pi_1}(\theta_1; \theta_2) \leq D_{\Pi_2}(\theta_1; \theta_2)$. This means the user can trade computability against conservatism by using a small finite policy set. Alternative policy sets (e.g., near-optimal policies that generalize better) may further reduce the signed divergence. Also, if $J(\cdot, \pi)$ is locally Lipschitz in θ for all $\pi \in \Pi(\theta_1)$, then $D(\theta_1; \theta_2)$ is locally Lipschitz in θ_2. Finally, the signed divergence is inherently asymmetric, i.e., $D(\theta_1; \theta_2) \neq D(\theta_2; \theta_1)$, in general. This directionality correctly captures that the difficulty of transferring from one context to another is not necessarily symmetric.[1]

4.2. Relation to Model Smoothness in CMDPs

An alternative approach to analyzing contextual systems is to assume that the underlying MDP parameters vary smoothly with context (Modi et al., 2018).

Definition 2 (Smoothness (Modi et al., 2018)) *Given a CMDP* $(\Theta, \mathcal{S}, \mathcal{A}, \mathcal{M})$ *where* Θ *is the context space,* $\mathcal{S}$ *is the state space,* $\mathcal{A}$ *is the action space,* $\mathcal{M}$ *is a function which maps a context* $\theta \in \Theta$ *to MDP parameters* $\mathcal{M}(\theta) = \{p^\theta(\cdot|\cdot,\cdot), r^\theta(\cdot,\cdot), \mu^\theta(\cdot)\}$ *and a distance metric over the context space* $\phi(\cdot,\cdot)$*, if for any two contexts* $\theta_1, \theta_2 \in \Theta$*, we have the following constraints:*

$$\|p^{\theta_1}(\cdot|s,a) - p^{\theta_2}(\cdot|s,a)\|_1 \leq L_p \phi(\theta_1, \theta_2), \tag{3a}$$

$$|r^{\theta_1}(s,a) - r^{\theta_2}(s,a)| \leq L_r \phi(\theta_1, \theta_2). \tag{3b}$$

Then the CMDP is smooth with smoothness parameters L_p *and* L_r.

While both approaches aim to bound performance variation, our signed divergence offers several practical advantages. First, our measure is policy-class dependent, whereas model-based smoothness is not. The smoothness definition in Modi et al. (2018) requires that the transition and reward functions be Lipschitz continuous over the entire state-action space and the Lipschitz constants be known. This can lead to pessimistic bounds if large changes in the dynamics occur in regions that are irrelevant to the policies of interest. In contrast, our signed divergence is defined with respect to a specific policy class $\Pi(\theta_1)$. This distinction is critical when changes in the system dynamics occur in parts of the state-action space that are irrelevant to the policies of interest. In such cases, the model parameters may change dramatically, but the performance of the relevant policies remains unaffected. The following proposition formalizes this idea.

1. In addition, the signed divergence is not a formal distance metric—it need not satisfy the triangle inequality.

Proposition 3 *There exists a family of CMDPs for which the parameter variations* $\|p^{\theta_1}(\cdot|s,a) - p^{\theta_2}(\cdot|s,a)\|_1$ *can be large for some* (s,a)*, while the divergence* $D(\theta_1;\theta_2)$ *remains small.*

This proposition shows that assuming model parameter smoothness is not a necessary condition for performance generalization. Conversely, it is a sufficient condition under this local Lipschitz assumption on $J(\theta,\pi)$, as formalized below.

Proposition 4 (Model vs. performance continuity) *Let* L_{model} *be a bound on* $|J(\theta_1,\pi)-J(\theta_2,\pi)|$ *implied by Lipschitz constants* (L_p, L_r) *as in Definition 2. Then for any policy set* Π*,*

$$L_{\text{perf}}(\Pi) := \sup_{\pi\in\Pi}\sup_{\theta_1\neq\theta_2} \frac{|J(\theta_1,\pi) - J(\theta_2,\pi)|}{\phi(\theta_1,\theta_2)} \leq L_{\text{model}}. \tag{4}$$

Together, Propositions 3 and 4 show that performance continuity can be strictly weaker than model smoothness. We refer the reader to Appendix A and B for the complete proof.

Second, our signed divergence is computable for the practically relevant policy classes considered in this work and captures the directional nature of policy transfer. As defined in Section 3, the policy set $\Pi(\theta_1)$ represents the finite collection of policies given exogenously (e.g., training a DRL agent with multiple random seeds). In this setting, the supremum in Definition 1 becomes a maximum over a finite set, making the signed divergence $D(\theta_1;\theta_2)$ directly computable once the performance of each policy in the set is evaluated as $D(\theta_1;\theta_2) = \max_{\pi\in\Pi(\theta_1)} (J(\theta_1,\pi) - J(\theta_2,\pi))$. In this formulation, $J(\cdot,\pi)$ can be estimated from system rollouts. In addition, the generalization gap is inherently directional by definition. Any symmetric metric used in standard heuristic symmetric (e.g., Euclidean/Wasserstein) cannot encode this thus can misguide source selection. Our signed divergence is asymmetric by design to capture this effect. Details are explained in Appendix C.

5. Task-Space Complexity

This section operationalizes the signed divergence to answer the following question: how many source contexts (or policies) suffice to guarantee at most ε-tolerance performance everywhere in Θ?

5.1. Tolerance Sets and Set Cover

For a source context $\theta \in \Theta$, define its *ε-tolerance set* as the targets on which the source policy sets loses at most ε.

