AIPR: ADAPTIVE ISOMETRIC PLASTICITY REGULARIZATION FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep reinforcement learning encounters a balance breakdown between stability and plasticity during long–term training, which causes a degradation in the network's continuous optimization capability and impairs the agent's long–term learning ability. Some existing approaches employ the Deviation from Isometry (DfI) to measure how far a weight matrix strays from an isometric structure. However, discrete parameter reinitialization interrupts the continuous optimization process, while abrupt parameter updates may disrupt the policy distribution. To address this issue, we propose Adaptive Isometric Plasticity Regularization (AIPR). This method transforms DfI from a trigger metric for discrete parameter reinitialization into a continuously differentiable geometric regularization term. By jointly optimizing this term alongside the reinforcement learning task objective, AIPR continuously constrains the evolving geometric state of network weights throughout training. Concurrently, AIPR incorporates a DfI–aware adaptive gating mechanism that dynamically regulates regularization intensity based on the network's current weight geometric state. Considering potential discrepancies in weight geometry between the Actor and Critic, AIPR separately characterizes their geometric states, generates corresponding adaptive regularization weights, and preserves the original task coupling within the Actor–Critic framework. AIPR is compatible with both on-policy and off-policy learning, enabling seamless integration with PPO and SAC. Empirical results demonstrate that AIPR achieves competitive learning performance across multiple continuous control benchmarks.

1 INTRODUCTION

In recent years, deep reinforcement learning (DRL) has achieved remarkable success in a wide range of applications, including complex decision–making and continuous control tasks such as game playing and robotic manipulation (Mnih et al., 2015; Wołczyk et al., 2021). However, as training progresses, neural networks may gradually suffer from representation degradation, leading to impaired adaptability and reduced long–term learning capability of the agent (Tang et al., 2025; Wu et al., 2026; Zhou et al., 2025). Existing research (Dohare et al., 2024) demonstrates that an imbalance between network stability and plasticity during long–term training can lead to plasticity loss, thereby weakening the agent's ability to sustain continuous learning and optimization.

Recent work (Han et al., 2026) addresses the above challenge from a weight geometry perspective by formulating parameter intervention as a stability–plasticity trade–off. Stability is measured by the squared Frobenius error (SFE), which captures parameter changes after geometric intervention, while plasticity is characterized by DfI, which measures deviations from near–isometric weight structures. When DfI indicates substantial geometric degradation, Han et al. (Han et al., 2026) perform parameter reinitialization to restore the degraded weight geometry. Related studies (Lewandowski et al., 2025) further show that maintaining orthogonal or near–isometric structures improves singular value distributions and optimization stability, providing theoretical support for geometry–based approaches. However, existing geometric intervention methods (Han et al., 2026; Chung et al., 2024) have limitations. Parameter reinitialization relies on discrete interventions triggered only after substantial geometric degradation, preventing smooth adaptation to evolving weight dynamics and potentially introducing optimization instability.

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.

To address this issue, we propose Adaptive Isometric Plasticity Regularization (AIPR), which reformulates the DfI metric from a discrete reinitialization criterion into a continuous and differentiable geometric regularization term. This formulation enables weight geometric constraints to be directly incorporated into the gradient–based optimization process within the Actor–Critic framework. AIPR continuously adapts weight constraints according to the evolving geometric state of the network, enabling an adaptive balance between preserving desirable weight geometry and optimizing the task objective. The main contributions of this paper are summarized as follows:

1. We reformulate DfI from a discrete reinitialization criterion into a continuous differentiable geometric regularization objective, enabling gradient–based optimization toward near–isometric weight structures.

2. We propose a DfI–aware adaptive gating mechanism that dynamically adjusts regularization strength according to weight geometry. It reduces constraints when weights remain near–isometric and strengthens regularization when geometric degradation occurs, balancing geometric preservation and task–driven optimization.

3. We conduct extensive experiments on MuJoCo, Robosuite, and CARL benchmarks, and further extend AIPR from PPO to SAC to evaluate its generality beyond on–policy learning. Results demonstrate consistent improvements in learning performance across diverse environments.

2 Related Work

2.1 Representation Degradation

Plasticity loss in deep reinforcement learning (DRL) has recently attracted increasing attention (Klein et al., 2024). Reinforcement learning agents continuously collect data through environment interactions, resulting in non–stationary optimization dynamics and distribution shifts that may gradually reduce network plasticity during long–term training (Lyle et al., 2023; Juliani and Ash, 2024), ultimately limiting policy optimization and future adaptability (Tang et al., 2025; Zhou et al., 2025). To address this issue, recent works have explored mechanisms for restoring or preserving network adaptability. Dohare et al. (Dohare et al., 2024) analyzed plasticity loss in deep continuous learning, while DASH (Shin et al., 2024) and FIRE (Han et al., 2026) investigated recovery strategies based on network updates and weight geometry, respectively. However, existing recovery approaches mainly rely on discrete intervention mechanisms, abrupt parameter reinitialization may disrupt optimization trajectories and policy distributions, increasing training instability. Therefore, an open challenge remains, how to transform DfI from a discrete recovery criterion into a continuous geometric regularization signal and adaptively adjust constraint strength according to evolving weight geometry.

2.2 Continuous Geometric Constraints

Existing studies improve optimization stability and representation quality in deep reinforcement learning by explicitly regulating weight geometry. Representative approaches include spectral normalization (Bjorck et al., 2021), spectral regularization (Nauman et al., 2024), and Parseval regularization (Chung et al., 2024). Spectral normalization controls the spectral norm to regulate transformation scales, while spectral and Parseval regularization impose continuous constraints on spectral properties. However, existing continuous geometric regularization methods typically employ fixed regularization strengths (Chung et al., 2024; Nauman et al., 2024), limiting their adaptability to evolving weight geometry. Strong constraints may unnecessarily restrict task–driven updates when weights remain geometrically favorable, whereas insufficient constraints may fail to correct severe degradation. Therefore, an adaptive mechanism that dynamically adjusts geometric regularization according to the current geometric state remains necessary. To address this limitation, we propose AIPR, which reformulates DfI as a continuous differentiable regularization signal and introduces a state–aware soft gating mechanism for adaptive regulation throughout training.

2.3 Geometric Differences in the Actor–Critic Framework

Recent studies (Juliani and Ash, 2024; Lewandowski et al., 2025) show that long–term Actor–Critic training can suffer from plasticity loss and representation degradation. In the Actor network, degraded representations reduce state discriminability and limit policy expressiveness, while in the Critic network, they impair value estimation and the reliability of learning signals such as advantages and TD errors (Moalla et al., 2024; Wu et al., 2026). Consequently, representation degradation in both networks can negatively affect policy optimization and value learning dynamics (Lewandowski et al., 2025; Garcin et al., 2025). Although Actor and Critic are jointly optimized within the same framework, they pursue different objectives and may exhibit distinct weight geometric states and evolution patterns (Garcin et al., 2025). Therefore, applying identical geometric constraints may not adequately match their individual optimization dynamics (Garcin et al., 2025; Raileanu and Fergus, 2021). Motivated by this observation, we model Actor and Critic geometric states separately, allowing each network to adaptively adjust regularization strength according to its own geometric condition while preserving the original Actor–Critic optimization structure.

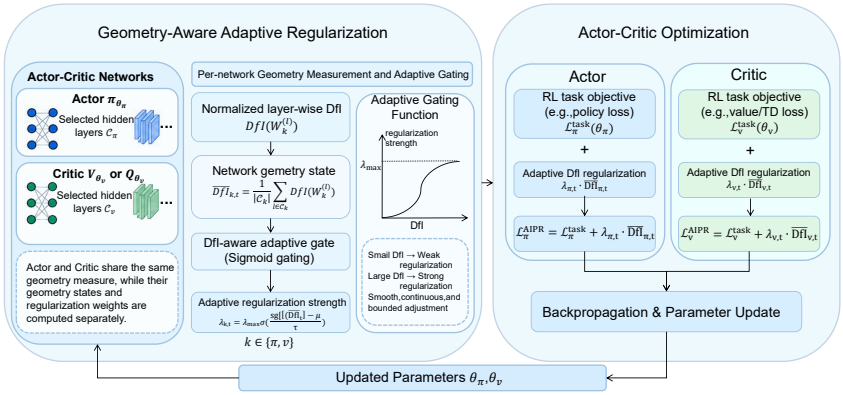


Figure 1: Overall framework of AIPR. AIPR independently characterizes the weight geometry of selected hidden layers in the Actor and Critic networks and dynamically adjusts the regularization strength according to the network–level DfI. The adaptive DfI regularization terms are jointly optimized with the Actor and Critic objectives through gradient descent, enabling smooth preservation of near–isometric weight structures without introducing discrete parameter reinitialization.

