QUADRATIC ADVANTAGE RESPONSE POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy optimization methods such as trust region policy optimization (TRPO) and proximal policy optimization (PPO) improve training stability through policy change constraints. However, their surrogate objectives mainly regulate policy update magnitudes, without explicitly characterizing the relationship between advantage magnitude and policy response. We derive an analytical mapping from sample-wise quadratic curvature to advantage-conditioned responses under a general smooth quadratic probability-ratio surrogate objective. Based on this analysis, we propose Quadratic Advantage Response Policy Optimization (QARPO). QARPO assigns quadratic curvature according to relative advantage magnitude, suppressing response growth under extreme advantages. It preserves response ordering across different advantage magnitudes while retaining a smooth first-order optimization form. Theoretical analysis shows that QARPO preserves first-order gradient consistency at the old-policy point and characterizes the effect of policy normalization on the resulting scalar targets. Meanwhile, its advantage-dependent curvature reshapes the local second-order optimization structure. The quadratic restoration mechanism further provides analytical control of the PPO-Clip clipping gap, establishing an ordering between the smooth quadratic and clipped surrogates. Experiments on continuous-control and discrete-control benchmarks show that QARPO effectively regulates advantage-conditioned responses through advantage-dependent curvature design. Compared with representative policy optimization methods, QARPO achieves response differentiation and competitive performance.

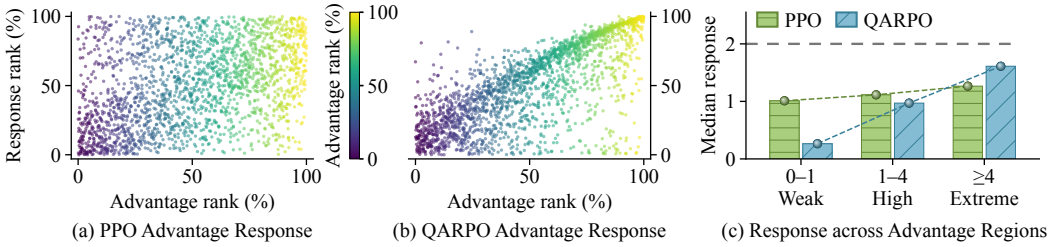
1 INTRODUCTION

In recent years, deep reinforcement learning has used deep neural networks to address complex decision-making problems, achieving substantial progress in areas such as game playing and robotic control (Hansen et al., 2024; Tirumala et al., 2025; Hafner et al., 2025). Policy-gradient methods take the expected return of a parameterized policy as the optimization objective, with policy parameters updated through gradient estimation (Sutton et al., 1999; Agarwal et al., 2021a). They provide a general optimization framework for stochastic-policy learning and continuous control. On this basis, actor-critic methods combine value-function estimation with policy optimization, where the advantage function measures the relative value of an action with respect to the current policy average (Konda & Tsitsiklis, 1999; Wu et al., 2018; Mei et al., 2022). In policy-gradient updates, the sign of the advantage determines the update direction, whereas its magnitude modulates the strength of a sample's contribution. The advantage-response relationship is therefore an important factor affecting policy-optimization behavior.

However, while policy gradients use advantage signals to guide policy updates, they must also control the magnitude of change between the new and old policies. Excessively large updates can cause drastic changes in the policy distribution, whereas overly conservative updates limit optimization efficiency. Accordingly, TRPO constructs a trust region through a Kullback–Leibler (KL) divergence constraint to limit policy changes explicitly (Schulman et al., 2015), whereas PPO uses probability-ratio clipping to obtain a simpler first-order optimization form (Schulman et al., 2017). Subsequent studies further show that design choices involving the surrogate objective, gradient estimation, normalization scheme, and optimizer all substantially affect the practical optimization behavior of policy gradients (Ilyas et al., 2020; Engstrom et al., 2020; Andrychowicz et al., 2021). The specific structure of a surrogate objective therefore not only prescribes the constraints on policy updates, but also affects the policy responses associated with different advantage signals. Existing surrogate objectives primarily focus

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.

on regulating policy update magnitudes, with the relationship between advantage magnitude and policy response remaining insufficiently characterized.



(a) PPO Advantage Response (b) QARPO Advantage Response (c) Response across Advantage Regions

Figure 1: Advantage-conditioned response scaling under a fixed-batch HalfCheetah-v4 setting. PPO and QARPO start from the same trained policy and use the same 2,048 rollout samples. (a-b) compare the advantage-response ordering of PPO and QARPO. (c) compares response strengths across advantage regions. PPO exhibits limited differentiation in response strengths across advantage magnitudes, whereas QARPO assigns advantage-dependent response strengths with ordered and bounded scaling behavior.

In probability-ratio policy optimization, different surrogate objectives limit policy changes through clipping (Schulman et al., 2017; Chen et al., 2018; Colletti & Holzinger, 2026), divergence regularization (Tomar et al., 2022; Grudzien et al., 2022), proximal constraints (Schulman et al., 2015; Li & Tan, 2025), and smooth restoration mechanisms (Kobayashi, 2021; Xie et al., 2025). These designs correspondingly form different probability-ratio update structures. For a smooth quadratic surrogate objective, this relationship can be further reduced to the relative interaction between an advantage-driven term and sample-wise quadratic curvature. The advantage signal determines the direction of policy improvement, whereas quadratic curvature regulates the restoration strength when the probability ratio departs from the old policy point. Together, they determine the independent scalar probability-ratio responses of different advantage samples. Accordingly, different curvature rules produce different response structures and trade-offs between preserving advantage-magnitude information and controlling response scale.

In this paper, we propose a novel policy optimization method, *quadratic advantage response policy optimization* (QARPO). The key difference in optimization behavior between PPO and QARPO is illustrated in Figure 1. PPO applies a uniform clipping threshold across samples with different advantage magnitudes, resulting in limited differentiation of response strengths. QARPO allocates response strengths according to relative advantage magnitudes, preserving advantage-response ordering while bounding responses for extreme advantages. Our main contributions are summarized as follows:

- We derive an analytical relationship between sample-wise quadratic curvature and independent scalar probability-ratio responses under a smooth quadratic objective. This analysis reveals the response structures associated with different curvature designs.
- We propose QARPO with advantage-dependent curvature allocation to regulate probability-ratio responses. The resulting surrogate preserves advantage ordering, bounds response growth, and maintains a smooth first-order optimization form.
- We analyze QARPO from local optimization properties and surrogate objective relationships. Experiments demonstrate that its advantage-dependent curvature design improves response regulation and optimization performance across tasks.

2 Related Work

Trust-region and clipped policy optimization. TRPO establishes a trust region by constraining the expected KL divergence between the new and old policies, thereby limiting excessively large policy updates. Its conjugate-gradient and line-search procedures introduce additional computational overhead (Schulman et al., 2015). PPO-Clip further constructs a first-order surrogate objective through probability-ratio clipping, which enables simpler control of policy changes (Schulman et al., 2017). However, hard clipping cannot strictly constrain the probability ratios of all samples or provide the same explicit trust-region constraint as TRPO (Wang et al., 2020). Achieving smoother and finer-grained control of policy updates while maintaining optimization efficiency has therefore become an important direction for subsequent research.

Refinements of proximal policy updates. Research on improving PPO's update constraints has gradually expanded from fixed clipping boundaries to redesigning surrogate objectives and update

signals. To address the limitations of the clipping mechanism in PPO-Clip, existing methods adjust the range of policy updates through state-dependent boundaries, adaptive clipping, and dynamic probability-ratio thresholds (Chen et al., 2018; 2023; Li & Tan, 2025). Meanwhile, MDPO, PPO-RPE, and SPO modify the surrogate objective with a mirror-descent proximal term, relative Pearson divergence, and smooth quadratic restoration term, respectively (Tomar et al., 2022; Kobayashi, 2021; Xie et al., 2025). Recent studies further consider fine-grained update mechanisms such as advantage modulation, per-sample penalty coefficients, and probability-ratio distribution constraints (Sane, 2025; Colletti & Holzinger, 2026; Ao et al., 2026; Luo et al., 2026). However, these methods primarily focus on the form of update constraints rather than the explicit relationship between advantage magnitude and response structure.

Advantage-response geometry. We revisit the relationship between advantage magnitude and policy response in smooth quadratic probability-ratio objectives. PPO-Clip applies a uniform clipping scale across different advantage magnitudes. After the probability ratio reaches the clipping plateau, the surrogate objective no longer explicitly adjusts the response according to additional advantage differences. Thus, it does not explicitly characterize the response scales associated with different advantage strengths. This limitation motivates redesigning the curvature structure to enable differentiated responses across advantage strengths while preserving smooth first-order optimization.