Definition 5 (ε-tolerance set) *For a context* $\theta \in \Theta$ *and with a tolerance* $\varepsilon > 0$*, its* ε-tolerance set *is defined as* $S_\varepsilon(\theta) = \{\tilde{\theta} \in \Theta :\ D(\theta;\tilde{\theta}) \leq \varepsilon\}$.

Since the divergence takes a supremum over the policy set, $S_\varepsilon(\theta)$ is a policy set level certificate: every policy in the retained source family incurs at most ε loss on targets in $S_\varepsilon(\theta)$. In the singleton case, it reduces to the usual region of competence of one source controller.

We may regard an oracle $O : \mathbb{R}\times\Theta \mapsto \mathcal{P}(\Theta)$ that maps a tolerance level $\varepsilon \in \mathbb{R}$ and a context $\theta \in \Theta$ to an ε-tolerance set $O(\epsilon,\theta) = S_\varepsilon(\theta) \subset \Theta$. When ε is clear from context, we simply write $S(\theta)$. In practice, we operate on a finite grid of contexts rather than the continuous space; the technical treatment of the resulting discretization gap and radius adjustment is deferred to Appendix D.

Our goal is to choose the smallest possible subset of source contexts, Θ_{src}, whose tolerance sets cover Θ. This construction connects generalization guarantees directly to a well-studied set-cover problem, enabling complexity results and approximation algorithms. We directly pose source-task selection as a set cover instance over $\{S_\varepsilon(\theta)\}_{\theta\in\Theta}$ with the finite universe (Karp, 1972).

Definition 6 (Source task selection via set cover) *The source task selection (STS) problem (Cho et al., 2024) with ε-tolerance set becomes* a set cover problem *that seeks the smallest subset of source contexts whose covers blanket Θ:*

$$\min_{\Theta_{\text{src}}\subseteq\Theta} |\Theta_{\text{src}}| \quad \textit{s.t.} \quad \Theta \subseteq \bigcup_{\theta\in\Theta_{\text{src}}} S_\varepsilon(\theta). \tag{P}$$

Why cover contexts rather than policies? One could alternatively attempt to cover the policy space by selecting a representative subset of controllers. That viewpoint is useful for removing redundancy among controllers, but by itself it does not certify which target tasks are served: controllers that are close in parameter space can fail on different regions of Θ. Our notion asks whether every target context lies in some certified tolerance set.

5.2. Task-Space Complexity

Definition 7 (Task-space complexity) *We define task-space complexity as the minimal number of source contexts whose policies, when zero-shot transferred, collectively achieve a generalization gap $\leq \varepsilon$ across the task space Θ. Formally, task-space complexity at tolerance ε as*

$$\mathsf{Comp}_\varepsilon(\Theta;\Pi) \;:=\; \min_{\Theta_{\text{src}}\subseteq\Theta} \Big\{ |\Theta_{\text{src}}| :\; \Theta \subseteq \textstyle\bigcup_{\theta\in\Theta_{\text{src}}} S_\varepsilon(\theta) \Big\}. \tag{5}$$

Thus, $\mathsf{Comp}_\varepsilon(\Theta;\Pi)$ is precisely the optimum of (P).

Task-space complexity $\mathsf{Comp}_\varepsilon$ is a performance-based complexity of the task space at tolerance ε, parameterized by the chosen policy classes. It quantifies "how modular" the system must be: smoother generalization landscapes require fewer modules (controllers), while highly heterogeneous ones demand more. It upper-bounds the number of policies one must keep in a policy set to meet the tolerance everywhere.

5.3. Local Loss Rates and Geometric Certificates

We next show that local performance smoothness around each context induces a pair of geometric inner and outer certificates that bound the ε-tolerance set. Specifically, we define local upper and lower loss rates, $L^r(\theta)$ and $M^r(\theta)$, which quantify the steepest local performance decrease and increase, respectively.

Definition 8 (Local upper/lower loss rates at radius r) *For $r > 0$, define the ball $B_r(\theta) = \{\tilde\theta : \|\tilde\theta - \theta\|_2 \leq r\}$ and*

$$L^r(\theta) := \sup_{\tilde\theta\in B_r(\theta),\,\tilde\theta\neq\theta} \sup_{\pi\in\Pi(\theta)} \frac{J(\theta,\pi)-J(\tilde\theta,\pi)}{\|\theta-\tilde\theta\|_2}, \quad M^r(\theta) := \inf_{\tilde\theta\in B_r(\theta),\,\tilde\theta\neq\theta} \inf_{\pi\in\Pi(\theta)} \frac{J(\theta,\pi)-J(\tilde\theta,\pi)}{\|\theta-\tilde\theta\|_2}. \tag{6}$$

Remarks. (i) $r \mapsto L^r(\theta)$ is nondecreasing; $r \mapsto M^r(\theta)$ is nonincreasing. (ii) If $J(\cdot, \pi)$ is locally Lipschitz for each π and $\Pi(\theta)$ is finite, then $L^r(\theta) < \infty$ for small r. (iii) $M^r(\theta)$ can certainly be nonpositive. This occurs whenever transfer is beneficial in some nearby directions. A positive lower rate is needed only for the *outer* certificate below and for the corresponding lower bound on complexity; the inner certificate depends only on $L^r(\theta)$.

The local upper and lower rates translate the signed divergence into measurable slopes of the generalization performance surface, allowing geometric reasoning in context space. These quantities bound how far we can move in context space before exceeding ε-loss.