3 Method

During long–term deep reinforcement learning training, the continuous evolution of network parameters may gradually drive weight matrices away from their desired near–isometric configurations, resulting in geometric degradation and reduced plasticity. Unlike existing approaches (Han et al., 2026) that rely on discrete parameter reinitialization, we propose Adaptive Isometric Plasticity Regularization (AIPR). AIPR reformulates DfI from a discrete reinitialization criterion into a continuous and differentiable geometric regularization signal, enabling weight geometric preservation to be directly integrated into the gradient–based optimization process of the Actor–Critic framework without parameter reinitialization. The overall framework of AIPR is illustrated in Figure 1.

Specifically, AIPR integrates DfI–based geometric regularization terms into the original Actor and Critic objectives, allowing task optimization and weight geometric preservation to be jointly optimized within a unified differentiable framework. To overcome the limitation of fixed regularization intensity, AIPR further introduces a DfI–aware state-dependent soft gating mechanism that dynamically modulates the regularization strength according to the evolving geometric states of the Actor and Critic networks. When weights remain close to the desired near–isometric structure, the mechanism relaxes geometric constraints to avoid unnecessary interference with task–driven updates. In contrast, stronger regularization is applied when geometric degradation increases, providing additional constraints to preserve favorable weight structures. Consequently, AIPR adaptively balances geometric preservation and task optimization while retaining the original policy–value coupling in the Actor–Critic architecture.

3.1 Weight Geometry Preservation

Reinforcement learning problems are commonly formulated as Markov decision processes (MDPs), defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively, P represents the state transition probability, r is the immediate reward function, and $\gamma \in (0, 1)$ is the discount factor. In the Actor–Critic framework, the policy network $\pi_{\theta_\pi}(a|s)$ models the action distribution conditioned on the current state, while the value network $V_{\theta_v}(s)$ and action–value network $Q_{\theta_v}(s, a)$ estimate the expected return of states and state–action pairs, respectively. To accommodate different Actor–Critic algorithms, their task–specific optimization objectives can be generally formulated as

$$\mathcal{L}_{\mathrm{RL}}(\theta_\pi, \theta_v) = \mathcal{L}_{\mathrm{actor}}(\theta_\pi) + \beta_v \mathcal{L}_{\mathrm{critic}}(\theta_v) \tag{1}$$

where $\mathcal{L}_{\mathrm{actor}}$ denotes the policy optimization objective, $\mathcal{L}_{\mathrm{critic}}$ represents the value estimation objective, and β_v is the coefficient weighting the value loss. This standard optimization framework is solely designed to maximize the expected return of the task and does not impose explicit constraints on the geometric properties of network parameters.

During long–term training, the evolution of the policy continuously changes the agent's state–action sampling distribution. Under the combined effects of non–stationary data distributions, bootstrapped learning objectives, and iterative parameter updates, network weight structures may gradually drift away from desirable near–isometric configurations. However, conventional reinforcement learning objectives focus solely on task performance optimization and do not explicitly regulate the geometric properties of network parameters. Therefore, we reformulate the reinforcement learning objective as a joint optimization problem that simultaneously considers task optimization and weight geometric preservation

$$\min_{\theta_\pi, \theta_v} \mathcal{L}_{\mathrm{total}}(\theta_\pi, \theta_v) = \mathcal{L}_{\mathrm{RL}}(\theta_\pi, \theta_v) + \Omega_{\mathrm{reg}}(\theta_\pi, \theta_v) \tag{2}$$

where Ω_{reg} denotes the weight geometric regularization term, which constrains the deviation of network parameters from the desired near–isometric structure. Unlike existing approaches that restore weight geometry through discrete parameter reinitialization, AIPR reformulates DfI as a continuous and differentiable geometric regularization signal, allowing geometric constraints to be directly incorporated into the gradient–based optimization process. AIPR introduces no additional auxiliary networks and only relies on the existing parameters of the Actor and Critic networks to characterize their geometric states and impose corresponding constraints during gradient updates. Furthermore, considering that the Actor and Critic may exhibit heterogeneous geometric evolution patterns, AIPR independently quantifies their geometric degradation and adaptively adjusts the regularization strength through a state–dependent gating mechanism. This design unifies task optimization and weight geometric preservation within a continuous gradient-based optimization framework while maintaining the original Actor–Critic learning structure.

3.2 Continuous Geometric Regularization Based on DfI

To quantify the deviation of weight matrices from desirable near–isometric structures, we adopt DfI as a measure of weight geometric degradation. Unlike diagnostic metrics that depend on input samples or activation statistics, DfI is determined solely by network parameters and has a differentiable formulation, making it directly applicable as a geometric regularization term within end–to–end gradient–based optimization. Previous studies (Chung et al., 2024; Lewandowski et al., 2025) have shown that weight matrices with near–isometric or well–conditioned spectral structures can maintain a more balanced singular value spectrum and improve optimization stability. Therefore, we utilize DfI to quantify the deviation of weight matrices from near–isometric configurations and construct a continuous geometric regularization signal. Specifically, consider an arbitrary weight matrix $W \in \mathbb{R}^{m \times n}$, where m and n denote the output and input dimensions, respectively. Since a non–square matrix cannot simultaneously satisfy both $W^\top W = I_n$ and $WW^\top = I_m$, we adopt column–wise or row–wise isometric constraints depending on the matrix dimensionality and define the normalized DfI as follows

$$\mathrm{DfI}(W) = \begin{cases} \frac{1}{n} \left\| W^\top W - I_n \right\|_F^2, & m \geq n, \\ \frac{1}{m} \left\| WW^\top - I_m \right\|_F^2, & m < n. \end{cases} \tag{3}$$

where I_n and I_m denote identity matrices of the corresponding dimensions, and $\| \cdot \|_F$ represents the Frobenius norm. DfI regularization promotes orthonormality of weight matrices, encouraging

approximate isometry in non-dimensionality-reducing settings ($m \geq n$) and subspace isometry in dimensionality-reducing settings ($m < n$). A detailed analysis of its geometric interpretation under different dimensional configurations is provided in Appendix A.1. Importantly, DfI is continuously differentiable with respect to the weight matrix, allowing it to be directly optimized via backpropagation. The explicit gradient expressions are provided in Appendix A.2.

To theoretically justify how DfI regularization maintains the geometric structure and stabilizes training, we summarize its core theoretical properties from the perspectives of static weight spectra and dynamic feature propagation:

Property 1(Singular Value Bounds and Stable Rank Preservation): Under the DfI constraint, the singular values of the weight matrix W are bounded within a tight neighborhood around 1. This provides a non-trivial lower bound for the stable rank, preventing rank collapse during optimization.

These results characterize the geometric regulation effect of DfI in the parameter space. However, maintaining a stable weight spectrum does not directly characterize whether feature representations remain stable during forward propagation. We therefore further analyze the influence of DfI on the feature covariance structure across layers.

Property 2 (Pre-activation Covariance Non-degeneracy): Assuming non-degenerate input covariance and a sufficiently small DfI (e.g, $\mathrm{DfI}(W) < 1/m$), DfI regularization guarantees a strictly positive lower bound on the minimum eigenvalue of the pre-activation covariance matrix for a single layer, thereby mitigating feature space collapse and covariance degeneration.

Due to space limitations, the formal mathematical statements and detailed proofs for Properties 1 and 2 are provided in Appendix B.

3.3 DfI-aware Adaptive Regularization

Existing orthogonal and spectral regularization approaches typically employ predefined fixed regularization coefficients to continuously constrain weight geometry with a constant intensity during training (Bjorck et al., 2021; Lewandowski et al., 2025). Although these methods can improve the spectral properties of weight matrices and enhance optimization stability, they cannot adapt the constraint intensity to the dynamically changing geometric condition of the network. When weight matrices already maintain favorable near-isometric structures, excessively strong geometric constraints may interfere with task-driven optimization and limit adaptation to the current data distribution. Conversely, when severe geometric degradation occurs, weak fixed-strength constraints may provide insufficient correction to prevent further distortion of the singular value spectrum. Therefore, developing a mechanism that adaptively adjusts regularization strength according to the current geometric state of network parameters remains an open challenge. Based on the network-level geometric state $\overline{\mathrm{DfI}}_t$ defined in Section 3.2, we propose a DfI-aware adaptive gating mechanism that dynamically generates regularization coefficients according to the evolving weight geometry

$$\lambda_t = \lambda_{\max} \sigma \left(\frac{\mathrm{sg}\left[\overline{\mathrm{DfI}}_t\right] - \mu}{\tau} \right) \tag{4}$$

where $\lambda_{\max}$ represents the maximum regularization strength, μ defines the gating center, $\tau > 0$ controls the temperature of the sigmoid transition, and $\sigma(\cdot)$ denotes the sigmoid function. The parameter μ specifies the reference location at which the regularization strength gradually shifts from weak to strong, whereas τ determines the smoothness of this transition. Different from approaches that learn gating parameters through additional optimization, AIPR achieves adaptivity by dynamically modulating the regularization coefficient λ_t as a function of the current geometric state $\overline{\mathrm{DfI}}_t$. As $\overline{\mathrm{DfI}}_t$ evolves during training, λ_t is continuously adjusted through the sigmoid mapping. Consequently, crossing the gating center μ induces only a continuous adjustment of the regularization strength, so moderate inaccuracies in μ lead to gradual changes in regularization behavior, whereas FIRE triggers an abrupt parameter update upon recovery.