3 Preliminaries

3.1 Policy Gradient

In online reinforcement learning, a sequential decision process is commonly modeled as a Markov decision process (MDP), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces; $R(s_t, a_t)$ is the reward function; $P(s' \mid s, a)$ is the state-transition probability; and $\gamma \in (0, 1)$ is the discount factor. Given a stochastic policy $\pi_\theta(a \mid s)$, the policy-optimization objective is

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right], \qquad \tau = (s_0, a_0, s_1, a_1, \ldots). \tag{1}$$

To characterize the relative value of an action with respect to the average behavior of the current policy, define the advantage function as $A_{\pi_\theta}(s_t, a_t) = Q_{\pi_\theta}(s_t, a_t) - V_{\pi_\theta}(s_t)$, where $Q_{\pi_\theta}(s_t, a_t)$, $V_{\pi_\theta}(s_t)$, and d_{π_θ} denote the action-value function, state-value function, and discounted state-visitation distribution, respectively. By the policy gradient theorem, the gradient of the policy objective with respect to θ is

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{\substack{s\sim d^{\pi_\theta} \\ a\sim\pi_\theta(\cdot|s)}}\left[A_{\pi_\theta}(s, a)\nabla_\theta \log \pi_\theta(a \mid s)\right]. \tag{2}$$

3.2 Trust Region Policy Optimization

Policy-gradient methods directly update policy parameters using gradient estimates, but excessively large policy changes can cause a substantial shift between the new and old policy distributions. To limit the magnitude of a single update, TRPO constructs a trust region by constraining the KL divergence between the new and old policies. It then optimizes the following surrogate objective:

$$\max_\theta \mathbb{E}_{(s_t,a_t)\sim\mu_{\text{old}}}\left[r_t(\theta)\widehat{A}_t\right], \text{s.t. } \mathbb{E}_{(s_t,a_t)\sim\mu_{\text{old}}}[D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot \mid s_t) \,\|\, \pi_\theta(\cdot \mid s_t))] \le \delta_{\text{KL}}. \tag{3}$$

Here, $\mu_{\text{old}}(s, a) = d_{\pi_{\theta_{\text{old}}}}(s)\pi_{\theta_{\text{old}}}(a \mid s)$ denotes the discounted old-policy state–action distribution, $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ denotes the new-to-old policy ratio, $\hat{A}_t$ is the advantage estimate, and $\delta_{\text{KL}} > 0$ specifies the trust-region constraint scale. However, the constrained optimization in TRPO typically relies on second-order approximations and incurs high computational costs.

3.3 Proximal Policy Optimization

PPO suppresses excessively large policy changes within a single update by clipping the probability-ratio surrogate objective, enabling stable first-order policy optimization. PPO-Clip uses the following clipped surrogate objective:

$$J_{\text{clip}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\mu_{\text{old}}}\left[\min\Big(r_t(\theta)\widehat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\,\widehat{A}_t\Big)\right]. \tag{4}$$

Here, $\epsilon > 0$ is the probability-ratio clipping radius. PPO-Clip limits policy updates through fixed sample-wise probability-ratio boundaries. When a probability ratio moves beyond the clipping boundary in the direction encouraged by the advantage, the corresponding sample's gradient contribution is truncated to zero. This truncation stops the sample from contributing additional gradient signals and may discard information that is useful for optimization.

4 METHODOLOGY

To explicitly characterize the relationship between advantage magnitude and policy response, we formulate a general smooth quadratic probability-ratio objective. This formulation provides a basis for analyzing the connection between sample-wise quadratic curvature and advantage response.

4.1 GENERAL QUADRATIC OBJECTIVES

For the population analysis, let $\hat{A}_i = \hat{A}(s_i, a_i)$ denote the normalized advantage estimate and $c_i = c(s_i, a_i) > 0$ the sample-wise curvature, and let $\epsilon > 0$ be the response scale. These quantities are fixed with respect to θ. Expectations are taken under the fixed old-policy state-action distribution, include advantage-estimation randomness, and are assumed finite. Consider the quadratic surrogate

$$L_c(\theta) = \mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}\left[r_i(\theta)\widehat{A}_i - \frac{c_i}{2\epsilon}(r_i(\theta) - 1)^2\right]. \tag{5}$$

This objective separates the advantage-driven term from the quadratic restoration term centered at $r = 1$, with c_i determining the relative restoration strength for sample i. To characterize the sample-wise response, treat r_i as an independent scalar. For $c_i > 0$, the corresponding single-sample objective has a unique unconstrained maximizer (see Appendix A.1 for the complete proof):

$$r_i^\circ - 1 = \epsilon\frac{\widehat{A}_i}{c_i}, \qquad |r_i^\circ - 1| = \epsilon\frac{|\widehat{A}_i|}{c_i}. \tag{6}$$

Thus, the advantage sign determines the target direction, while $|\hat{A}_i|/c_i$ determines its displacement. Completing the square gives a curvature-weighted objective that fits the parameterized ratios toward the corresponding scalar targets, where C is independent of θ (see Appendix A.1).

$$L_c(\theta) = C - \frac{1}{2\epsilon}\mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}\left[c_i(r_i(\theta) - r_i^\circ)^2\right]. \tag{7}$$

To separate the overall advantage scale under the old-policy distribution from the relative advantage structure, define

$$\kappa_A = \mathbb{E}_{(s,a)\sim\mu_{\text{old}}}[|\widehat{A}|], \qquad \alpha_i = \frac{|\widehat{A}_i|}{\kappa_A}, \qquad \kappa_A > 0. \tag{8}$$

Further parameterize the curvature as $c_i = \kappa_A\phi(\alpha_i)$, with $\phi(\alpha) > 0$, where $\phi(\alpha)$ is the curvature-allocation function relative to the old-policy advantage reference scale. Define the dimensionless probability-ratio response function as $q(\alpha) = \alpha/\phi(\alpha)$. Then

$$r_i^\circ - 1 = \epsilon\,\text{sign}(\widehat{A}_i)q(\alpha_i), \qquad |r_i^\circ - 1| = \epsilon q(\alpha_i). \tag{9}$$

The response function $q(\alpha)$ characterizes the scaling of probability-ratio responses with respect to relative advantage magnitudes, providing a unified view of different curvature designs.

4.2 ADVANTAGE RESPONSE DESIGN

PPO-Clip modifies the surrogate incentive through the clipping thresholds $1 \pm \epsilon$, without imposing hard bounds on realized probability ratios. For nonzero advantages, its single-sample surrogate becomes flat once the advantage-aligned ratio displacement exceeds ϵ, although advantage magnitude still scales the gradient before the plateau. Thus, ϵ specifies an advantage-independent clipping threshold for the onset of the plateau, not a unique scalar maximizer. For comparison within the quadratic family, $\phi(\alpha) = \alpha$ yields the constant target response $q(\alpha) = 1$ for $\alpha > 0$, with $|r_i^\circ - 1| = \epsilon q(\alpha_i)$. This quadratic target matches the clipping scale of PPO-Clip, but the objectives have different maximizer structures. We instead seek quadratic targets whose displacements vary with advantage magnitude.

To retain advantage-strength information while eliminating curvature differences across samples, consider shared curvature $\phi(\alpha) = 1$, giving $q(\alpha) = \alpha$. The response then increases strictly with advantage magnitude, assigning smaller displacements to weak advantages and larger displacements to strong advantages, but its magnitude also grows without bound as α increases.

To limit the unbounded growth of the linear response in the high-advantage region while preserving the magnitude relationship for weak advantages, consider $q(\alpha) = \min(\alpha, 1)$ and $\phi(\alpha) = \max(1, \alpha)$. This form restores a bounded response, but $q(\alpha) = 1$ for $\alpha > 1$, preventing it from distinguishing high-advantage samples of different strengths.

$$q(0) = 0, \qquad q(\alpha_1) < q(\alpha_2) \quad (0 \le \alpha_1 < \alpha_2), \qquad \sup q(\alpha) < \infty \quad (\alpha \ge 0). \tag{10}$$

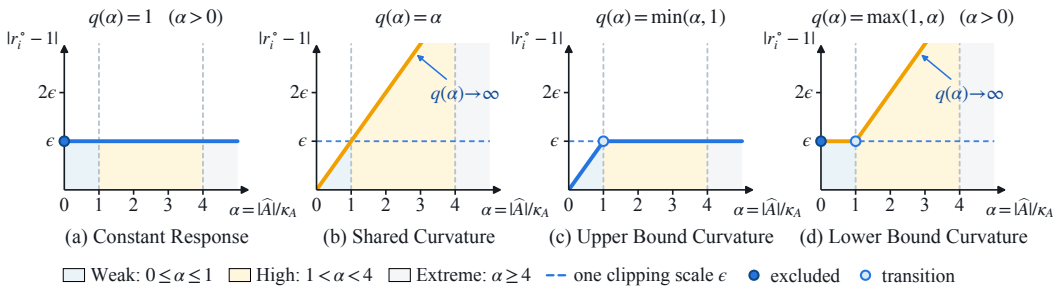


Figure 2: Independent scalar target responses of quadratic surrogates under $|r_i^\circ - 1| = \epsilon q(\alpha_i)$. (a) The constant response $q(\alpha) = 1$ for $\alpha > 0$ matches the advantage-independent clipping scale of PPO-Clip. (b) Shared curvature yields $q(\alpha) = \alpha$. (c–d) The other curvature designs yield $q(\alpha) = \min(\alpha, 1)$ and $q(\alpha) = \max(1, \alpha)$, respectively (the latter for $\alpha > 0$).