Lemma 9 (Geometric bounds on ε-tolerance set) *By Definition 8, for any $\varepsilon > 0$, the ε-tolerance set satisfies $\widehat{S}_\epsilon^-(\theta) = B_{\varepsilon/L^r(\theta)}(\theta) \subseteq S_\varepsilon(\theta)$, and if $M^r(\theta) > 0$, $S_\varepsilon(\theta) \subseteq B_{\varepsilon/M^r(\theta)}(\theta) = \widehat{S}_\epsilon^+(\theta)$.*

Proof For any $\tilde{\theta} \in B_r(\theta)$, $D(\theta; \tilde{\theta}) = \sup_{\pi \in \Pi(\theta)} \big(J(\theta, \pi) - J(\tilde{\theta}, \pi)\big) \le L^r(\theta) \|\theta - \tilde{\theta}\|_2$, so $\|\theta - \tilde{\theta}\| \le \varepsilon / L^r(\theta)$ implies $D(\theta; \tilde{\theta}) \le \varepsilon$. The outer bound follows similarly from the lower rate. ■

Interpretation. The inner coverage set provides a certified region where the source policy class meets the tolerance; the outer coverage set bounds how far tolerance can possibly extend. An upper (resp. lower) bound rate yields an *inner* (resp. *outer*) geometric certificate. This result provides geometric certificates for local generalization: the inner ball certifies a guaranteed region of ε-performance, while the outer ball bounds how far such performance can extend.

Theorem 10 (Bounds on task-space complexity) *Suppose $\Theta \subset \mathbb{R}^d$ is compact with $\mathrm{Vol}(\Theta) < \infty$. Assume that for some $r > 0$, $L_{\max} := \sup_{\theta \in \Theta} L^r(\theta) < \infty$ and $M_{\min} := \inf_{\theta \in \Theta} M^r(\theta) > 0$. Fix $\varepsilon > 0$ with $\varepsilon / L_{\max}^r \le r$ and $\varepsilon / M_{\min}^r \le r$. Then the ε-complexity of the task space satisfies*

$$\frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}\big(B_{\varepsilon/M_{\min}}\big)} \;\le\; \mathsf{Comp}_\varepsilon \;\le\; C_d \, \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}\big(B_{\varepsilon/L_{\max}}\big)}, \tag{7}$$

where B_r denotes a d-dimensional Euclidean ball of radius r and C_d is a geometric constant, e.g., $C_d \le 3^d$ for a lattice cover.

Proof The volume of a d-dimensional Euclidean ball of radius r is $\mathrm{Vol}(B_r) = v_d r^d$, where $v_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$. This follows from the Gamma identity $\int_{\mathbb{R}^d} e^{-\|x\|^2} dx = \pi^{d/2} = \int_0^\infty e^{-r^2} S_{d-1} r^{d-1} dr = S_{d-1} \frac{1}{2} \Gamma(\frac{d}{2})$, which implies $S_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ and hence $v_d = \frac{S_{d-1}}{d}$. *Upper bound.* Tile the domain Θ by a lattice with spacing proportional to $r_{\text{in}} = \varepsilon / L_{\max}$. Each lattice point corresponds to a ball $B_{r_{\text{in}}}$, which together cover Θ. The number of required centers is therefore upper bounded by the ratio of volumes, up to a packing constant: $\mathsf{Comp}_\varepsilon \le C_d \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(B_{r_{\text{in}}})} = C_d \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(B_{\varepsilon/L_{\max}})}$. *Lower bound.* If N sources suffice to cover Θ, then the covered region is a union of N balls of radius at most $r_{\text{out}} = \varepsilon / M_{\min}$. By subadditivity of volume, $\mathrm{Vol}(\Theta) \le N \, \mathrm{Vol}(B_{r_{\text{out}}}) = N \, \mathrm{Vol}(B_{\varepsilon/M_{\min}})$, which gives the lower bound $\mathsf{Comp}_\varepsilon \ge \mathrm{Vol}(\Theta) / \mathrm{Vol}(B_{\varepsilon/M_{\min}})$. ■

Each trained policy certifies a region where it performs well. The ratio of total context-space volume to the volume of one certified region tells us roughly how many policies are needed. The signed divergence can be negative when a transferred policy improves performance; the bounds above use its magnitude through L^r, M^r. Although (7) uses aggregated quantities $L_{\max}$ and $M_{\min}$ over the task family, the result is still instance-specific: these constants are computed from the given contextual system and chosen source families, rather than from universal model-class Lipschitz constants as in Modi et al. (2018).

5.4. Greedy Set Cover

Having established that task-space complexity reduces to a set-cover instance, we now describe a greedy algorithm (Algorithm 1), a practical and tractable algorithm for solving it. The problem (P) is NP-hard even with information about the oracle O. This method iteratively builds the set of source tasks, Θ_{src}, by selecting the source task that covers the greatest number of currently uncovered contexts at each step. The SELECT procedure can be implemented in various ways, such as greedy selection with oracle information, random selection to explore the space, or a grid-based approach.