The stop-gradient operation in Eq. 4 is applied only to the gating pathway. During the t-th back propagation step, $\overline{\mathrm{DfI}}_t$ is still used to compute the geometric regularization gradient, while the generated λ_t is treated as a constant. Therefore, the optimization process does not introduce an additional gradient term $\nabla_\theta \lambda_t$. Let the reinforcement learning task gradient and the geometric regularization

gradient be denoted as $g_{\mathrm{RL}} = \nabla_\theta \mathcal{L}_{\mathrm{RL}}(\theta_t)$ and $g_{\mathrm{reg}} = \nabla_\theta \overline{\mathrm{DfI}}_t$, respectively. The parameter update rule under adaptive geometric regularization is given by

$$\theta_{t+1} = \theta_t - \eta \left(g_{\mathrm{RL},t} + \lambda_t g_{\mathrm{reg},t} \right) \tag{5}$$

where $\eta > 0$ denotes the learning rate.

It should be noted that the stop-gradient operation does not imply that λ_t remains constant across different optimization steps. Although the gating coefficient in AIPR dynamically evolves throughout training, the Sigmoid-based mapping prevents unbounded variations in the regularization strength. The Lipschitz continuity analysis of this mapping is provided in Appendix C.

To theoretically characterize how DfI-aware adaptive gating regulates geometric regularization while limiting its interference with task optimization, we summarize its key theoretical properties from the perspectives of adaptive regularization and first-order task-gradient perturbation:

Property 3 (Suppression of Task Gradient Interference by Adaptive Gating): The adaptive gating mechanism dynamically adjusts the regularization strength according to the current geometric state. When the average DfI remains below the gating center, it yields a decreasing upper bound on the worst-case first-order task-gradient perturbation as the deviation from the gating center increases.

This result characterizes how AIPR avoids unnecessary optimization perturbations through geometry-aware regularization. However, it does not directly quantify the advantage over fixed-strength geometric regularization. We therefore further compare the adaptive mechanism with a fixed-strength DfI regularizer.

Corollary 3.1 (First-Order Perturbation Reduction over Fixed-Strength DfI Regularization): Under the same parameter conditions and maximum regularization strength, AIPR yields a tighter upper bound on the first-order task-gradient perturbation than fixed-strength DfI regularization when the average DfI remains below the gating center. Since both methods share the same task and regularization gradients at a given parameter state, the ratio of their worst-case first-order perturbation bounds is exactly $\lambda_t / \lambda_{\max}$.

The formal statements and detailed proofs of Property 3 and Corollary 3.1 are provided in Appendix D.

3.4 Decoupled Adaptive Regularization for Actor–Critic Networks

In the Actor–Critic architecture, the Actor and Critic networks play different roles in the optimization process, the two networks may exhibit different parameter update patterns and heterogeneous evolution of weight geometry during training. Therefore, applying a unified geometric state estimation and regularization strength to both networks may fail to adapt the constraint intensity to their respective geometric conditions. The geometric states of the two networks are independently characterized using their respective network-level DfI values, and individual adaptive regularization strengths are determined through the DfI-aware gating mechanism.

Let $\overline{\mathrm{DfI}}_{\pi,t}$ and $\overline{\mathrm{DfI}}_{v,t}$ denote the network-level geometric states of the Actor and Critic at the t-th parameter update, respectively. Following Eq. 4, the corresponding adaptive regularization coefficients are defined as $\lambda_{\pi,t} = \lambda^{\pi}_{\max} \sigma \left(\frac{\mathrm{sg}\left[\overline{\mathrm{DfI}}_{\pi,t}\right] - \mu_\pi}{\tau_\pi} \right)$, $\lambda_{v,t} = \lambda^{v}_{\max} \sigma \left(\frac{\mathrm{sg}\left[\overline{\mathrm{DfI}}_{v,t}\right] - \mu_v}{\tau_v} \right)$.

Property 3 is established under a local parameter block-wise analysis within the Actor-Critic framework. Specifically, when analyzing the single-step update of one network, the parameters of the other network are considered fixed, and only the corresponding parameter block participates in the optimization. This allows both networks to independently regulate the influence of geometric regularization on task optimization according to their respective weight geometric states. The detailed mathematical proofs are deferred to Appendix E. The AIPR optimization objectives for the Actor and Critic are defined as follows

$$\mathcal{L}^{\mathrm{AIPR}}_{\pi} = \mathcal{L}_{\pi} + \lambda_{\pi,t} \overline{\mathrm{DfI}}_{\pi,t} \tag{6}$$

$$\mathcal{L}^{\mathrm{AIPR}}_{v} = \beta_v \mathcal{L}_{v} + \lambda_{v,t} \overline{\mathrm{DfI}}_{v,t} \tag{7}$$

For notational simplicity, the regularization objectives of the Actor and Critic networks are equivalently expressed in the following unified form

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\pi} + \beta_v \mathcal{L}_{v} + \lambda_{\pi,t} \overline{\mathrm{DfI}}_{\pi,t} + \lambda_{v,t} \overline{\mathrm{DfI}}_{v,t} \tag{8}$$

To illustrate how AIPR integrates into the Actor–Critic framework, Algorithm 1 summarizes the overall optimization procedure (see Appendix F). The additional computational cost introduced by AIPR is dominated by the DfI computation over the selected layers and is given by $\mathcal{O}\left(\sum_{l\in\mathcal{C}} m_l n_l \min(m_l, n_l)\right)$, where $\mathcal{C}$ denotes the set of layers to which AIPR is applied, and m_l and n_l denote the dimensions of the weight matrix at layer l, respectively. Detailed complexity analysis is provided in Appendix F.

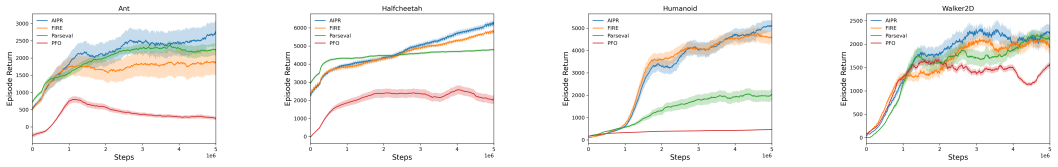
4 Experiments

4.1 Experimental Setup

To systematically evaluate AIPR, we conduct experiments on three representative reinforcement learning benchmarks, MuJoCo (Todorov et al., 2012), Robosuite (Zhu et al., 2020), and CARL (Benjamins et al., 2021), covering standard continuous control, robotic manipulation, and context-dependent environments, respectively. This diverse set of benchmarks allows us to evaluate AIPR across varying task dynamics, environment complexities, and learning conditions. We compare AIPR against representative methods for mitigating weight geometric degradation and plasticity loss in deep reinforcement learning, including FIRE (Han et al., 2026), Parseval regularization (Chung et al., 2024), and PFO (Moalla et al., 2024). Detailed descriptions of the benchmarks, baselines, implementation details, and hyperparameter settings are provided in Appendix G.

4.2 Main Experiment

We adopt PPO as the base reinforcement learning algorithm and incorporate AIPR and all comparison methods into the PPO framework for evaluation. We evaluate the performance of AIPR on the MuJoCo continuous control benchmark and compare it with FIRE, Parseval regularization, and PFO. The detailed experimental results are presented in Figure 2.



(a) Ant (b) HalfCheetah (c) Humanoid (d) Walker2D

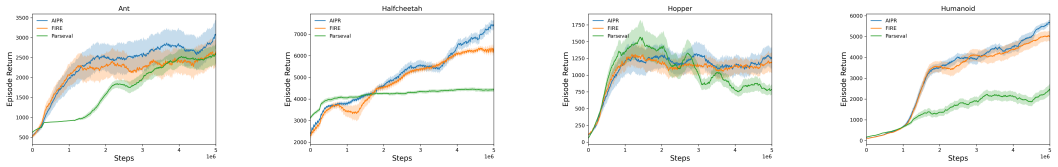
Figure 2: Learning curves of AIPR and baselines on MuJoCo continuous control tasks.

As shown in Figure 2, AIPR consistently achieves competitive performance across the MuJoCo tasks. On Ant and HalfCheetah, AIPR continues improving while FIRE and Parseval reach earlier plateaus, suggesting that fixed constraints or discrete interventions may become less compatible with evolving network states. In contrast, AIPR adaptively adjusts regularization strength based on DfI, preserving optimization flexibility while maintaining favorable weight geometry. On Walker2D, AIPR achieves the best final performance and maintains stable returns, whereas Parseval suffers from longer plateaus and fluctuations. On Humanoid, AIPR and FIRE show similar trends, but AIPR achieves higher final returns, while Parseval and PFO provide limited improvements. Overall, AIPR balances representation preservation and optimization flexibility through DfI-aware adaptive regularization. By dynamically adjusting constraint strength according to evolving weight geometry, AIPR avoids excessive restrictions and achieves stronger performance across diverse MuJoCo tasks. We further evaluate AIPR on CARL to examine its robustness under contextual variations, detailed results are provided in Appendix H.