As shown in Figure 2, following the preceding curvature analysis and the requirements in Eq. (10), we consider an advantage response with three desirable properties. It approaches zero as the advantage magnitude of weak-advantage samples decreases and maintains strict response ordering across different advantage strengths. The probability-ratio response remains finite for extreme advantages, avoiding unbounded response growth.

4.3 Quadratic Advantage Response Policy Optimization

To satisfy the response requirements in Section 4.2, we retain shared curvature in the low-advantage region and redesign the high-advantage region. For $0 \le \alpha \le 1$, set $\phi(\alpha) = 1$, and for $\alpha > 1$, use the affine function $\phi(\alpha) = u + v\alpha$. Continuity at $\alpha = 1$ requires $u + v = 1$, and the high-advantage response is then $q(\alpha) = \alpha/(u + v\alpha)$. A finite tail response requires $v > 0$, whereas strict response ordering requires $u > 0$. Hence,

$$q_\lambda(\alpha) = \frac{\alpha}{\phi_\lambda(\alpha)}, \quad \phi_\lambda(\alpha) = \begin{cases} 1, & 0 \le \alpha \le 1 \\ 1 + \lambda(\alpha - 1), & \alpha > 1 \end{cases}. \tag{11}$$

For every $0 < \lambda < 1$, $q_\lambda(\alpha)$ is continuous and strictly increasing on $\alpha \ge 0$, and satisfies $\lim_{\alpha\to\infty} q_\lambda(\alpha) = 1/\lambda$. Thus, the response retains magnitude-dependent linear variation in the weak-advantage region and strict ordering in the high-advantage region, while remaining asymptotically finite. The complete proof is provided in Appendix A.2.

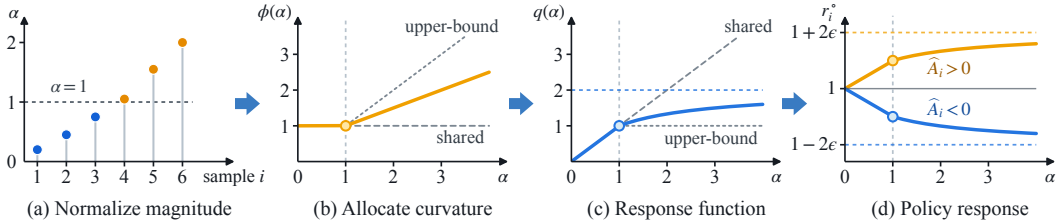


Figure 3: The quadratic advantage-response design process of QARPO. (a) Scale normalized advantage magnitudes by κ_A to define α_i. (b) Allocate sample-wise curvature $\phi_\lambda(\alpha)$ according to advantage magnitude. (c) Derive the corresponding probability-ratio response function $q_\lambda(\alpha)$. (d) Convert the response function into independent probability-ratio targets r_i°, with samples having positive and negative advantages producing symmetric updates around $r = 1$.

To give the quadratic restoration term both stability in the weak-advantage region and the ability to distinguish responses in the high-advantage region, the sample-wise quadratic curvature is

$$c_i^\lambda = \kappa_A + \lambda \max\left[(|\widehat{A}_i| - \kappa_A), 0\right]. \tag{12}$$

When $|\widehat{A}_i| \le \kappa_A$, the curvature remains at the base scale κ_A, preventing excessively strong corrections for low-magnitude advantage samples. When $|\widehat{A}_i| > \kappa_A$, the curvature increases with advantage magnitude, strengthening restoration for high-advantage samples.

Using the same old-policy sampling distribution, the QARPO surrogate objective is therefore

$$L_{\text{QARPO}}(\theta) = \mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}\left[r_i(\theta)\widehat{A}_i - \frac{\kappa_A + \lambda \max[(|\widehat{A}_i| - \kappa_A), 0]}{2\epsilon}(r_i(\theta) - 1)^2\right]. \tag{13}$$

The independent probability-ratio displacement induced by QARPO for sample i is $|r_i^\circ - 1| = \epsilon q(\alpha_i)$. We use $\lambda = 1/2$ as the default advantage-response adjustment parameter, for which $\lim_{\alpha\to\infty} q_{1/2}(\alpha) = 2$. Thus, weak-advantage samples receive responses that approach zero as advantage magnitude decreases. High-advantage samples retain strict ordering, with a supremum of 2ϵ for their independent probability-ratio displacements. Figure 3 shows the complete QARPO construction from relative advantage scaling and curvature allocation to probability-ratio response.

5 Theoretical Results

QARPO regulates sample-wise probability-ratio responses through advantage-dependent curvature. This section further analyzes the resulting local optimization structure and response characteristics, together with its relationship to PPO clipping.

5.1 Local Optimization Structure

QARPO's sample-wise curvature acts on the quadratic restoration term, so its effect is primarily reflected in the local optimization structure around the reference policy. Fix the old-policy distribution and treat $\widehat{A}_i$ and c_i as constants with respect to the policy parameters θ. Assume that the ratios are twice differentiable and that differentiation and expectation can be interchanged. Denote the unclipped probability-ratio objective by $L_{\mathrm{unclip}}(\theta)$ and the policy score vector at the reference policy π_{old} by z_i:

$$L_{\mathrm{unclip}}(\theta) = \mathbb{E}_{(s_i,a_i)\sim\mu_{\mathrm{old}}}[r_i(\theta)\widehat{A}_i], \qquad z_i = \nabla_\theta \log \pi_\theta(a_i \mid s_i)|_{\theta=\theta_{\mathrm{old}}}. \tag{14}$$

At the old-policy point $\theta = \theta_{\mathrm{old}}$, we have $r_i(\theta_{\mathrm{old}}) = 1$. Differentiating the quadratic surrogate with the advantage estimates and curvature coefficients held fixed gives

$$\nabla_\theta L_c = \mathbb{E}_{(s_i,a_i)\sim\mu_{\mathrm{old}}}[\widehat{A}_i z_i], \quad \nabla_\theta^2 L_c = \nabla_\theta^2 L_{\mathrm{unclip}} - \frac{1}{\epsilon}\mathbb{E}_{(s_i,a_i)\sim\mu_{\mathrm{old}}}[c_i z_i z_i^\top]. \tag{15}$$

Since $z_i z_i^\top$ is positive semidefinite, the curvature term introduces a sample-weighted negative-semidefinite second-order correction without changing the first-order gradient. PPO-Clip is inactive at $r_i = 1$, so QARPO and PPO-Clip have identical surrogate gradients at the old-policy point under the same advantages and sampling distribution. QARPO therefore preserves the first-order update direction while modifying the local optimization geometry through sample-dependent second-order restoration. The complete proof is provided in Appendix A.3.

5.2 Relation to PPO Clipping

Since PPO-Clip induces opposite probability-ratio movements for positive and negative advantages, we define a unified displacement by aligning the ratio change with the advantage direction. For nonzero-advantage samples, define $y_i = \mathrm{sign}(\widehat{A}_i)(r_i(\theta) - 1)$. The single-sample clipping gap can then be expressed uniformly as $\Delta_i = |\widehat{A}_i| \max[(y_i - \epsilon), 0]$.
Establishing a quadratic upper bound in terms of the probability-ratio displacement gives

$$\Delta_i \le \frac{|\widehat{A}_i|}{4\epsilon}(r_i(\theta) - 1)^2, \qquad \max[(x - \epsilon), 0] \le \frac{x^2}{4\epsilon}. \tag{16}$$

By the curvature definition in Section 4.3, $c_i^\lambda \ge \lambda|\widehat{A}_i|$. Thus, when $\lambda \ge 1/2$, the quadratic restoration term of QARPO covers the clipping gap introduced by PPO clipping. We therefore choose $\lambda = 1/2$ as the default setting, since it is the minimum value that guarantees this coverage while avoiding unnecessarily strong curvature amplification.

$$L_{\mathrm{QARPO}}(\theta) \le L_{\mathrm{PPO}}(\theta) \le L_{\mathrm{unclip}}(\theta). \tag{17}$$

The inequalities follow by taking the same old-policy expectation of the pointwise bounds. As shown in Eq. (17), QARPO establishes a controlled lower-bounded relationship with the PPO-Clip surrogate objective, where the objective difference induced by clipping is compensated through a continuous quadratic restoration term. The complete proof is provided in Appendix A.4.