Algorithm 1 Greedy Meta-Algorithm for Set Cover

Require: Context space Θ, ε-tolerance set S_ε
Initialize uncovered set $U \leftarrow \Theta$ and source set $\Theta_{\text{src}} \leftarrow \emptyset$
while $U \neq \emptyset$ **do**
  Select $\theta^* \in \Theta$ maximizing $|S_\varepsilon(\theta) \cap U|$
  Update $U \leftarrow U \setminus S_\varepsilon(\theta^*)$
  Update $\Theta_{\text{src}} \leftarrow \Theta_{\text{src}} \cup \{\theta^*\}$
end while
return Θ_{src}

If tolerance sets are constructed naively on a finite grid of size $n = |\Theta|$ with at most m source policies per context, exact cover construction may require on the order of $O(n^2 m)$ performance evaluations. The point of the algorithm is once tolerance sets are available, it compresses a dense candidate set into a smaller deployable subset and reveals the associated task-space complexity.

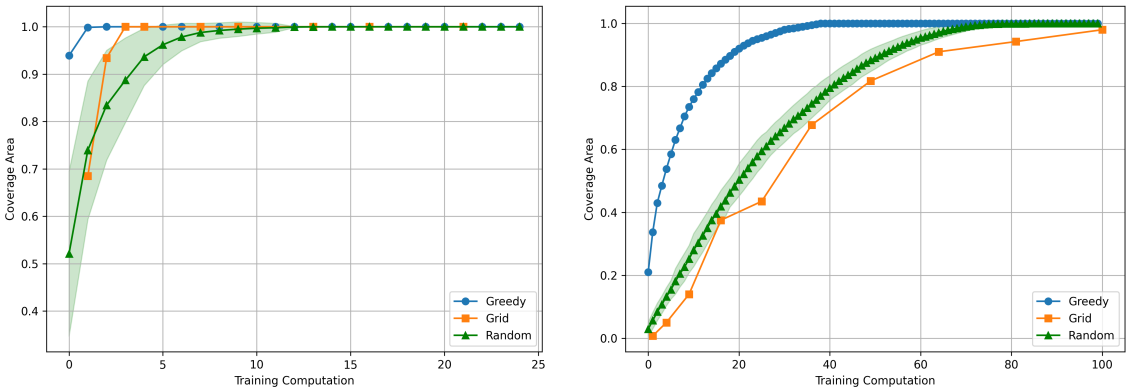
Theorem 11 (Greedy Algorithm Performance Guarantee) *For a discretized context space Θ of size $n = |\Theta|$, the greedy strategy (Algorithm 1) returns at most $f \cdot (\ln n + 1)$ source tasks, where f is the optimal value of* (P).

Proof Algorithm 1 is exactly the classical greedy algorithm for set cover problem on the universe Θ. At each step, the algorithm selects the source task θ^* whose cover $\widehat{S}(\theta^*)$ includes the maximum number of previously uncovered contexts. The standard approximation guarantee states that greedy set cover returns at most $H(n)$ times the optimum, where $H(n) \leq \ln n + 1$ is the nth harmonic number (Vazirani, 2001). This provides a strong theoretical guarantee on the performance of our task selection method. ■

6. Numerical Experiments

In this section, we evaluate the efficiency–how many policies are needed to achieve certified coverage versus standard heuristics, accounting for both training and evaluation costs. We validate our framework on two representative contextual dynamical systems: a continuous-time linear Mass-Spring-Damper (MSD) system with LQR controllers and a nonlinear CartPole system with deep RL controllers. For each system, we construct a discretized context space and corresponding policy sets to evaluate certified coverage, following the procedure detailed in Appendix E.

We evaluate our task selection framework on the number of policies that must be trained. The primary cost in solving a family of contextual tasks stems from training individual controllers, which is computationally expensive. Therefore, a crucial metric of success is minimizing the total number of policies required to cover the entire context space. We compare our greedy algorithm (Algorithm 1) against baselines: a uniform grid selection and a random selection. A key limitation of the



(a) Mass-Spring-Damper (MSD)

(b) CartPole

Figure 1: **Coverage vs. training computation.** (a) Mass-Spring-Damper with LQR policies; (b) CartPole with PPO policies. Greedy selection (blue) covers the context space with fewer trained policies than grid (orange) and random (green). Shaded bands indicate variability across seeds.

uniform grid baseline is its dependence on a pre-selected grid resolution. A coarse grid may fail to place policies in critical regions, while a fine grid is inefficient. Greedy approach avoids this hyper-parameter tuning by adaptively placing policies based on the local performance landscape. Figure 1 shows coverage versus compute for MSD and CartPole. Across both systems, greedy selection achieves a given coverage level using fewer trained policies than uniform or random selection. The advantage is largest in regions where the performance landscape is anisotropic or exhibits sharp changes (visible as early steep gains for Greedy). Importantly, the gap persists until near-complete coverage, indicating that gains do not rely on cherry-picking easy regions.

7. Conclusion

We introduced a performance-centric, directional task dissimilarity—the signed divergence—and used it to define task-space complexity: the minimum number of context-independent policies needed to guarantee ε-level performance throughout a context space. Under mild local performance smoothness, we derived inner/outer geometric certificates and instance-dependent volume bounds. Casting source selection as set cover leads to a simple greedy algorithm with a harmonic-factor guarantee that, in practice, achieves the same ε-coverage with substantially fewer trained policies than uniform or random baselines on both linear and nonlinear systems. The main limitation of the present work is explicit: exact greedy selection currently relies on oracle or exhaustive tolerance-set construction. We view this not as a flaw in the complexity definition, but as the natural separation between a structural theory of task-space complexity and the estimation problem required to make that theory scalable.