4.3 Cross-Algorithm Applicability

To evaluate the generality of AIPR beyond PPO, we extend it to the off-policy SAC framework and conduct experiments on MuJoCo continuous control tasks. Unlike PPO, SAC uses experience replay and off-policy updates, providing a complementary setting with different optimization dynamics. Since PFO is designed for PPO's on-policy update mechanism and cannot be directly applied

to SAC, it is not considered in this comparison. In contrast, AIPR regularizes network weight geometry independently of the optimization procedure, allowing direct integration with SAC. AIPR is independently applied to the policy and online Q-networks, while target Q-networks are excluded from regularization. Results are shown in Figure 3.



(a) Ant (b) HalfCheetah (c) Hopper (d) Humanoid

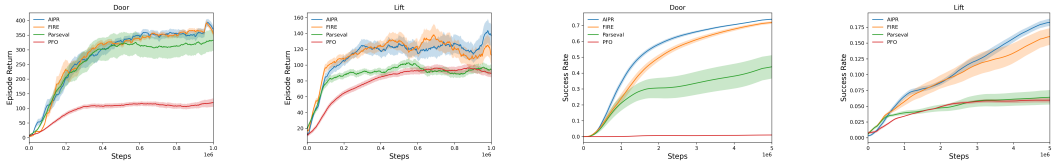
Figure 3: Learning curves of AIPR and baselines on MuJoCo SAC tasks.

As shown in Figure 3, AIPR achieves competitive performance across MuJoCo tasks under SAC, with clear advantages in later training stages. On Ant, AIPR and FIRE show similar early trends, but AIPR gradually surpasses FIRE and achieves the highest final return, while Parseval converges more slowly with lower final performance. These results suggest that fixed geometric constraints or discrete interventions may be less adaptive to evolving network states under off-policy optimization with replay-induced distribution shifts. Similar advantages are observed on HalfCheetah, Humanoid and Hopper. These results demonstrate the effectiveness of DfI-aware adaptive regularization under changing optimization dynamics.

Extending AIPR from PPO to SAC demonstrates that its effectiveness is not limited to on-policy optimization, but generalizes to off-policy Actor–Critic frameworks with experience replay. By adaptively balancing geometric constraint and parameter flexibility, AIPR provides consistent long-horizon learning improvements across diverse continuous control tasks.

4.4 Complex Robotic Manipulation Tasks

To further evaluate the applicability of AIPR in complex robotic manipulation scenarios, we conduct experiments on the Robosuite Door and Lift tasks. Unlike standard continuous control benchmarks, robotic manipulation tasks require both reward optimization and successful task completion under explicit objectives. Therefore, we use Success Rate as the primary metric and report Episode Return as a complementary measure of policy performance. Results are shown in Figure 4.



(a) Door (b) Lift (c) Door (success rate) (d) Lift (success rate)

Figure 4: Performance comparison of AIPR and baselines on Robosuite tasks. The first two panels show Episode Return, while the last two present Success Rate.

As shown in Figure 4, AIPR achieves strong final success rates on both Door and Lift tasks. On Door, AIPR continuously improves and surpasses FIRE during later training stages, achieving the best task completion performance, while Parseval and PFO obtain lower success rates. Although AIPR and FIRE show similar Episode Return trends, their Success Rate differs significantly, indicating that complex manipulation requires not only reward optimization but also effective exploration and precise control. By dynamically adjusting regularization strength through DfI-aware gating, AIPR maintains geometric stability while preserving optimization flexibility. Similar improvements are observed on Lift, where AIPR achieves stable success-rate growth and higher final returns than competing methods. The Robosuite experiments demonstrate that AIPR generalizes beyond standard continuous control tasks to complex robotic manipulation scenarios. By balancing

weight geometric preservation and task-driven optimization flexibility through DfI-aware adaptive regulation, AIPR enables more effective long-term policy refinement and improves task completion performance.

4.5 Ablation Experiments

To investigate the contribution of each key component in AIPR, we design four variants along two orthogonal dimensions, the geometric constraint mechanism (Discrete vs. Continuous) and the regularization strength adaptation strategy (Fixed vs. Adaptive). The resulting variants include Discrete+Fixed DfI, Discrete+Adaptive DfI, Continuous+Fixed DfI, and AIPR. We evaluate these variants on 10 tasks spanning MuJoCo, Robosuite, and CARL.

Table 1: Ablation study of key AIPR components. Results denote mean $\pm$ standard deviation of episode return over the final 10% evaluation period across 10 random seeds. Best performance is highlighted in bold.

Task	Environment	Discrete+Fixed DfI	Discrete+Adaptive DfI	Continuous+Fixed DfI	AIPR
MuJoCo	Ant	1924.10 ± 972.27	-26.96 ± 6.33	-19.88 ± 0.76	$\mathbf{2486.45 \pm 498.35}$
	HalfCheetah	5490.81 ± 153.05	1541.00 ± 876.99	20.06 ± 83.99	$\mathbf{5890.76 \pm 432.78}$
	Hopper	$\mathbf{1311.60 \pm 114.89}$	1087.99 ± 98.66	678.81 ± 213.52	1204.86 ± 57.79
	Humanoid	4651.99 ± 193.29	362.53 ± 11.62	375.22 ± 10.17	$\mathbf{5006.78 \pm 208.35}$
Robosuite	Door	359.65 ± 8.98	64.17 ± 2.54	77.33 ± 4.74	$\mathbf{361.44 \pm 19.61}$
	Lift	112.45 ± 4.85	56.12 ± 1.16	64.83 ± 8.65	$\mathbf{121.77 \pm 5.51}$
	Nut Assembly	$\mathbf{29.22 \pm 11.49}$	3.71 ± 0.25	4.20 ± 0.14	23.91 ± 10.90
	Stack	16.62 ± 4.18	15.90 ± 0.87	16.45 ± 0.09	$\mathbf{26.87 \pm 2.57}$
CARL	Acrobot	478.35 ± 12.75	473.11 ± 0.60	480.31 ± 5.50	$\mathbf{486.13 \pm 4.89}$
	MountainCarContinuous	482.19 ± 6.77	474.89 ± 19.31	482.63 ± 10.17	$\mathbf{487.90 \pm 5.92}$

As shown in Table 1, AIPR achieves the highest average return on 8 out of 10 tasks. Compared with Continuous + Fixed DfI, AIPR consistently improves performance, especially on Ant, HalfCheetah, and Humanoid. This indicates that fixed regularization strength may impose unnecessary constraints during training, whereas DfI-aware gating adaptively adjusts the regularization intensity according to the current geometric state, preserving optimization flexibility while maintaining geometric stability. However, Discrete + Adaptive DfI performs worse than Discrete + Fixed DfI on most tasks, suggesting that adaptive modulation alone is insufficient without a suitable continuous constraint mechanism. Moreover, the comparison between Discrete + Fixed DfI and Continuous + Fixed DfI shows that continuous regularization alone cannot guarantee improvements when the constraint strength is fixed.

The ablation results demonstrate a synergistic effect between the two components. Continuous regularization preserves favorable weight geometry, while adaptive modulation prevents unnecessary restrictions during training. Together, they enable AIPR to maintain a better balance between geometric stability and task-driven optimization, leading to consistent improvements across diverse tasks.

5 Conclusion

This paper proposes Adaptive Isometric Plasticity Regularization (AIPR), a continuous geometric regularization framework for mitigating plasticity degradation in deep reinforcement learning. By reformulating DfI from a discrete intervention criterion into a continuously differentiable weight geometric regularizer and introducing a DfI-aware adaptive gating mechanism, AIPR dynamically adjusts the regularization strength of the Actor and Critic networks according to their evolving geometric states. Extensive experiments on MuJoCo, Robosuite, and CARL demonstrate that AIPR achieves competitive performance and consistent improvements across a wide range of tasks. Moreover, its extension from PPO to SAC verifies that AIPR is not limited to a specific Actor–Critic optimization paradigm. Overall, AIPR provides an effective and general framework for continuously preserving weight geometry while maintaining task-driven optimization flexibility in reinforcement learning. Future work will investigate the applicability of AIPR to high-dimensional visual reinforcement learning tasks and broader classes of reinforcement learning algorithms. We will also explore more automatic strategies for determining adaptive gating parameters.

AI USAGE STATEMENT

In this research, generative AI tools were used for manuscript language polishing. Generative AI tools were not used for research idea design, experimental planning, result analysis or formulation of core claims of this paper. Generative AI was not applied for dataset construction and figure plotting in this work. Additionally, generative AI assisted with grammar correction and sentence readability optimization. We have reviewed all AI-assisted work. All AI-polished manuscript text was reviewed and edited paragraph by paragraph. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We guarantee the reproducibility of all experimental results. The code, training and inference scripts, hyperparameter configurations, and environment setup scripts are available at URL:https://anonymous.4open.science/r/AIPR-4BE7. We utilize only publicly available simulation platforms. During the double-blind review phase, we will submit a comprehensive reproduction package for the committee's verification. Upon acceptance, the relevant materials will be hosted in a permanent open-source repository. We take full responsibility for the authenticity of the experimental results.