5.3 Policy-Normalization Scalar Targets

At a fixed state s, write $p(a) = \pi_{\mathrm{old}}(a \mid s)$ for the old-policy probability mass or density, use $\hat{A}(a)$ and $c(a)$ instead of $\hat{A}(s,a)$ and $c(s,a)$. Under the conditions in Appendix A.5, the normalization-

constrained optimum takes the form

$$r^*(a) = \max\left[(1 + \epsilon\frac{\widehat{A}(a) - \tau_s^*}{c(a)}), 0\right], \quad \sum_a p(a)r^*(a) = 1 \quad \text{or} \int p(a)r^*(a)\,\mathrm{d}a = 1. \tag{18}$$

Here, τ_s^* enforces normalization via the sum for discrete actions or the integral for continuous actions. Let $\mathbb{E}_p = \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(\cdot|s)}$ denote the corresponding old-policy action expectation and define $\|f\|_{c,s}^2 = \mathbb{E}_p[cf^2]$. When truncation is inactive, any normalized policy ratio r_θ with finite weighted error satisfies

$$\tau_s^* = \frac{\mathbb{E}_p[\widehat{A}/c]}{\mathbb{E}_p[1/c]}, \quad r^\circ = 1 + \epsilon\frac{\widehat{A}}{c}, \quad \|r_\theta - r^\circ\|_{c,s}^2 = \|r_\theta - r^*\|_{c,s}^2 + \epsilon^2(\tau_s^*)^2\mathbb{E}_p[1/c]. \tag{19}$$

The final term of Eq. (19) is independent of θ, so fitting the independent and normalized targets induces equivalent objectives over the same policy family. Therefore, under this condition, the normalization multiplier τ_s^* need not be explicitly computed during optimization. Appendix A.5 provides the complete normalization analysis. Appendix A.6 further characterizes the corresponding shared-parameter tangent-space projection and error decomposition.

6 Experiments

We evaluate QARPO on MuJoCo continuous-control environments (Todorov et al., 2012) and Atari 2600 discrete-control environments (Bellemare et al., 2013; Machado et al., 2018). All environments are implemented through the Gymnasium interface (Towers et al., 2025), with a unified training framework and experimental protocol to ensure comparability across algorithms.

6.1 Algorithm Comparison

To systematically evaluate QARPO's overall performance, we compare it with representative policy-optimization methods, including PPO-Clip (Schulman et al., 2017), PPO-Penalty (Schulman et al., 2017), PPO-RB (Wang et al., 2020), TR-PPO (Wang et al., 2020), and SPO (Xie et al., 2025). QARPO uses $\lambda = 1/2$ across all environments without environment-level curvature-parameter tuning. Except for the policy surrogate objective and its method-specific parameters, the training procedure and main optimization configurations are identical. Complete hyperparameters are provided in Appendix B.1.

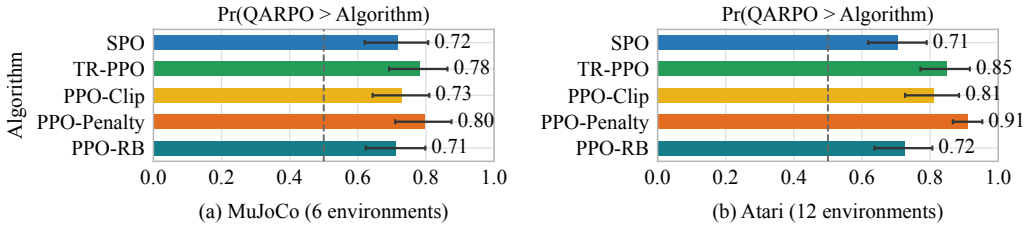


Figure 4: Probability of improvement (POI) of QARPO over representative baseline algorithms, with 95% confidence intervals. Appendix C.1 details the computation. Confidence intervals entirely above 0.5 indicate consistent improvement over the corresponding baseline (Agarwal et al., 2021b).

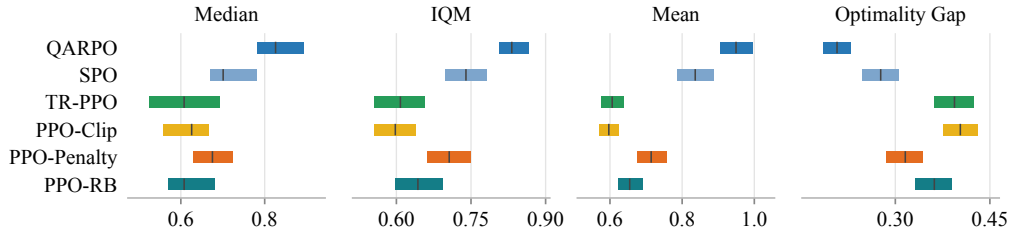


Figure 5: Cross-task aggregate performance statistics and optimality gap of the algorithms on MuJoCo. Median, Interquartile Mean (IQM), Mean, and Optimality Gap are computed over returns normalized with PPO-Clip as the reference algorithm across 6 MuJoCo tasks with 10 independent random seeds. Error bars indicate 95% confidence intervals. Detailed normalization procedures and metric definitions are provided in Appendix C.1.

As shown in Figure 4–6, QARPO demonstrates consistent improvements in aggregate performance metrics over representative policy-optimization baselines across the evaluated MuJoCo and Atari 2600 benchmarks. Complete per-environment performance comparisons and additional learning curves for representative methods are provided in Appendix C.1.

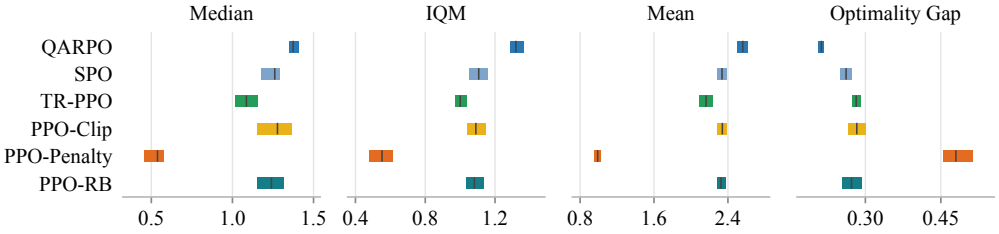


Figure 6: Cross-task aggregate performance statistics and optimality gap of the algorithms on Atari 2600. Median, IQM, Mean, and Optimality Gap are computed over human-normalized scores across 12 Atari 2600 games with 5 independent random seeds. Error bars indicate 95% confidence intervals. Detailed normalization procedures and metric definitions are provided in Appendix C.1.

6.2 Ablation Studies

To analyze the effects of different design choices in QARPO, we conduct ablations covering curvature structures, hyperparameter settings, and curvature functional forms. Except for the ablated factor and the number of random seeds, all other training configurations are kept identical to those in the algorithm comparison. This section reports ablation results on representative MuJoCo and Atari environments. Complete results for all evaluated environments are provided in Appendix C.2. Solid lines show the mean over 3 seeds, with shaded regions indicating standard deviation.

Curvature Structures. We compare the effects of different curvature-allocation strategies on QARPO's performance and analyze the role of advantage-dependent curvature design in regulating responses across different advantage samples. Representative results are shown in Figure 7 and Figure 8. These results show that advantage-dependent curvature provides more effective response regulation by maintaining stronger differentiation across advantage magnitudes.

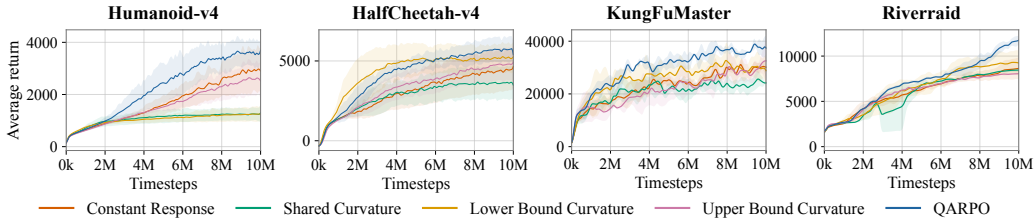


Figure 7: Training curves for the curvature-structure ablation in representative MuJoCo and Atari environments, comparing different advantage-dependent curvature-allocation strategies. The results indicate that advantage-dependent curvature allocation improves optimization by providing more effective response differentiation across advantage magnitudes.

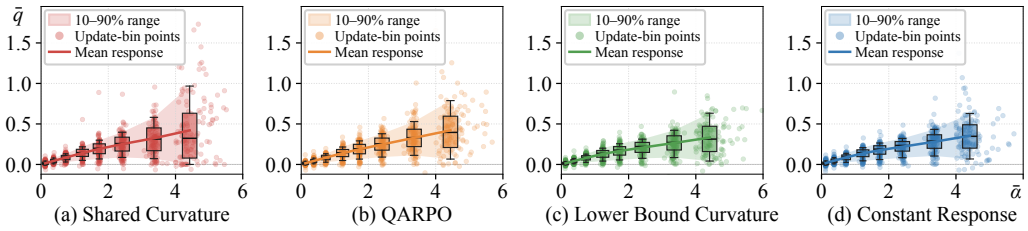


Figure 8: Advantage-conditioned training responses under four curvature designs in KungFuMaster over three random seeds. Samples are grouped into eight relative-advantage bins. Solid lines show mean advantage-aligned responses, shaded regions indicate the 10th–90th percentile range, and box plots summarize response distributions. Scatter points show sampled update-bin records.