Future work can develop statistically efficient estimators of local loss rates and certified ε-tolerance sets via a geometric set-cover formulation with LP relaxation and VC-dimension-based guarantees. One could also extend this framework to multi-task or meta-learning settings, where the goal is to train a single, adaptable policy rather than a discrete set. An important direction is to integrate certificate-driven selection with multi-task or switching policies inside Π, thereby tightening complexity further while preserving certifiability.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF) under the CAREER award (#2239566) and CCF (#2236484), and the Kwanjeong Educational Foundation Ph.D. scholarship program.

References

Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021.

Haitham Bou Ammar, Eric Eaton, Matthew E Taylor, Decebal Constantin Mocanu, Kurt Driessens, Gerhard Weiss, and Karl Tüyls. An automated measure of mdp similarity for transfer in reinforcement learning. In *28th AAAI Conference on Artificial Intelligence, AAAI 2014*, pages 31–37. AI Access Foundation, 2014.

Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me–the case for context in reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10069–10076, 2020.

Jung-Hoon Cho, Sirui Li, Jeongyun Kim, and Cathy Wu. Temporal transfer learning for traffic optimization with coarse-grained advisory autonomy. *arXiv preprint arXiv:2312.09436*, 2023.

Jung-Hoon Cho, Vindula Jayawardana, Sirui Li, and Cathy Wu. Model-based transfer learning for contextual reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 88279–88319, 2024.

Jeffery Dick, Saptarshi Nath, Christos Peridis, Eseoghene Ben-Iwhiwhu, Soheil Kolouri, and Andrea Soltoggio. Statistical context detection for deep lifelong reinforcement learning. In *Conference on Lifelong Learning Agents*, pages 1013–1031. PMLR, 2025.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *UAI*, volume 4, pages 162–169, 2004.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

T Grandon Gill and W Murphy. Task complexity and design science. In *9th International Conference on Education and Information Systems, Technologies and Applications (EISTA 2011)*, 2011.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Ahmed Hendawy, Jan Peters, and Carlo D'Eramo. Multi-task reinforcement learning with mixture of orthogonal experts. In *The Twelfth International Conference on Learning Representations*, 2024.

Petros A Ioannou and Jing Sun. *Robust adaptive control*, volume 1. PTR Prentice-Hall Upper Saddle River, NJ, 1996.

Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

Richard M Karp. Reducibility among combinatorial problems. 1972.

Dongjae Kim, Geon Yeong Park, John P O Doherty, and Sang Wan Lee. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nature communications*, 10(1):5738, 2019.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

Ang Li, Zhihang Yuan, Yang Zhang, Shouda Liu, and Yisen Wang. Know when to explore: Difficulty-aware certainty as a guide for llm reinforcement learning. *arXiv preprint arXiv:2509.00125*, 2025.

Dhruv Malik, Yuanzhi Li, and Pradeep Ravikumar. When is generalizable reinforcement learning tractable? *Advances in Neural Information Processing Systems*, 34:8032–8045, 2021.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. In *ICLR*, 2016.

James A Preiss and Gaurav S Sukhatme. Suboptimal coverings for continuous spaces of control tasks. In *Learning for Dynamics and Control*, pages 547–558. PMLR, 2021.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Wilson J Rugh and Jeff S Shamma. Research on gain scheduling. *Automatica*, 36(10):1401–1425, 2000.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

Shivam Vats, Oliver Kroemer, and Maxim Likhachev. Synergistic scheduling of learning and allocation of tasks in human-robot teams. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2789–2795. IEEE, 2022.

Vijay V Vazirani. *Approximation algorithms*, volume 1. Springer, 2001.

Zhiyong Wang, Chen Yang, John Lui, and Dongruo Zhou. Provable zero-shot generalization in offline reinforcement learning. *arXiv preprint arXiv:2503.07988*, 2025.

Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*, pages 1015–1022, 2007.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

Chi Zhang, Ziying Jia, George K Atia, Sihong He, and Yue Wang. Pessimism principle can be effective: Towards a framework for zero-shot transfer reinforcement learning. *arXiv preprint arXiv:2505.18447*, 2025.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

Kemin Zhou and John Comstock Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

Appendix

Appendix A. Proof of Proposition 3

Proof We construct a CMDP family where transition kernels differ maximally at an unreachable state while the signed divergence is zero.

Consider a finite-horizon MDP with $\mathcal{S} = \{s_{\text{start}}, s_{\text{safe}}, s_{\text{critical}}, s_{\text{end}}\}$, $\mathcal{A} = \{a_1, a_2\}$, horizon $T = 2$, and reward 1 iff s_{end} is reached. Let the policy class under consideration $\Pi = \{\pi\}$ such that $\pi(s_{\text{start}}) = a_1$. The transition dynamics from the start state and action a_1 is $p(s_{\text{safe}}|s_{\text{start}}, a_1) = 1$, which is independent of the context θ. Any other action from s_{start}, or any action from s_{safe}, leads to the terminal state s_{end} with probability 1. Therefore, for the policy π, the trajectory is always $s_{\text{start}} \xrightarrow{a_1} s_{\text{safe}} \xrightarrow{a_k} s_{\text{end}}$, yielding a total reward of 1. Let contexts differ only at s_{critical}, with $p^{\theta_1}(s_{\text{end}} \mid s_{\text{critical}}, a) = 1$ and $p^{\theta_2}(s_{\text{start}} \mid s_{\text{critical}}, a) = 1$ for all a. Under θ_1, the critical state leads to termination, while under θ_2, it leads back to the start, creating a loop and yielding 0 reward.