REFERENCES

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, David Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Michał Wołczyk, Mateusz Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 28496–28510, 2021.

Haoran Tang, Joaquin Obando-Ceron, Pablo S. Castro, Aaron Courville, and Georgia Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 58883–58904, 2025.

Ziming Wu, Haoran Tang, Yuhang Ma, Jiahao Liu, Yifan Zheng, and Jianye Hao. The rank and gradient lost in non-stationarity: Sample weight decay for mitigating plasticity loss in reinforcement learning. In *International Conference on Learning Representations*, 2026.

Haoran Zhou, Zhiwei Zhuang, and Deqing Wang. Stay hungry, keep learning: Sustainable plasticity for deep reinforcement learning. In *Proceedings of the 42nd International Conference on Learning Representations*, volume 267, pages 79644–79672, 2025.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingyang Lan, Parsa Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

Ilwoo Han, Seonghwan Park, Seungwoo Oh, Donghyun Kim, Haejun Lee, and K.-J. Kim. FIRE: Frobenius-isometry reinitialization for balancing the stability–plasticity tradeoff. In *International Conference on Learning Representations*, 2026.

William Chung, Lamia Cherif, David Meger, and Doina Precup. Parseval regularization for continual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pages 127937–127967, 2024.

Tobias Klein, Lukas Miklautz, Kamil Sidak, Claudia Plant, and Stefan Tschiatschek. Plasticity loss in deep reinforcement learning: A survey. *arXiv preprint arXiv:2411.04832*, 2024.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bruno A. Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 23190–23211, 2023.

Arthur Juliani and Jordan T. Ash. A study of plasticity loss in on-policy deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sami Moalla, Alessandro Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No representation, no trust: Connecting representation, collapse, and trust issues in PPO. In *Advances in Neural Information Processing Systems*, volume 38, pages 69652–69699, 2024.

Byungjun Shin, Jaehoon Oh, Hyojin Cho, and Se-Young Yun. DASH: Warm-starting neural network training in stationary settings without loss of plasticity. In *Advances in Neural Information Processing Systems*, volume 37, pages 43300–43340, 2024.

Joakim Bjorck, Carla P. Gomes, and Kilian Q. Weinberger. Towards deeper deep reinforcement learning with spectral normalization. In *Advances in Neural Information Processing Systems*, volume 35, pages 8242–8255, 2021.

Mohammad Nauman, Michał Bortkiewicz, Piotr Miłos, Tomasz Trzciński, Maciej Ostaszewski, and Marek Cygan. Overestimation, overfitting, and plasticity in actor-critic: The bitter lesson of reinforcement learning. In *International Conference on Machine Learning*, 2024.

Adam Lewandowski, Michał Bortkiewicz, Shankar Kumar, András György, Dale Schuurmans, Maciej Ostaszewski, and Marlos C. Machado. Learning continually by spectral regularization. In *International Conference on Learning Representations*, 2025.

Sébastien Garcin, Tyler McInroe, Pablo S. Castro, Prakash Panangaden, Christopher G. Lucas, David Abel, and Stefano V. Albrecht. Studying the interplay between the actor and critic representations in reinforcement learning. In *International Conference on Learning Representations*, 2025.

Roberta Raileanu and Rob Fergus. Decoupling value and policy for generalization in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2021.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

Yuke Zhu, Josiah Wong, Animesh Mandlekar, Roberto Martín-Martín, Ankur Joshi, Kuan Lin, Abhishek Maddukuri, Shubham Nasiriany, and Yuke Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

Christoph Benjamins, André Eimer, Frederik Schubert, Jan Biedenkapp, Bastian Rosenhahn, Frank Hutter, and Marius Lindauer. CARL: A benchmark for contextual and adaptive reinforcement learning. In *Workshop on Ecological Theory of Reinforcement Learning, NeurIPS*, 2021.

A Additional Analysis of DfI

A.1 Geometric Interpretation of DfI

We first provide additional geometric interpretation of the DfI definition in Eq. 3. When $m \geq n$, DfI encourages the column vectors of the weight matrix to remain approximately orthonormal by constraining $W^\top W$ toward the identity matrix I_n, thereby preserving approximate isometry in the input space. When $m < n$, DfI instead constrains $WW^\top$ toward the identity matrix I_m, encouraging the row vectors of the weight matrix to maintain approximate orthogonality. Since the output dimension is smaller than the input dimension in this case, the transformation cannot preserve distances along all directions of the original input space. Nevertheless, it can maintain approximate isometry within the effective subspace spanned by the row space of the weight matrix.

A.2 Gradient Derivation of DfI

The geometric implication of DfI can be further characterized through the singular value spectrum of the weight matrix. Let $r = \min(m, n)$. When $\mathrm{DfI}(W) = 0$, the r non–zero singular values of W are all equal to one. When $\mathrm{DfI}(W)$ remains small, these singular values are constrained to stay close to 1, thereby limiting excessive amplification or attenuation of feature representations

during inter-layer propagation. During long-term reinforcement learning training, the interaction between non-stationary sampling distributions, bootstrapped objectives, and continual parameter updates can gradually distort the singular value spectrum of weight matrices, causing an increase in geometric deviation. Consequently, larger $\mathrm{DfI}(W)$ reflects a greater departure from near-isometric transformations and potentially unfavorable signal propagation and gradient flow characteristics. Therefore, AIPR incorporates DfI as a differentiable geometric regularization term to continuously preserve desirable weight geometry throughout training.

We next extend DfI from individual weight matrices to characterize the geometric state of the entire network. Let $\mathcal{C}$ denote the set of layers selected for geometric regularization. In this work, $\mathcal{C}$ includes only the linear layers within the hidden representation modules, while excluding task output layers, bias parameters, and normalization parameters. For the Actor and Critic networks, the corresponding regularized layer sets are denoted as $\mathcal{C}_\pi$ and $\mathcal{C}_v$, respectively.AIPR computes the normalized DfI of the selected hidden-layer weight matrices and producing a network–level geometric state. Output layers are excluded from regularization to avoid imposing unnecessary geometric constraints on task-specific mappings, such as action distributions and value predictions. Unless otherwise specified, all experiments follow this layer selection strategy. The network-level weight geometric state is therefore defined as the average DfI across all regularized layers

$$\overline{\mathrm{DfI}}_t = \frac{1}{|\mathcal{C}|} \sum_{l \in \mathcal{C}} \mathrm{DfI}\left(W_t^{(l)}\right) \tag{9}$$

where $W_t^{(l)}$ denotes the weight matrix of the l-th regularized layer at training step t. In AIPR, $\overline{\mathrm{DfI}}_t$ is used as the current geometric state of network parameters to quantify the deviation of the overall weight structure from the desired near-isometric configuration during training. Since DfI is continuously differentiable, its gradients can be directly propagated through backpropagation and incorporated into network parameter updates. For a weight matrix $W \in \mathbb{R}^{m \times n}$, the gradient of DfI can be generally formulated as follows

$$\nabla_W \mathrm{DfI}(W) = \begin{cases} \frac{4}{n} W(W^\top W - I_n), & m \geq n, \\ \frac{4}{m}(WW^\top - I_m)W, & m < n. \end{cases} \tag{10}$$

Therefore, DfI serves as a differentiable geometric regularization term that can be directly incorporated into network parameter updates through backpropagation, without requiring additional projection modules or non-differentiable interventions.

Unlike prior work (Han et al., 2026), which employs DfI as a criterion for triggering discrete parameter reinitialization, AIPR reformulates DfI as a continuous geometric regularizer that directly participates in gradient-based optimization. AIPR jointly optimizes the DfI regularization term with the Actor and Critic task objectives throughout training, thereby replacing discrete recovery with continuous geometric regulation.