Hyperparameter Sensitivity. We analyze the sensitivity of QARPO to two key hyperparameters governing the curvature strength and response constraint. Specifically, we evaluate different values

of the curvature parameter $\lambda_{\text{QARPO}} \in \{0.25, 0.5, 0.75\}$ and the response-constraint scale $\epsilon \in \{0.1, 0.2, 0.3\}$. Representative results are shown in Figure 9. The results show that $\lambda_{\text{QARPO}} = 0.5$ achieves consistently strong performance, indicating an effective balance between response adaptation and regularization strength. Varying ϵ preserves similar performance trends across tasks, demonstrating that QARPO remains robust to moderate changes in the response-constraint scale.

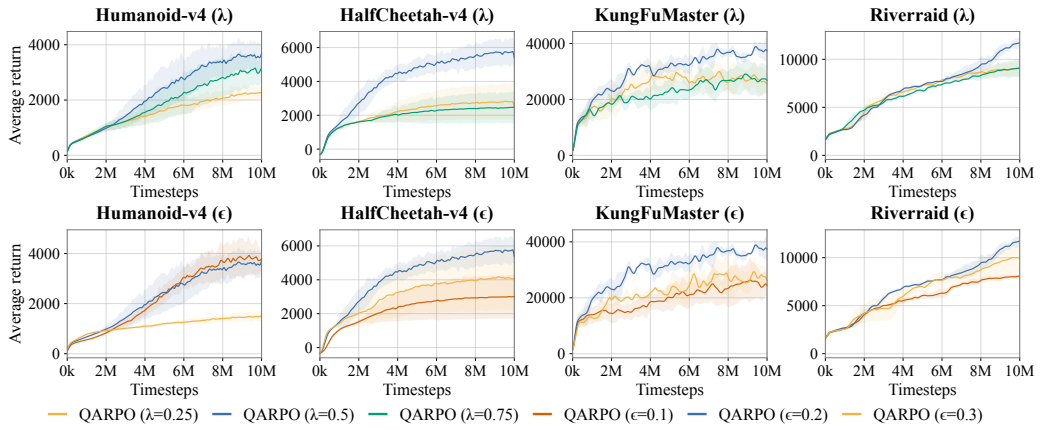


Figure 9: Hyperparameter sensitivity analysis of QARPO. QARPO remains robust to moderate changes in λ and ϵ, with the default configuration $(\lambda = 0.5, \epsilon = 0.2)$ balancing curvature strength and response constraint while providing consistently strong performance. Moderate variations of both parameters preserve similar optimization trends.

Curvature Functional Forms. We compare the piecewise curvature with the globally affine form $c_i^{\text{global}} = 1/2(|\widehat{A}_i| + \kappa_A)$. At the default $\lambda = 0.5$, both designs use sample-wise curvature and coincide for $\alpha_i \geq 1$, isolating their difference in the weak-advantage region. Representative results are shown in Figure 10. The return gains are consistent with the smaller independent scalar target displacements induced by the piecewise design for $0 < \alpha_i < 1$.

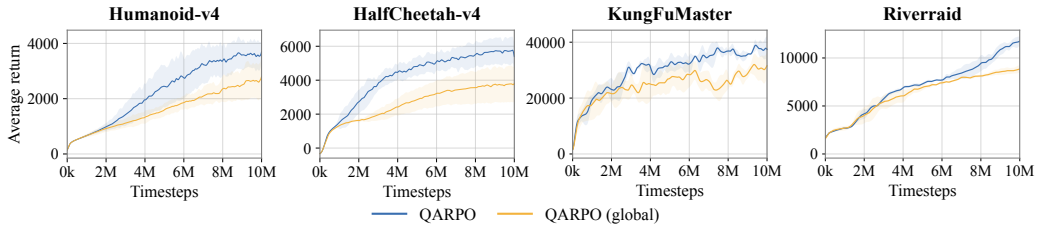


Figure 10: Curvature functional-form ablation on representative MuJoCo and Atari environments. The piecewise design achieves higher final mean returns than the globally affine variant in the displayed tasks.

7 Conclusion

In this paper, we propose QARPO, which replaces the heuristic hard clipping mechanism in PPO with advantage-dependent quadratic curvature while retaining a smooth first-order optimization form. This design produces differentiated probability-ratio responses across samples with different advantage strengths. Theoretical analysis shows that curvature structure effectively governs the probability-ratio response form and clarifies the effect of policy normalization on the corresponding scalar targets. Different curvature designs introduce different trade-offs among advantage-order preservation, response constraints, and optimization stability. Experimental results on MuJoCo and Atari 2600 demonstrate competitive performance against PPO and its variants. Ablation studies further verify the effects of different curvature designs on policy optimization behavior. These findings indicate that advantage-dependent curvature offers a structured approach to differentiating policy responses beyond uniform clipping. Future work may explore more general advantage-to-curvature mappings and investigate their interactions with policy parameterization, regularization mechanisms, and large-scale reinforcement learning tasks.

Reproducibility Statement

We provide the algorithmic details and implementation settings in the main text and appendix. The appendix includes the complete proofs of the theoretical results, experimental configurations, and hyperparameter settings to facilitate the reproduction of our results. Anonymous source code is available at: https://anonymous.4open.science/r/QARPO/.

AI Use Statement

Generative AI tools were used to improve the clarity, grammar, and readability of the manuscript. All AI-assisted revisions were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021a.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021b.

Matthew Aitchison, Penny Sweetser, and Marcus Hutter. Atari-5: Distilling the arcade learning environment down to five games. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 421–438, 2023.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021.

Yunke Ao, Le Chen, Bruce D. Lee, Assefa S. Wahd, Aline Czarnobai, Philipp Fürnstahl, Bernhard Schölkopf, and Andreas Krause. Bounded ratio reinforcement learning. *arXiv preprint arXiv:2604.18578*, 2026.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013. doi: 10.1613/jair.3912.

Gang Chen, Yiming Peng, and Mengjie Zhang. An adaptive clipping approach for proximal policy optimization. *arXiv preprint arXiv:1804.06461*, 2018.

Xing Chen, Dongcui Diao, Hechang Chen, Hengshuai Yao, Haiyin Piao, Zhixiao Sun, Zhiwei Yang, Randy Goebel, Bei Jiang, and Yi Chang. The sufficiency of off-policyness and soft clipping: PPO is still insufficient according to an off-policy measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7078–7086, 2023. doi: 10.1609/aaai.v37i6.25864.

Riccardo Colletti and Robin Holzinger. KLip-PPO: A per-sample KL perspective on PPO-clip. *arXiv preprint arXiv:2606.23932*, 2026.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2020.

Jakub Grudzien, Christian A. Schroeder De Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7825–7844, 2022.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020.

Taisuke Kobayashi. Proximal policy optimization with relative pearson divergence. In *2021 IEEE International Conference on Robotics and Automation*, pp. 8416–8421, 2021. doi: 10.1109/ICRA48506.2021.9560856.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems 12*, pp. 1008–1014, 1999.

Yao Li and Xiaoyang Tan. Candidate ratio guided proximal policy optimization. *Engineering Applications of Artificial Intelligence*, 152:110576, 2025. doi: 10.1016/j.engappai.2025.110576.

Yu Luo, Shuo Han, Yihan Hu, Lei Lv, Huaping Liu, Fuchun Sun, Jianye Hao, and Dong Li. Ratio-variance regularized policy optimization. *arXiv preprint arXiv:2605.26784*, 2026.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018. doi: 10.1613/jair.5699.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. The role of baselines in policy gradient optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17818–17830, 2022. doi: 10.52202/068431-1295.

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, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. doi: 10.1038/nature14236.

Soham Sane. AM-PPO: (advantage) alpha-modulation with proximal policy optimization. *arXiv preprint arXiv:2505.15514*, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015.

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

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057–1063, 1999.

Dhruva Tirumala, Markus Wulfmeier, Ben Moran, Sandy Huang, Jan Humplik, Guy Lever, Tuomas Haarnoja, Leonard Hasenclever, Arunkumar Byravan, Nathan Batchelor, Neil Sreendra, Kushal Patel, Marlon Gwira, Francesco Nori, Martin Riedmiller, and Nicolas Heess. Learning robot soccer from egocentric vision with deep reinforcement learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 165–184, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022.

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A Standard Interface for Reinforcement Learning Environments. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4916.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 113–122, 2020.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations*, 2018.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68813–68824, 2025.