Since $\mu_\pi(s_{\text{critical}}) = 0$, $D(\theta_1; \theta_2) = \sup_{\pi' \in \Pi}(J(\theta_1, \pi') - J(\theta_2, \pi')) = J(\theta_1, \pi) - J(\theta_2, \pi) = 0$. Yet for any a, $\|p^{\theta_1}(\cdot \mid s_{\text{critical}}, a) - p^{\theta_2}(\cdot \mid s_{\text{critical}}, a)\|_1 = \|\delta_{s_{\text{end}}} - \delta_{s_{\text{start}}}\|_1 = 2$. Moreover, along a continuous interpolation between θ_1 and θ_2, the local Lipschitz constant of $\theta \mapsto p^\theta(\cdot \mid s_{\text{critical}}, a)$ can be made arbitrarily large near the switching point while $D(\theta_1; \theta_2) = 0$. This shows that parameter-based smoothness can be a poor indicator of performance generalization, whereas our divergence metric correctly identifies that tasks θ_1 and θ_2 are equivalent from the perspective of the policy set Π. ■

Appendix B. Proof of Proposition 4

Proof Let $\pi \in \Pi$ be an arbitrary policy. The performance is given by $J(\theta, \pi) = \mathbb{E}_{s_0 \sim q^\theta}[V_\pi^\theta(s_0)]$. Assuming a fixed initial state distribution q across contexts, the performance difference is bounded by the maximum difference in the value function as $|J(\theta_1, \pi) - J(\theta_2, \pi)| \leq \|V_\pi^{\theta_1} - V_\pi^{\theta_2}\|_\infty$, where V_π^θ is the value function for policy π in context θ. The value function is the unique fixed point of the Bellman operator T_π^θ, defined as $(T_\pi^\theta V)(s) = r^\theta(s, \pi(s)) + \gamma \sum_{s'} p^\theta(s'|s, \pi(s))V(s')$. Using the fixed-point property, $V_\pi^\theta = T_\pi^\theta V_\pi^\theta$, we analyze the difference as: $\|V_\pi^{\theta_1} - V_\pi^{\theta_2}\|_\infty = \|T_\pi^{\theta_1} V_\pi^{\theta_1} - T_\pi^{\theta_2} V_\pi^{\theta_2}\|_\infty \leq \|T_\pi^{\theta_1} V_\pi^{\theta_1} - T_\pi^{\theta_1} V_\pi^{\theta_2}\|_\infty + \|T_\pi^{\theta_1} V_\pi^{\theta_2} - T_\pi^{\theta_2} V_\pi^{\theta_2}\|_\infty$. Using the γ-contraction of $T_\pi^{\theta_1}$ with respect to the infinity norm, rearranging gives:

$$(1 - \gamma)\|V_\pi^{\theta_1} - V_\pi^{\theta_2}\|_\infty \leq \|T_\pi^{\theta_1} V_\pi^{\theta_2} - T_\pi^{\theta_2} V_\pi^{\theta_2}\|_\infty. \tag{8}$$

Now, we bound the term on RHS. For any state s:

$$\begin{aligned} \left|(T_\pi^{\theta_1} - T_\pi^{\theta_2})\, V_\pi^{\theta_2}(s)\right| &= \left| r^{\theta_1}(s,a) - r^{\theta_2}(s,a) + \gamma \textstyle\sum_{s'} \left(p^{\theta_1}(s'|s,a) - p^{\theta_2}(s'|s,a)\right) V_\pi^{\theta_2}(s') \right| \\ &\leq \left| r^{\theta_1}(s,a) - r^{\theta_2}(s,a) \right| + \gamma \left| \textstyle\sum_{s'} \left(p^{\theta_1}(s'|s,a) - p^{\theta_2}(s'|s,a)\right) V_\pi^{\theta_2}(s') \right|. \end{aligned} \tag{9}$$

where $a = \pi(s)$. Using the smoothness assumptions from Definition 2 and the fact that the value function is bounded by $\|V_\pi^{\theta_2}\|_\infty \leq \frac{R_{\max}}{1-\gamma}$, where $R_{\max} = \sup_{\theta,s,a} |r^\theta(s,a)|$:

$$\begin{aligned} |(T_\pi^{\theta_1} V_\pi^{\theta_2})(s) - (T_\pi^{\theta_2} V_\pi^{\theta_2})(s)| &\leq L_r \phi(\theta_1, \theta_2) + \gamma \|p^{\theta_1}(\cdot|s,a) - p^{\theta_2}(\cdot|s,a)\|_1 \|V_\pi^{\theta_2}\|_\infty \\ &\leq L_r \phi(\theta_1, \theta_2) + \gamma L_p \phi(\theta_1, \theta_2) \tfrac{R_{\max}}{1-\gamma}. \end{aligned} \tag{10}$$

Taking the supremum over all states s, we get:

$$\|T_\pi^{\theta_1} V_\pi^{\theta_2} - T_\pi^{\theta_2} V_\pi^{\theta_2}\|_\infty \leq (L_r + \tfrac{\gamma L_p R_{\max}}{1-\gamma})\phi(\theta_1, \theta_2). \tag{11}$$

Substituting this back into Equation 8:

$$\|V_\pi^{\theta_1} - V_\pi^{\theta_2}\|_\infty \leq (\tfrac{L_r}{1-\gamma} + \tfrac{\gamma L_p R_{\max}}{(1-\gamma)^2})\phi(\theta_1, \theta_2). \tag{12}$$