B Theoretical Properties of DfI

B.1 Property 1: Singular Value Bounds and Stable Rank Preservation

Let $r = \min(m, n)$, and denote the singular values of the weight matrix W as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. According to the normalized definition in Eq. 3, when $m \geq n$, we have $r = n$ and DfI adopts a column-wise isometric constraint. When $m < n$, we have $r = m$ and DfI adopts a row-wise isometric constraint. Therefore, regardless of the dimensional configuration of the weight matrix, Eq. 3 can be uniformly expressed as

$$\mathrm{DfI}(W) = \frac{1}{r} \sum_{i=1}^{r} \left(\sigma_i^2 - 1\right)^2 \tag{11}$$

According to Eq. 11, for any $i \in \{1, \cdots, r\}$, the squared deviation of an individual singular value from one is upper-bounded by the total deviation of all singular values $(\sigma_i^2 - 1)^2 \leq \sum_{j=1}^{r} (\sigma_j^2 - 1)^2 = r \cdot \mathrm{DfI}(W)$. Therefore:

$$\left|\sigma_i^2 - 1\right| \leq \sqrt{r\,\mathrm{DfI}(W)} \tag{12}$$

The non-zero singular values of the weight matrix are bounded as follows

$$1 - \sqrt{r\mathrm{DfI}(W)} \le \sigma_i^2 \le 1 + \sqrt{r\mathrm{DfI}(W)} \tag{13}$$

Eq. 13 shows that when DfI lies in the small–DfI regime, i.e., $\mathrm{DfI}(W) < \frac{1}{r}$, the lower bound $1 - \sqrt{r\mathrm{DfI}(W)} > 0$ remains strictly positive. Therefore, for all $i = 1, \ldots, r$, the singular values satisfy $\sigma_i^2 \ge 1 - \sqrt{r\,\mathrm{DfI}(W)}$, which guarantees that $\mathrm{rank}(W) = r$. This result indicates that, under the condition $\mathrm{DfI}(W) < \frac{1}{r}$, DfI not only constrains the deviation of singular values from one but also prevents rank collapse, ensuring that the weight matrix maintains its maximum attainable rank.

Furthermore, the stable rank of the weight matrix is defined as $\mathrm{srank}(W) = \frac{\|W\|_F^2}{\|W\|_2^2} = \frac{\sum_{i=1}^r \sigma_i^2}{\sigma_1^2}$. Let $\delta_i = \sigma_i^2 - 1$. Applying the Cauchy–Schwarz inequality, the numerator can be lower–bounded as $\sum_{i=1}^r \sigma_i^2 = r + \sum_{i=1}^r \delta_i \ge r - |\sum_{i=1}^r \delta_i| \ge r - \sqrt{r \sum_{i=1}^r \delta_i^2} = r\left(1 - \sqrt{\mathrm{DfI}(W)}\right)$. On the other hand, according to Eq. 13, the denominator satisfies the upper bound $\sigma_1^2 \le 1 + \sqrt{r\,\mathrm{DfI}(W)}$. Therefore, within the small–DfI regime, combining the lower bound of the numerator with the upper bound of the denominator yields

$$\mathrm{srank}(W) \ge \frac{r\left(1 - \sqrt{\mathrm{DfI}(W)}\right)}{1 + \sqrt{r\,\mathrm{DfI}(W)}} \tag{14}$$

Eq. 14 shows that, within the aforementioned small–DfI regime, the lower bound of the stable rank increases as DfI decreases. Specifically, when $\mathrm{DfI}(W) \to 0$, $\frac{r\left(1-\sqrt{\mathrm{DfI}(W)}\right)}{1+\sqrt{r\,\mathrm{DfI}(W)}} \to r$. In the ideal case where $\mathrm{DfI}(W) = 0$, all singular values satisfy $\sigma_1 = \cdots = \sigma_r = 1$, resulting in $\mathrm{srank}(W) = r$. Therefore, minimizing DfI effectively constrains the deviation of weight singular values from one and provides a theoretical basis for preserving a more balanced and well–conditioned singular value spectrum.

B.2 Property 2: Pre-activation Covariance Non-degeneracy

To further characterize the effect of DfI on feature propagation, consider a linear layer with input dimension n and output dimension m, where $m \le n$. Under this setting, Eq. 3 can be uniformly expressed as $\mathrm{DfI}(W) = \frac{1}{m}\left\|WW^\top - I_m\right\|_F^2 = \frac{1}{m}\sum_{i=1}^m (\sigma_i^2 - 1)^2$, where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_m \ge 0$ denote the singular values of W. Let $x \in \mathbb{R}^n$ be a zero-mean input random vector with covariance matrix $\Sigma_{\mathrm{in}} = \mathbb{E}\left[xx^\top\right]$. We assume that the input covariance is non–degenerate, i.e., $\lambda_{\min}(\Sigma_{\mathrm{in}}) > 0$, where Σ_{in} represents the population covariance of the input distribution rather than the empirical covariance estimated from a finite mini–batch. For the linear transformation $z = Wx$, the pre–activation covariance is given by $\Sigma_{\mathrm{pre}} = \mathbb{E}\left[zz^\top\right] = W\Sigma_{\mathrm{in}}W^\top$. For any unit vector $\boldsymbol{u} \in \mathbb{R}^m$ satisfying $\|\boldsymbol{u}\|_2 = 1$, since $\Sigma_{\mathrm{in}} \succeq \lambda_{\min}(\Sigma_{\mathrm{in}})\, I_n$, we have $\boldsymbol{u}^\top\Sigma_{\mathrm{pre}}\boldsymbol{u} = (W^\top\boldsymbol{u})^\top\Sigma_{\mathrm{in}}(W^\top\boldsymbol{u}) \ge \lambda_{\min}(\Sigma_{\mathrm{in}})\left\|W^\top\boldsymbol{u}\right\|_2^2 \ge \lambda_{\min}(\Sigma_{\mathrm{in}})\,\sigma_{\min}^2(W)$, where the last inequality follows from $\left\|W^\top\boldsymbol{u}\right\|_2^2 \ge \sigma_{\min}^2(W)\,\|\boldsymbol{u}\|_2^2 = \sigma_{\min}^2(W)$. Taking the minimum over all unit vectors yields

$$\lambda_{\min}(\Sigma_{\mathrm{pre}}) \ge \lambda_{\min}(\Sigma_{\mathrm{in}})\,\sigma_{\min}^2(W) \tag{15}$$

Furthermore, according to Eq. 13 in Appendix B.1, for the case $m \le n$ where $r = m$, the smallest singular value satisfies $\sigma_{\min}^2(W) \ge 1 - \sqrt{m\,\mathrm{DfI}(W)}$. When $\mathrm{DfI}(W) < \frac{1}{m}$, substituting this bound into Eq. 15 yields the following lower bound on the minimum eigenvalue of the pre–activation covariance matrix

$$\lambda_{\min}(\Sigma_{\mathrm{pre}}) \ge \lambda_{\min}(\Sigma_{\mathrm{in}})\left(1 - \sqrt{m\,\mathrm{DfI}(W)}\right) > 0 \tag{16}$$

As shown in Eq. 16, when $m \le n$, $\lambda_{\min}(\Sigma_{\mathrm{in}}) > 0$, and $\mathrm{DfI}(W) < \frac{1}{m}$, the DfI constraint guarantees a strictly positive lower bound for the minimum eigenvalue of the single–layer pre–activation covariance matrix. This result indicates that controlling weight geometry through DfI can alleviate covariance degeneration caused by geometric deviation of the weight matrix. However, this property only provides a conditional local guarantee for individual layers and does not imply that DfI can recover representation directions that have already been lost in upstream layers. When $m > n$,

the output covariance matrix necessarily contains zero eigenvalues, making positive definiteness unattainable in this setting. Therefore, we do not analyze covariance positive definiteness for this case, but instead rely on Property 1 to constrain the non–zero singular value structure of the weight matrix.

C Lipschitz Continuity of the Adaptive DfI Gate

According to Eq. 4, the adaptive regularization coefficient is defined as $\lambda_t = \lambda_{\max}\sigma\left(\frac{\mathrm{sg}\left[\overline{\mathrm{DfI}}_t\right]-\mu}{\tau}\right)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ denotes the sigmoid function. The derivative of the sigmoid function satisfies $\sigma'(z) = \sigma(z)\left(1-\sigma(z)\right) \leq \frac{1}{4}$, where the maximum value $\frac{1}{4}$ is attained when $\sigma(z) = \frac{1}{2}$. Therefore, according to Eq. 4, the gating function with respect to DfI is $\frac{\lambda_{\max}}{4\tau}$–Lipschitz continuous, which gives

$$\left|\lambda_{t+1} - \lambda_t\right| \leq \frac{\lambda_{\max}}{4\tau}\left|\overline{\mathrm{DfI}}_{t+1} - \overline{\mathrm{DfI}}_t\right| \tag{17}$$

Furthermore, assuming that $\overline{\mathrm{DfI}}(\theta)$ is locally L_{DfI}–Lipschitz continuous, the composite gating function satisfies

$$\left|\lambda_{t+1} - \lambda_t\right| \leq \frac{\lambda_{\max} L_{\mathrm{DfI}}}{4\tau}\left\|\theta_{t+1} - \theta_t\right\|_2 \tag{18}$$

This result indicates that the variation of the gating coefficient is controlled by the magnitude of parameter updates. Although the AIPR gating mechanism evolves dynamically throughout training, the Sigmoid–based mapping prevents abrupt or unbounded changes in the regularization strength. Consequently, AIPR avoids discontinuities in the optimization process. It should be emphasized that this analysis only characterizes local changes in the adaptive regularization process and does not provide a global convergence guarantee for the full non–convex reinforcement learning optimization problem.