APPENDIX

A THEORETICAL ANALYSES

A.1 DERIVATION OF THE SCALAR RESPONSE RELATION

Fix the old-policy sampling distribution. At a state–action pair (s_i, a_i), treat the advantage $\widehat{A}_i$ and curvature c_i as constants with respect to θ. Consider the pointwise quadratic objective

$$\ell_i(r) = r\widehat{A}_i - \frac{c_i}{2\epsilon}(r-1)^2, \qquad c_i > 0, \quad \epsilon > 0. \tag{A.1}$$

Its first and second derivatives are

$$\ell_i'(r) = \widehat{A}_i - \frac{c_i}{\epsilon}(r-1), \qquad \ell_i''(r) = -\frac{c_i}{\epsilon} < 0. \tag{A.2}$$

Thus, $\ell_i(r)$ is strictly concave in r and has a unique unconstrained maximizer. Setting $\ell_i'(r) = 0$ gives

$$r_i^\circ = 1 + \epsilon\frac{\widehat{A}_i}{c_i}, \qquad r_i^\circ - 1 = \epsilon\frac{\widehat{A}_i}{c_i}, \qquad |r_i^\circ - 1| = \epsilon\frac{|\widehat{A}_i|}{c_i}. \tag{A.3}$$

This establishes the scalar response relation used in the main text.
To obtain an equivalent fitting interpretation, we rewrite the quadratic objective around its maximizer. Expanding Eq. (A.1) gives

$$\ell_i(r) = -\frac{c_i}{2\epsilon}r^2 + \left(\hat{A}_i + \frac{c_i}{\epsilon}\right)r - \frac{c_i}{2\epsilon}. \tag{A.4}$$

Using the maximizer in Eq. (A.3) and completing the square yields

$$\ell_i(r) = \ell_i(r_i^\circ) - \frac{c_i}{2\epsilon}(r - r_i^\circ)^2. \tag{A.5}$$

The value at the optimum is Eq. (A.6)

$$\ell_i(r_i^\circ) = \hat{A}_i + \frac{\epsilon\hat{A}_i^2}{2c_i}. \tag{A.6}$$

Therefore, completing the square around the independent scalar target r_i° gives

$$\ell_i(r) = \widehat{A}_i + \frac{\epsilon\widehat{A}_i^2}{2c_i} - \frac{c_i}{2\epsilon}(r - r_i^\circ)^2. \tag{A.7}$$

Substituting the parameterized ratios $r_i(\theta)$ and taking expectation under the old-policy sampling distribution yields

$$L_c(\theta) = C - \frac{1}{2\epsilon}\mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}\left[c_i\big(r_i(\theta) - r_i^\circ\big)^2\right], \tag{A.8}$$

where

$$C = \mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}\left[\widehat{A}_i + \frac{\epsilon\widehat{A}_i^2}{2c_i}\right] \tag{A.9}$$

is independent of θ. Thus, maximizing the quadratic policy surrogate is equivalent to minimizing a curvature-weighted squared error between the parameterized ratios and the independent scalar targets.

A.2 PROOF OF THE ADVANTAGE-RESPONSE FUNCTION PROPERTIES

For $\phi_\lambda(\alpha) = 1 + \lambda\max[(\alpha-1), 0]$, with $0 < \lambda < 1$, the response function is

$$q_\lambda(\alpha) = \frac{\alpha}{\phi_\lambda(\alpha)} = \begin{cases} \alpha, & 0 \le \alpha \le 1, \\ \dfrac{\alpha}{1+\lambda(\alpha-1)}, & \alpha > 1. \end{cases} \tag{A.10}$$

At the junction $\alpha = 1$,

$$q_\lambda(1^-) = 1, \qquad q_\lambda(1^+) = \frac{1}{1+\lambda(1-1)} = 1. \tag{A.11}$$

Thus, $q_\lambda(\alpha)$ is continuous at $\alpha = 1$ and therefore over the entire domain $\alpha \geq 0$. For $0 \leq \alpha < 1$, $q'_\lambda(\alpha) = 1 > 0$. For $\alpha > 1$, differentiation gives

$$q'_\lambda(\alpha) = \frac{1 + \lambda(\alpha - 1) - \lambda\alpha}{[1 + \lambda(\alpha - 1)]^2} = \frac{1 - \lambda}{[1 + \lambda(\alpha - 1)]^2}. \tag{A.12}$$

Since $0 < \lambda < 1$, we have $q'_\lambda(\alpha) > 0$. Together with continuity at $\alpha = 1$, this establishes that $q_\lambda(\alpha)$ is strictly increasing over the entire domain $\alpha \geq 0$. Moreover,

$$\lim_{\alpha\to\infty} q_\lambda(\alpha) = \lim_{\alpha\to\infty} \frac{\alpha}{1 + \lambda(\alpha - 1)} = \frac{1}{\lambda}. \tag{A.13}$$

Thus, for $0 < \lambda < 1$, QARPO's response is continuous, strictly monotonic, and has a finite asymptotic upper bound.

A.3 Proof of Local Optimization Properties at the Old-Policy Point

Fix the old-policy sampling distribution and treat $\widehat{A}_i$ and c_i as constants with respect to the policy parameters θ. Assume that the ratios are twice differentiable and that the first and second derivatives can be passed through the expectation. Consider the pointwise quadratic surrogate objective

$$\ell_i(\theta) = r_i(\theta)\widehat{A}_i - \frac{c_i}{2\epsilon}(r_i(\theta) - 1)^2. \tag{A.14}$$

Taking the gradient with respect to θ yields

$$\nabla_\theta \ell_i(\theta) = \left[\widehat{A}_i - \frac{c_i}{\epsilon}(r_i(\theta) - 1)\right] \nabla_\theta r_i(\theta). \tag{A.15}$$

By the definition of the policy probability ratio $r_i(\theta)$ and the log-derivative identity, we have

$$\nabla_\theta r_i(\theta) = r_i(\theta)\nabla_\theta \log \pi_\theta(a_i \mid s_i). \tag{A.16}$$

At the old-policy point θ_{old}, $r_i(\theta_{\text{old}}) = 1$, so $\nabla_\theta r_i(\theta_{\text{old}}) = z_i$, where $z_i = \nabla_\theta \log \pi_\theta(a_i \mid s_i)|_{\theta=\theta_{\text{old}}}$. Since $r_i(\theta_{\text{old}}) - 1 = 0$, the quadratic restoration term vanishes from the first-order gradient, giving $\nabla_\theta \ell_i(\theta_{\text{old}}) = \widehat{A}_i z_i$. Taking the old-policy expectation gives

$$\nabla_\theta L_c(\theta_{\text{old}}) = \mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}[\widehat{A}_i z_i]. \tag{A.17}$$

Thus, at the old-policy point, QARPO has the same first-order gradient direction as the probability-ratio objective without the quadratic restoration term. The sample-wise curvature coefficient c_i does not affect the initial policy-update direction.

Taking the second derivative gives

$$\nabla^2_\theta \ell_i(\theta) = \left[\widehat{A}_i - \frac{c_i}{\epsilon}(r_i(\theta) - 1)\right] \nabla^2_\theta r_i(\theta) - \frac{c_i}{\epsilon}\nabla_\theta r_i(\theta)\nabla_\theta r_i(\theta)^\top. \tag{A.18}$$

At the old-policy point $\theta = \theta_{\text{old}}$,

$$\nabla^2_\theta \ell_i(\theta_{\text{old}}) = \widehat{A}_i \nabla^2_\theta r_i(\theta_{\text{old}}) - \frac{c_i}{\epsilon} z_i z_i^\top. \tag{A.19}$$

Denote the corresponding unclipped probability-ratio objective by

$$L_{\text{unclip}}(\theta) = \mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}[r_i(\theta)\widehat{A}_i]. \tag{A.20}$$

The complete Hessian is therefore

$$\nabla^2_\theta L_c(\theta_{\text{old}}) = \nabla^2_\theta L_{\text{unclip}}(\theta_{\text{old}}) - \frac{1}{\epsilon}\mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}[c_i z_i z_i^\top]. \tag{A.21}$$

Since $z_i z_i^\top \succeq 0$, $c_i > 0$, and $\epsilon > 0$,

$$-\frac{1}{\epsilon}\mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}[c_i z_i z_i^\top] \preceq 0. \tag{A.22}$$

That is, sample-wise curvature regulates the local optimization geometry through a negative-semidefinite second-order correction, assigning stronger restoration constraints to samples with larger advantage magnitudes while maintaining first-order update consistency at the old-policy point.