Then, by definition, $L_{\text{model}} = \frac{L_r}{1-\gamma} + \frac{\gamma L_p R_{\max}}{(1-\gamma)^2}$. We have shown that for any policy π, $|J(\theta_1, \pi) - J(\theta_2, \pi)| \leq L_{\text{model}}\phi(\theta_1, \theta_2)$. Since this holds for all $\pi \in \Pi$, it must also hold for the supremum:

$$L_{\text{perf}}(\Pi) = \sup_{\pi \in \Pi} \sup_{\theta_1 \neq \theta_2} \tfrac{|J(\theta_1,\pi) - J(\theta_2,\pi)|}{\phi(\theta_1,\theta_2)} \leq L_{\text{model}}. \tag{13}$$

■

Appendix C. Why directionality matters

Generalization gap is inherently directional by definition. Notice that reversing the arrow asks a different question because the policy you deploy has been trained somewhere else. Any symmetric metric used in standard heuristic symmetric (e.g., Euclidean/Wasserstein on parameters or rewards) cannot encode this thus can misguide source selection. Our signed divergence is asymmetric by design to capture this effect.

Inherent Asymmetry of Signed Divergence. For exposition, fix a common policy set Π. Let $f_\Pi(\pi) := J(\theta_1, \pi) - J(\theta_2, \pi)$. Then, $D(\theta_1; \theta_2) = \sup_{\pi \in \Pi} f_\Pi(\pi)$, $D(\theta_2; \theta_1) = \sup_{\pi \in \Pi}\{-f_\Pi(\pi)\} = -\inf_{\pi \in \Pi} f_\Pi(\pi)$. Unless $f_\Pi(\pi)$ takes the same value for all π (a degenerate case), we have $\sup f_\Pi \neq -\inf f_\Pi$, hence $D(\theta_1; \theta_2) \neq D(\theta_2; \theta_1)$. In practice, the asymmetry can be even stronger because the admissible sets $\Pi(\theta_1)$ and $\Pi(\theta_2)$ need not coincide.

Toy CMDP illustrating directionality. Consider a one-step CMDP with start state s_0 and terminal states s_A, s_B. Action a_A (resp. a_B) deterministically reaches s_A (resp. s_B). Let $\Pi = \{\pi_A, \pi_B\}$ where $\pi_A(s_0) = a_A$ and $\pi_B(s_0) = a_B$. Rewards depend on context: at θ_1, $(r(s_A), r(s_B)) = (10, 0)$; at θ_2, $(5, 8)$. Then, $f_\Pi(\pi_A) = 10 - 5 = 5, f_\Pi(\pi_B) = 0 - 8 = -8$, so $D(\theta_1; \theta_2) = \sup\{5, -8\} = 5$ while $D(\theta_2; \theta_1) = -\inf\{5, -8\} = 8$. The generalization gap differs by direction.

Misguided source selection under symmetric distances. Let $\varepsilon \in (5, 8]$ in the toy CMDP above. Any chooser that relies only on a symmetric distance δ (e.g., Euclidean/Wasserstein on rewards/parameters) is indifferent between θ_1 and θ_2 since $\delta(\theta_1, \theta_2) = \delta(\theta_2, \theta_1)$, and may select θ_2 as the source. This fails the tolerance requirement because $D(\theta_2; \theta_1) = 8 > \varepsilon$, i.e., $\theta_1 \notin S_\varepsilon(\theta_2)$. In contrast, selecting θ_1 succeeds since $D(\theta_1; \theta_2) = 5 \leq \varepsilon$, so $\theta_2 \in S_\varepsilon(\theta_1)$.

Appendix D. Discretization Details

In practice, we solve the set cover problem on a finite grid of points, $\Theta_{\text{grid}} \subseteq \Theta$, rather than the entire continuous space. However, covering all points on this grid does not automatically guarantee that the spaces between the grid points are also covered, introducing a potential *discretization gap*. To address this gap while preserving the original guarantee, we can adjust our ε-tolerance set. For instance, if the set has the shape of a ball, we effectively shrink the radius of our balls to create a

buffer. Let δ_{grid} be the resolution of our grid. To ensure a cover for the grid translates to a valid cover for the entire continuous space, we adjust the radius by δ_{grid}. For instance, consider a d-dimensional hypercubic grid with uniform spacing h. The point furthest from any grid node is the center of a hypercube, which gives a maximum discretization error of $\delta_{\text{grid}} = \frac{h\sqrt{d}}{2}$. By using an adjusted radius, $r_{\text{adj}} = \max\{0,\, r - \delta_{\text{grid}}\}$, we ensure that if a grid point is covered, the entire region around it up to the next grid point is also covered. For notational simplicity, we will use Θ to refer to the relevant finite context sets, with the understanding that the discretization gap bridging step is implicitly handled.