D Theoretical Analysis of DfI-Aware Adaptive Gating

D.1 Property 3: Suppression of Task Gradient Interference by Adaptive Gating

Assume that, within a local neighborhood around the current parameter update, the reinforcement learning objective $\mathcal{L}_{\mathrm{RL}}(\theta)$ is differentiable and has a locally β–Lipschitz gradient (i.e., it is locally β–smooth). We note that although some reinforcement learning objectives, such as the clipped surrogate objective in PPO, are non–smooth at isolated points (e.g., clipping boundaries), the following analysis applies within the smooth regions and serves as a local first–order approximation of the optimization dynamics rather than a global characterization. According to the descent lemma, $\mathcal{L}_{\mathrm{RL}}(\theta_{t+1}) \leq \mathcal{L}_{\mathrm{RL}}(\theta_t) + \langle g_{\mathrm{RL}}, \theta_{t+1} - \theta_t\rangle + \frac{\beta}{2}\left\|\theta_{t+1} - \theta_t\right\|_2^2$. Substituting the parameter update rule in Eq. 5 into the above inequality yields

$$\mathcal{L}_{\mathrm{RL}}(\theta_{t+1}) - \mathcal{L}_{\mathrm{RL}}(\theta_t) \leq -\eta\|g_{\mathrm{RL}}\|_2^2 - \eta\lambda_t\left\langle g_{\mathrm{RL}}, g_{\mathrm{reg}}\right\rangle + \frac{\beta\eta^2}{2}\left\|g_{\mathrm{RL}} + \lambda_t g_{\mathrm{reg}}\right\|_2^2 \tag{19}$$

where $-\eta\lambda_t\left\langle g_{\mathrm{RL}}, g_{\mathrm{reg}}\right\rangle$ represents the first–order contribution of the geometric regularization term to the task objective. When $\left\langle g_{\mathrm{RL}}, g_{\mathrm{reg}}\right\rangle \geq 0$, this term is non–positive, indicating that the geometric regularization does not introduce adverse interference to task optimization in the first–order sense. Conversely, when $\left\langle g_{\mathrm{RL}}, g_{\mathrm{reg}}\right\rangle < 0$, the two gradients are negatively aligned, and the geometric regularization may cause unfavorable first–order perturbations to task optimization. To quantify this adverse effect, we define the task gradient interference at update step t as $\mathcal{I}_t = \eta\lambda_t\left[-\left\langle g_{\mathrm{RL}}, g_{\mathrm{reg}}\right\rangle\right]_+$, where $[x]_+ = \max(x, 0)$. By applying the Cauchy–Schwarz inequality, we obtain

$$\mathcal{I}_t \leq \eta\lambda_t\left\|g_{\mathrm{RL}}\right\|_2\left\|g_{\mathrm{reg}}\right\|_2 \tag{20}$$

Eq. 20 provides an upper bound on the worst–case first–order task gradient interference induced by the current geometric regularization term. The magnitude of this interference is directly controlled by the adaptive regularization coefficient λ_t.

Consider the case where the network–level average DfI is below the gating center, i.e., $\overline{\mathrm{DfI}}_t < \mu$. We define the distance from the current geometric state to the gating center as $\Delta_t = \mu - \overline{\mathrm{DfI}}_t > 0$. According to Eq. 4, the adaptive regularization coefficient can be rewritten as $\lambda_t = \lambda_{\max}\sigma\left(-\frac{\Delta_t}{\tau}\right) = \frac{\lambda_{\max}}{1+\exp(\Delta_t/\tau)}$. Since, for any $z > 0$, $\frac{1}{1+\exp(z)} \leq \exp(-z)$, we further obtain

$$\lambda_t \leq \lambda_{\max} e^{-\Delta_t/\tau} \tag{21}$$

Substituting Eq. 21 into Eq. 20, we obtain

$$\mathcal{I}_t \leq \eta\lambda_{\max} e^{-\Delta_t/\tau} \left\|g_{\mathrm{RL}}\right\|_2 \left\|g_{\mathrm{reg}}\right\|_2 \tag{22}$$

Eq. 22 shows that when the average DfI is below the gating center, the upper bound of the worst–case first–order task gradient interference in AIPR decreases exponentially with increasing Δ_t. As DfI approaches or exceeds the gating center, the adaptive regularization coefficient increases smoothly, thereby strengthening the geometric constraint. Furthermore, due to the boundedness of the sigmoid function, the regularization strength is always constrained within $0 < \lambda_t < \lambda_{\max}$. It should be emphasized that this result only characterizes a local upper bound of first–order perturbation under the worst–case gradient alignment and does not provide a global convergence guarantee for the full non–convex reinforcement learning optimization process.

D.2 Corollary 3.1: First-Order Perturbation Reduction

Consider a fixed–strength DfI regularization variant that adopts the same geometric regularization objective as AIPR but uses a constant coefficient $\lambda_{\mathrm{fix}} = \lambda_{\max}$ throughout training. At the same parameter state, both methods share identical task and regularization gradients, denoted by g_{RL} and g_{reg}, respectively. The corresponding worst-case first–order task-gradient perturbation bounds are

$$B_t^{\mathrm{fix}} = \eta\lambda_{\max} \left\|g_{\mathrm{RL}}\right\|_2 \left\|g_{\mathrm{reg}}\right\|_2 \tag{23}$$

Accordingly, the worst-case first–order interference bound induced by AIPR can be written as $B_t^{\mathrm{AIPR}} = \eta\lambda_t \left\|g_{\mathrm{RL}}\right\|_2 \left\|g_{\mathrm{reg}}\right\|_2$. When $\overline{\mathrm{DfI}}_t < \mu$ and $\Delta_t = \mu - \overline{\mathrm{DfI}}_t > 0$, applying Eq. 21 yields

$$B_t^{\mathrm{AIPR}} \leq e^{-\Delta_t/\tau} B_t^{\mathrm{fix}} \tag{24}$$

When both gradient norms are non-zero, this bound can be equivalently expressed as: $\frac{B_t^{\mathrm{AIPR}}}{B_t^{\mathrm{fix}}} = \frac{\lambda_t}{\lambda_{\max}} = \frac{1}{1+\exp(\Delta_t/\tau)} \leq e^{-\Delta_t/\tau}$

Corollary 3.1 demonstrates that, under identical parameter configurations, DfI regularization objectives, and maximum regularization strength, AIPR achieves a tighter upper bound on the first-order gradient perturbation than fixed-strength DfI regularization when the average DfI remains below the gating center. Moreover, this advantage becomes more pronounced as Δ_t increases. This result characterizes only the local single-step optimization behavior and does not imply a global convergence guarantee. The empirical differences between AIPR and other regularization methods, including Parseval regularization, are further validated through experiments.

E Proof of Decoupled Actor–Critic Regularization

To further characterize the effect of decoupled gating on Actor and Critic optimization, we denote the task gradients and geometric regularization gradients of the two networks as $g_\pi = \nabla_{\theta_\pi}\mathcal{L}_\pi$, $g_{\pi,\mathrm{reg}} = \nabla_{\theta_\pi}\overline{\mathrm{DfI}}_{\pi,t}$, and $g_v = \nabla_{\theta_v}\mathcal{L}_v$, $g_{v,\mathrm{reg}} = \nabla_{\theta_v}\overline{\mathrm{DfI}}_{v,t}$, respectively. When the network-level average DfI values of the Actor and Critic are below their corresponding gating centers, we define $\Delta_{\pi,t} = \mu_\pi - \overline{\mathrm{DfI}}_{\pi,t} > 0$, $\Delta_{v,t} = \mu_v - \overline{\mathrm{DfI}}_{v,t} > 0$. According to Appendix D.1, under the worst-case alignment between the task gradients and geometric regularization gradients, the first-order task gradient interference bounds of the Actor and Critic satisfy

$$\mathcal{I}_{\pi,t} \leq \eta\lambda_{\pi,t} \left\|g_\pi\right\|_2 \left\|g_{\pi,\mathrm{reg}}\right\|_2 \leq \eta\lambda_{\max}^{\pi} e^{-\Delta_{\pi,t}/\tau_\pi} \left\|g_\pi\right\|_2 \left\|g_{\pi,\mathrm{reg}}\right\|_2 \tag{25}$$

$$\mathcal{I}_{v,t} \leq \eta\lambda_{v,t} \left\|g_v\right\|_2 \left\|g_{v,\mathrm{reg}}\right\|_2 \leq \eta\lambda_{\max}^{v} e^{-\Delta_{v,t}/\tau_v} \left\|g_v\right\|_2 \left\|g_{v,\mathrm{reg}}\right\|_2 \tag{26}$$

Eq. 25 and 26 show that when the network-level average DfI of the Actor or Critic remains below its corresponding gating center, the worst-case task gradient interference is bounded by exponential terms involving $e^{-\Delta_{\pi,t}/\tau_\pi}$ and $e^{-\Delta_{v,t}/\tau_v}$, respectively. Therefore, the Actor and Critic networks can independently regulate the influence of geometric regularization on task optimization according to their respective weight geometric states. This decoupling only applies to the adaptive regularization mechanism and does not modify the original task objectives or the policy–value interaction structure of the Actor–Critic framework.