A.4 PPO-Clipping Gap Control Proof

Since PPO-Clip limits probability-ratio changes in the positive- and negative-advantage directions according to the advantage sign, define

$$y_i = \operatorname{sign}(\widehat{A}_i)(r_i(\theta) - 1). \tag{A.23}$$

For nonzero-advantage samples, the difference between the unclipped objective and the PPO-Clip objective can then be expressed uniformly as

$$\Delta_i = |\widehat{A}_i| \max[(y_i - \epsilon), 0]. \tag{A.24}$$

For any real number x, we have

$$\frac{x^2}{4\epsilon} - (x - \epsilon) = \frac{(x - 2\epsilon)^2}{4\epsilon} \geq 0 \quad (x > \epsilon), \qquad \max[(x - \epsilon), 0] \leq \frac{x^2}{4\epsilon}. \tag{A.25}$$

Setting $x = y_i$ gives $\Delta_i \leq \frac{|\widehat{A}_i|}{4\epsilon} y_i^2$. Since $y_i^2 = (r_i(\theta) - 1)^2$, it follows that

$$\Delta_i \leq \frac{|\widehat{A}_i|}{4\epsilon}(r_i(\theta) - 1)^2. \tag{A.26}$$

Further, by the QARPO curvature definition,

$$c_i^\lambda = \kappa_A + \lambda \max[(|\widehat{A}_i| - \kappa_A), 0], \qquad \kappa_A > 0, \quad 0 < \lambda < 1. \tag{A.27}$$

The curvature definition gives $c_i^\lambda \geq \lambda |\widehat{A}_i|$, so $|\widehat{A}_i| \leq c_i^\lambda / \lambda$. Substitution into the clipping-gap upper bound yields

$$\Delta_i \leq \frac{c_i^\lambda}{4\lambda\epsilon}(r_i(\theta) - 1)^2. \tag{A.28}$$

Taking the same old-policy expectation on both sides gives

$$0 \leq L_{\text{unclip}}(\theta) - L_{\text{PPO}}(\theta) \leq \frac{1}{4\lambda\epsilon} \mathbb{E}_{(s_i, a_i) \sim \mu_{\text{old}}} \left[c_i^\lambda (r_i(\theta) - 1)^2 \right]. \tag{A.29}$$

When $\lambda \geq 1/2$, the quadratic penalty strength provided by QARPO's restoration term is no smaller than the preceding clipping-gap upper bound. Hence,

$$L_{\text{QARPO}}(\theta) \leq L_{\text{PPO}}(\theta) \leq L_{\text{unclip}}(\theta). \tag{A.30}$$

QARPO's quadratic restoration term provides quadratic control of the PPO-Clip clipping gap, maintaining a controlled, consistent relationship between the surrogate objectives at the same probability-ratio scale.

A.5 Policy Normalization in Discrete and Continuous Action Spaces

We analyze how normalization constraints affect the independent scalar objective at a fixed state s, treating advantages and curvatures as given quantities. Expectations in this subsection are over the complete old-policy action distribution at that state, not over the sampled actions in a batch of rollout. We write for brevity that $\widehat{A}(a) = \widehat{A}(s, a), c(a) = c(s, a), r(a) := r_\theta(a) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}$, where we condition on deterministic functions $\widehat{A}$ and c; trajectory-based advantage estimates require an additional sampling model.

Discrete action spaces. Let $\mathcal{A}$ be a finite action set, and write $p_a = \pi_{\text{old}}(a \mid s) > 0$, with $\sum_{a \in \mathcal{A}} p_a = 1$. Consider the constrained optimization problem with $\epsilon > 0$:

$$\max_{\{r \geq 0\}} J_s(r) = \sum_{a \in \mathcal{A}} p_a \left[\widehat{A}(a) r(a) - \frac{c(a)}{2\epsilon}(r(a) - 1)^2 \right], \text{s.t.} \sum_{a \in \mathcal{A}} p_a r(a) = 1. \tag{A.31}$$

The unique optimizer takes the following form, with a unique normalization multiplier τ_s^*:

$$r^*(a) = \max\left(1 + \epsilon \frac{\widehat{A}(a) - \tau_s^*}{c(a)}, 0\right), \qquad \sum_a p_a \max\left(1 + \epsilon \frac{\widehat{A}(a) - \tau_s^*}{c(a)}, 0\right) = 1. \tag{A.32}$$

The corresponding normalized policy is then given by $\pi^*(a \mid s) = p_a r^*$. To prove this result, first observe that $r = 1$ is feasible and that feasibility implies $0 \le r \le 1/p_a$. The feasible set is therefore nonempty and compact, and strict concavity of the objective ensures the existence of a unique optimizer. Introducing a multiplier τ_s for the equality constraint and multipliers $\nu_a \ge 0$ for nonnegativity gives the Lagrangian

$$\mathcal{L}(r; \tau_s, \nu_a) = J_s(r) - \tau_s \left(\sum_a p_a r(a) - 1 \right) + \sum_a \nu_a r(a). \tag{A.33}$$

The Karush–Kuhn–Tucker (KKT) conditions for this constrained problem are

$$p_a \left[\widehat{A} - \frac{c}{\epsilon}(r-1) - \tau_s \right] + \nu_a = 0, \quad r \ge 0, \quad \nu_a \ge 0, \quad \nu_a r = 0, \quad \forall a. \tag{A.34}$$

If $r > 0$, complementary slackness gives $\nu_a = 0$ and hence $r = 1 + \epsilon(\widehat{A} - \tau_s)/c$. If $r = 0$, stationarity and $\nu_a \ge 0$ imply $1 + \epsilon(\widehat{A} - \tau_s)/c \le 0$. Combining these two cases yields the optimizer in Eq. (A.32). Define the normalization function and note its limiting values:

$$F_s(\tau) = \sum_a p_a \max\left(1 + \epsilon \frac{\widehat{A} - \tau}{c}, 0 \right), \quad \lim_{\tau \to -\infty} F_s(\tau) \quad = +\infty, \lim_{\tau \to +\infty} F_s(\tau) = 0. \tag{A.35}$$

Continuity and monotonic nonincrease of F_s ensure that a multiplier satisfying $F_s(\tau_s^*) = 1$ exists. If $\tau_1 < \tau_2$ and $F_s(\tau_2) > 0$, at least one positive term strictly decreases between τ_1 and τ_2, while all remaining terms are nonincreasing. Thus $F_s(\tau_1) > F_s(\tau_2)$, establishing uniqueness of the normalization multiplier and completing the proof.

Relative to the independent scalar target $r^\circ = 1 + \epsilon \widehat{A}/c$, the normalized solution satisfies

$$r^* = \max\left(r^\circ - \frac{\epsilon \tau_s^*}{c}, 0 \right). \tag{A.36}$$

Normalization therefore couples actions through a state-dependent multiplier, rather than requiring each action to attain its independent target r°.

Continuous action spaces. Let $p(a) = \pi_{\text{old}}(a \mid s)$ denote the old-policy density and $\mathbb{E}_p$ its expectation. Assume that these functions are measurable and finite, with

$$c(a) \ge c_{\min} > 0, \qquad \mathbb{E}_p[c] < \infty, \qquad \mathbb{E}_p[\widehat{A}^2/c] < \infty. \tag{A.37}$$

All functional relations below are understood in the p-almost-everywhere (a.e.) sense. Over the feasible set

$$\mathcal{F}_s = \left\{ r(a) \ge 0 : \ \mathbb{E}_p[r(a)] = 1, \ \mathbb{E}_p[c(r(a)-1)^2] < \infty \right\}, \tag{A.38}$$

we consider the following optimization problem under the normalization constraint:

$$\max_{r \in \mathcal{F}_s} J_s(r) = \mathbb{E}_p \left[\widehat{A} r - \frac{c}{2\epsilon}(r-1)^2 \right]. \tag{A.39}$$

The optimizer is unique up to $p-$ a.e. equality and has the same form as in the discrete case:

$$r^*(a) = \max\left(1 + \epsilon \frac{\widehat{A}(a) - \tau_s^*}{c(a)}, 0 \right), \qquad \mathbb{E}_p \left[\max\left(1 + \epsilon \frac{\widehat{A} - \tau_s^*}{c}, 0 \right) \right] = 1. \tag{A.40}$$

Let $F_s(\tau) = \mathbb{E}_p[\max(1 + \epsilon(\widehat{A} - \tau)/c, 0)]$. The integrability assumptions and the Cauchy–Schwarz inequality imply $\mathbb{E}_p[|\widehat{A}|/c] < \infty$, so F_s is finite and satisfies

$$|F_s(\tau_1) - F_s(\tau_2)| \le \epsilon |\tau_1 - \tau_2| \mathbb{E}_p[1/c]. \tag{A.41}$$

This bound establishes continuity of F_s as a function of τ. For $\tau \ge 0$, the integrand is bounded by the integrable function $1 + \epsilon|\widehat{A}|/c$. As $\tau \to -\infty$, the integrand increases monotonically to positive infinity pointwise. Dominated convergence and monotone convergence, respectively, therefore give the limits

$$\lim_{\tau \to +\infty} F_s(\tau) = 0, \qquad \lim_{\tau \to -\infty} F_s(\tau) = +\infty. \tag{A.42}$$

As in the discrete case, if $F_s(\tau_2) > 0$, the set on which the integrand is positive has positive p-measure. Consequently, $\tau_1 < \tau_2$ implies $F_s(\tau_1) > F_s(\tau_2)$, so a unique τ_s^* satisfies $F_s(\tau_s^*) = 1$. The resulting r^* is nonnegative and normalized, and it also satisfies

$$\mathbb{E}_p[c(r^* - 1)^2] \le \epsilon^2 \mathbb{E}_p[(\widehat{A} - \tau_s^*)^2/c] < \infty. \tag{A.43}$$

Thus the candidate r^* belongs to the feasible set $\mathcal{F}_s$.