Proposition 12 (Grid-to-continuum cover via radius shrink) *Let Θ_{grid} be a d-dimensional hypercubic grid with spacing h over Θ. If each grid point θ is covered by a ball $B_r(\theta)$, then the union of shrunken balls $B_{r-\delta_{\text{grid}}}(\theta)$ with $\delta_{\text{grid}} = \frac{\sqrt{d}}{2}h$ covers all of Θ (interpret negative radii as empty).*

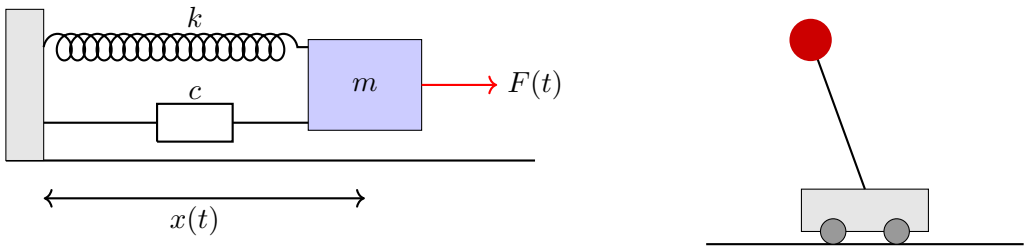
Proof Any $\tilde{\theta} \in \Theta$ lies within distance at most δ_{grid} of some grid node. If that node is in $B_r(\theta)$, then $\tilde{\theta} \in B_{r-\delta_{\text{grid}}}(\theta)$ by the triangle inequality. ■

Appendix E. Experimental Setup

Mass-Spring-Damper system with LQR controller The MSD system operates in a gravity-free environment and consists of a mass m connected to a spring with stiffness k and a damper with damping coefficient c, connected in parallel. We consider the system variation with different mass parameters. The dynamics of the Mass-Spring-Damper system are governed by Newton's second law and are described by the following equation:

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = F(t), \tag{14}$$

where $x(t)$ represents the displacement of the mass from its equilibrium position, $\dot{x}(t)$ is its velocity, and $\ddot{x}(t)$ denotes acceleration. The term $F(t)$ is the external force acting on the mass.



(a) Mass-Spring-Damper System. (b) CartPole System.

Figure 2: Diagrams of two dynamical systems.

Defining the state vector $\mathbf{x}(t)$ and input $u(t)$ as $\mathbf{x}(t) = \begin{bmatrix} x(t) & \dot{x}(t) \end{bmatrix}^\top$ and $\mathbf{u}(t) = \begin{bmatrix} 0 & F(t) \end{bmatrix}^\top$, the state-space equations are $\dot{\mathbf{x}}(t) = A_{(k,m,c)}\mathbf{x}(t) + B_{(k,m,c)}\mathbf{u}(t)$ with $A_{(k,m,c)} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}$ and $B_{(k,m,c)} = \begin{bmatrix} 0 & \frac{1}{m} \end{bmatrix}^\top$, where $x(t)$ represents the displacement of the mass from its equilibrium

position, $\dot{x}(t)$ is its velocity, and $F(t)$ is the external force acting on the mass. The output equation (measuring displacement) is $y(t) = x(t) = C\mathbf{x}(t) = \begin{bmatrix}1 & 0\end{bmatrix}\mathbf{x}(t)$. In our CMDP formulation for the MSD system, the context vector is defined by the spring stiffness and the damping coefficient, i.e., $\theta = [k, c]^\top$, while the mass is held constant at $m = 1.0$. The context space Θ is a two-dimensional grid where both k and c range from 0.1 to 8.0, discretized into 100 points each. We fix the initial condition distribution and rollout horizon to isolate context effects on performance.

For each system, an optimal controller can be derived using LQR, which minimizes the cost function $J = \int_0^\infty \left(x^\top(t)Qx(t) + u^\top(t)Ru(t)\right) dt$, where Q and R are positive semi-definite and positive definite matrices, respectively, used to weight state and control efforts. The LQR controller is computed by solving the Continuous-time Algebraic Riccati Equation (CARE): $A^\top P + PA - PBR^{-1}B^\top P + Q = 0$, where P is the unique positive definite solution to CARE. The optimal gain matrix K^* is then given by $K^* = R^{-1}B^\top P$. For this experiment, we set the LQR weighting matrices to $Q = \mathrm{diag}(10, 4)$ and $R = [1]$, penalizing the displacement more heavily than velocity. All controllers are evaluated under the same quadratic criterion to maintain comparability across contexts. To define the policy set $\Pi(\theta)$ for a given context θ, we model the variability inherent in practical controller synthesis. Instead of using only the single optimal gain K^*, we create an ensemble of controllers by adding Gaussian noise to the optimal gain. Specifically, for each context θ, the policy set is defined as $\Pi(\theta) = \{K_1, K_2, \ldots, K_N\} = \{K^* + \Delta_i\}_{i=1}^N$, where each Δ_i is a random perturbation sampled from a zero-mean normal distribution.

CartPole with DRL controller To further validate our algorithm, we examine the classic CartPole system with DRL controller as a test case. We use CARL, a CMDP variant of the standard OpenAI Gym implementation (Benjamins et al., 2023). In this scenario, we consider task variations by varying the pole length and the mass of cart as context parameters. The objective is to apply forces to the cart to balance the pole in an upright position while keeping the cart within the track boundaries. For the CartPole CMDP, the context vector is $\theta = [\text{pole length}, \text{cart mass}]^\top$. The context space Θ is a 20×20 grid where the pole length ranges from 0.25 to 5.0 and the cart mass ranges from 0.5 to 10.0. For each MDP, we train the policy for different random seeds using Proximal Policy Optimization (PPO) (Schulman et al., 2017) with Stable Baseline 3 (Raffin et al., 2021), to create the policy set Π. The performance $J(\theta, \pi)$ is the expected cumulative reward (balancing time) over an episode. This environment serves as a benchmark for nonlinear systems where the dynamics are not explicitly known. For all policies, we evaluate $J(\theta, \pi)$ as the expected episodic return under the same seeds used for certification estimates.