F Computational Complexity and Implementation Efficiency

To illustrate how AIPR integrates into the Actor–Critic framework, Algorithm 1 summarizes the overall optimization procedure. At each update step, AIPR computes the layer-wise and network-level DfI values for the selected hidden layers of the Actor and Critic networks. These estimates are then used to determine adaptive regularization coefficients via DfI-aware soft gating, yielding geometric regularization terms that are jointly optimized with the original task objectives.

Algorithm 1 AIPR Framework for Actor–Critic Algorithms

Require: Environment; Actor π_{θ_π}; Critic V_{θ_v}; Regularized layer sets $\mathcal{C}_\pi, \mathcal{C}_v$; Gating parameters $(\lambda^\pi_{\max}, \mu_\pi, \tau_\pi)$ and $(\lambda^v_{\max}, \mu_v, \tau_v)$.
Ensure: Optimized parameters θ_π, θ_v.

```
Initialize the Actor, Critic, and optimizer states;
while episode < Episode do
    Collect interaction data using the current policy;
    Construct the original task objectives according to the underlying Actor–Critic algorithm;
    for each gradient update step do
        Compute the original task losses L_{π,t} and L_{v,t};
        Compute DfI for each regularized layer: {DfI(W^(l)_{π,t})}_{l∈C_π},  {DfI(W^(l)_{v,t})}_{l∈C_v}
        Compute the network-level average DfI: overline{DfI}_{π,t}, overline{DfI}_{v,t};
        Compute adaptive coefficients via Eq. 4: λ_{π,t}, λ_{v,t}
        Construct the AIPR optimization objectives for the Actor and Critic: L^AIPR_π, L^AIPR_v
        Backpropagate and update θ_π and θ_v;
    end for
end while
```

AIPR introduces only lightweight geometric computations on the existing weight matrices of the Actor and Critic networks. Specifically, DfI is evaluated on selected hidden layers and incorporated into the task objective as a differentiable regularization term, without introducing additional auxiliary networks, target models, or input-dependent statistics. For a weight matrix $W \in \mathbb{R}^{m\times n}$, the computational overhead of DfI is dominated by the Gram matrix calculation $W^\top W$ or $WW^\top$. Denoting $r = \min(m, n)$, the single-layer DfI computation requires $\mathcal{O}(mnr)$ time and $\mathcal{O}(r^2)$ additional memory. Therefore, for the regularized layer set $\mathcal{C}$, the total additional computational cost can be expressed as

$$\mathcal{O}\left(\sum_{l\in\mathcal{C}} m_l n_l \min(m_l, n_l)\right) \tag{27}$$

Unlike discrete reinitialization-based approaches, AIPR does not require any explicit parameter reinitialization during training. Instead, it incorporates DfI as a continuous geometric regularization term into the gradient-based optimization process. This design preserves the original data sampling procedure, task objective formulation, and Actor–Critic interaction structure of the underlying algorithm. AIPR only augments the original Actor and Critic objectives with state-dependent geometric regularization terms, without modifying their fundamental learning mechanisms.

AIPR introduces only three lightweight operations into the standard Actor–Critic parameter update process. First, it computes the DfI values of selected layers to obtain the network-level geometric states. Then, adaptive regularization coefficients are determined according to the respective geometric states of the Actor and Critic networks. Finally, the corresponding DfI-based regularization

terms are incorporated into the task objectives, and the existing automatic differentiation pipeline is used to perform gradient-based parameter updates. Therefore, AIPR can be modularly integrated into various Actor–Critic algorithms without modifying their original training procedures.

G Experimental Details

This section provides additional details on the benchmark environments, baseline methods, and implementation settings used in our experiments.

MuJoCo (Todorov et al., 2012) is a widely adopted physics simulation platform for continuous control reinforcement learning, containing a variety of classical locomotion tasks with different state dimensions and dynamic properties. We select representative MuJoCo environments to evaluate the overall effectiveness of AIPR in standard continuous control settings.

Robosuite (Zhu et al., 2020) is a robot learning simulation framework focusing on complex manipulation tasks, including object interaction, robotic arm control, and contact-based operations. We use Robosuite to evaluate whether AIPR can maintain stable and effective optimization behavior in challenging control environments with high-dimensional dynamics.

CARL (Benjamins et al., 2021) is a benchmark framework for contextual reinforcement learning and robustness evaluation under environmental variations. It generates diverse task instances by varying environment context parameters, allowing us to assess the adaptability of AIPR under distribution shifts and dynamic changes.

We compare AIPR against three representative baseline methods targeting weight geometric degradation or plasticity loss in deep reinforcement learning, FIRE (Han et al., 2026), Parseval regularization (Chung et al., 2024), and PFO (Moalla et al., 2024).

FIRE (Han et al., 2026) adopts a discrete geometry-aware intervention mechanism, enabling recovery after substantial geometric degradation has occurred. Therefore, FIRE serves as a representative baseline for comparing AIPR's continuous adaptive regularization mechanism against discrete parameter recovery strategies.

Parseval Regularization (Chung et al., 2024) constrains weight matrices toward orthogonal or near-isometric structures, thereby reducing geometric distortion during network transformations and improving spectral properties throughout optimization. Unlike FIRE, it introduces geometric constraints continuously through gradient-based regularization. However, its regularization strength is typically predefined and fixed, making it an important baseline for evaluating AIPR's state-dependent adaptive regularization mechanism.

PFO (Moalla et al., 2024) addresses plasticity loss in long-term reinforcement learning by introducing regularization constraints that preserve the network's ability to continuously adapt during optimization. Rather than directly targeting weight geometry, PFO improves long-term learning capability from the perspective of plasticity preservation. Therefore, it provides a complementary baseline for evaluating the effectiveness of AIPR in maintaining network plasticity.

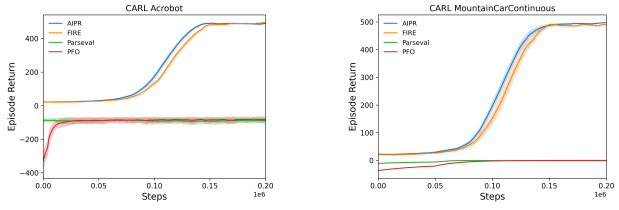
The aforementioned baselines represent three distinct categories of approaches, discrete parameter recovery, continuous geometric regularization, and plasticity preservation. To ensure a fair comparison, all baseline methods are evaluated using their original default hyperparameter settings. All learning curves in Figures 2, 3, 4, and 5 show the mean over 10 random seeds, with shaded regions indicating the corresponding standard deviation. All experiments are conducted on a workstation equipped with an NVIDIA RTX 4090D GPU, an AMD Ryzen 9 9950X CPU, and 64 GB DDR5 memory. The hyperparameter settings of AIPR are summarized in Table 2.

H Robustness under Changing Contexts

To evaluate the robustness and adaptability of AIPR under varying environmental conditions, we conduct experiments on the CARL benchmark. We select two representative contextual reinforcement learning tasks, CARL-Acrobot and CARL-MountainCarContinuous, and compare AIPR with FIRE, Parseval, and PFO. The experimental results are presented in Figure 5.

Table 2: Hyperparameter settings of AIPR.

Hyperparameter	Value
Shared hyperparameters	
Actor/Critic learning rate	1×10^{-4}
Actor/Critic hidden dimension	$128/512$
Discount factor γ	0.99
$\lambda^{\pi}_{\max}/\lambda^{v}_{\max}$	$0.01/0.01$
μ_{π}/μ_{v}	$1.0/1.0$
τ_{π}/τ_{v}	$0.5/0.5$
Random seeds	10
Training Budget	
MuJoCo	5×10^{6}
Robosuite	1×10^{6}
CARL	1×10^{6}
PPO-specific hyperparameters	
GAE parameter λ_{GAE}	0.95
Entropy coefficient	0.0
PPO rollout steps	2048
Mini-batch size	64
PPO update epochs	10
PPO clipping value	0.2
SAC-specific hyperparameters	
Target smoothing coefficient τ_{target}	0.005
Batch size	256
Replay buffer size	1×10^{6}
Target entropy coefficient	-0.5
Initial entropy temperature	0.01



(a) Acrobot (b) MountainCarContinuous

Figure 5: Learning curves of AIPR and baselines on CARL tasks.

As shown in Figure 5, AIPR consistently achieves competitive performance across CARL tasks with changing contextual conditions. In particular, for Acrobot and MountainCarContinuous, AIPR exhibits faster early-stage learning progress, entering the rapid improvement phase within the first 2×10^{5} environment steps and obtaining higher episode returns than FIRE. These results indicate that AIPR can effectively adapt geometric regularization to the evolving state of the network, it strengthens constraints when weight geometry becomes unstable and reduces unnecessary regularization when additional optimization flexibility is beneficial.

Overall, AIPR achieves stable and competitive performance across CARL tasks with varying environmental contexts, while demonstrating faster adaptation trends in tasks such as Acrobot and MountainCarContinuous. These results indicate that AIPR remains effective under contextual variations and can improve the efficiency of policy adaptation in dynamic environments, further validating its robustness and applicability beyond stationary task settings.