To establish global optimality, define $h = \widehat{A} - c(r^* - 1)/\epsilon$. The candidate solution implies $h = \tau_s^*$ wherever $r^* > 0$, and $h \le \tau_s^*$ wherever $r^* = 0$. For any feasible r, nonnegativity and the identity $\mathbb{E}_p[r - r^*] = 0$ give

$$\begin{aligned} J_s(r) - J_s(r^*) &= \mathbb{E}_p[h(r - r^*)] - \frac{1}{2\epsilon}\mathbb{E}_p[c(r - r^*)^2] \\ &\le -\frac{1}{2\epsilon}\mathbb{E}_p[c(r - r^*)^2] \le 0. \end{aligned} \tag{A.44}$$

Equality holds only when $r = r^*$ is $p-$ a.e., establishing both optimality and uniqueness.

There is also a geometric proof valid when the maximum operation is active. Under Eq. (A.37), we denote $\mathcal{H}_s = L^2(p, c)$ a Hilbert space with inner product $\langle f, f' \rangle_{c,s} = \mathbb{E}_{a\sim p}[c(a)f(a)f'(a)]$ and induced norm $\|f\|_{c,s} = \sqrt{\langle f, f\rangle_{c,s}}$. The set $\mathcal{F}_s$ is nonempty, closed, and convex: $1 \in \mathcal{F}_s$, nonnegativity defines a closed cone, and $\mathbb{E}_p[r]$ is continuous because $\mathbb{E}_p[1/c] < \infty$. Completing the square makes r^* the metric projection of $r^\circ = 1 + \epsilon\widehat{A}/c$ onto $\mathcal{F}_s$. Its variational inequality $\langle r^\circ - r^*, r - r^* \rangle_{c,s} \le 0$ for every $r \in \mathcal{F}_s$ gives

$$\|r - r^\circ\|_{c,s}^2 \ge \|r - r^*\|_{c,s}^2 + \|r^* - r^\circ\|_{c,s}^2. \tag{A.45}$$

This bounds distance to the normalized optimum by distance to the independent target, including at an active nonnegativity boundary. The omitted cross term is nonnegative and generally depends on r, so this inequality does not imply that the two fitting problems have the same minimizer over a restricted parameterized policy family. Exact equality is established below only when the normalization correction is orthogonal.

For QARPO, use the fixed distribution-level reference scale $\kappa_A > 0$ from Eq. (8) and choose the curvature function

$$c(a) = \kappa_A + \lambda \max\left(|\widehat{A}(a)| - \kappa_A, 0\right), \qquad 0 < \lambda < 1. \tag{A.46}$$

Then $c \ge \kappa_A$, $c \ge \lambda|\widehat{A}|$, and $c \le \kappa_A + \lambda|\widehat{A}|$. Therefore, $\mathbb{E}_p[|\widehat{A}|] < \infty$ suffices to ensure the integrability conditions in Eq. (A.37).

The result also applies when the old policy is a nondegenerate Gaussian distribution, although the optimal normalized density $\pi^*(\cdot \mid s) = pr^*$ is generally non-Gaussian. Any valid parametric density $\pi_\theta(\cdot \mid s)$ that is absolutely continuous with respect to the old density is automatically normalized through its distributional parameterization; Gaussian policies are one example:

$$\mathbb{E}_{a\sim\pi_{\text{old}}(\cdot|s)}[r_\theta(a)] = \int \pi_\theta(a \mid s)\, da = 1. \tag{A.47}$$

The Gaussian family and parameter sharing across states further restrict the realizable ratios. For both discrete and continuous actions, their relation to the independent targets admits a sharper characterization when the maximum operation in Eqs. (A.32) and (A.40) is inactive. At a fixed state s, we have

$$r^\circ = 1 + \epsilon\frac{\widehat{A}}{c}, \qquad r^* = r^\circ - \frac{\epsilon\tau_s^*}{c}, \qquad \tau_s^* = \frac{\mathbb{E}_p[\widehat{A}/c]}{\mathbb{E}_p[1/c]}. \tag{A.48}$$

For any normalized policy ratio r_θ with finite weighted error, the normalization correction is orthogonal to $r_\theta - r^*$:

$$\langle r_\theta - r^*, r^\circ - r^* \rangle_{c,s} = \epsilon\tau_s^* \mathbb{E}_p[r_\theta - r^*] = 0. \tag{A.49}$$

Expanding the squared error therefore yields

$$\|r_\theta - r^\circ\|_{c,s}^2 = \|r_\theta - r^*\|_{c,s}^2 + \epsilon^2(\tau_s^*)^2\mathbb{E}_p[1/c]. \tag{A.50}$$

The normalization correction term $\epsilon^2(\tau_s^*)^2\mathbb{E}_p[1/c]$ is independent of the policy parameters. Therefore, whenever the orthogonality condition in Eq. (A.49) holds, fitting the independent target and fitting the normalized target induce equivalent optimization objectives over the same policy family. If the condition holds at almost every state, the equivalence also holds after averaging over a fixed state distribution, provided the expectations are finite. This does not require the normalized target to be realizable by the parameterized policy. The exact equality need not hold when the maximum operation is active, and it concerns expectations under the old policy rather than exact identities on a finite minibatch.

A.6 Local Tangent-Space Projection and Error Decomposition

Fix the old-policy joint measure $\mu_{\text{old}}(s,a) = d^{\pi_{\theta_{\text{old}}}}(s)\pi_{\theta_{\text{old}}}(a \mid s)$, where $d^{\pi_{\theta_{\text{old}}}}$ matches the rollout's time weighting, and a finite-dimensional policy parameter θ. For deterministic functions $\widehat{A}(s,a)$ and $c(s,a) > 0$ fixed independently of the minibatch. A matching old-policy minibatch estimates this expectation without another factor of $\pi_{\text{old}}(a \mid s)$; Algorithm 1 instead freezes data-dependent finite-rollout coefficients; the deterministic population objective here is an idealized counterpart, not an exact finite-rollout identity. Assume that $\pi_\theta(\cdot \mid s)$ has fixed support near θ_{old}, is differentiable with differentiation permitted under the normalization integral, and that its ratio r_θ in $L^2(\mu_{\text{old}}, c)$ has a first-order expansion

$$r_\theta(s,a) = 1 + z(s,a)^\top \Delta + O(\|\Delta\|^2), \tag{A.51}$$

where $\Delta = \theta - \theta_{\text{old}}$ and $z(s,a) = \nabla_\theta \log \pi_\theta(a|s)\big|_{\theta=\theta_{\text{old}}}$. Assume $\mathbb{E}_{\mu_{\text{old}}}[c] < \infty$, $\mathbb{E}_{\mu_{\text{old}}}[\widehat{A}^2/c] < \infty$, and $\mathbb{E}_{\mu_{\text{old}}}[c\|z\|^2] < \infty$. These conditions make the following weighted fitting problem finite and give the conditional score identity $\mathbb{E}_{\pi_{\text{old}}}[z \mid s] = 0$,

$$\min_{\Delta} \ \mathbb{E}_{\mu_{\text{old}}}\left[c\left(z^\top\Delta - \epsilon\frac{\widehat{A}}{c}\right)^2\right]. \tag{A.52}$$

Expanding Eq. (A.52) gives $\Delta^\top F_c\Delta - 2\epsilon g^\top\Delta + \epsilon^2\mathbb{E}_{\mu_{\text{old}}}[\widehat{A}^2/c]$, where $F_c = \mathbb{E}_{\mu_{\text{old}}}[czz^\top], g = \mathbb{E}_{\mu_{\text{old}}}[\widehat{A}z]$. Its normal equation is $F_c\Delta = \epsilon g$. If $v \in \ker F_c$, then $\mathbb{E}_{\mu_{\text{old}}}[c(v^\top z)^2] = 0$, so $v^\top z = 0$ almost surely and $v^\top g = 0$. Consequently $g \in \text{range}(F_c)$ even when F_c is singular. Every minimizer has the form $\epsilon F_c^\dagger g + v$ for $v \in \ker F_c$; the minimum-norm choice is

$$\Delta_{\min} = \epsilon F_c^\dagger g. \tag{A.53}$$

If the population surrogate is twice differentiable and differentiation can be interchanged with expectation, its Hessian at θ_{old} is $\nabla_\theta^2 L_{\text{unclip}}(\theta_{\text{old}}) - F_c/\epsilon$, where $L_{\text{unclip}}(\theta) = \mathbb{E}_{\mu_{\text{old}}}[r_\theta\widehat{A}]$. If $F_c/\epsilon \succ \nabla_\theta^2 L_{\text{unclip}}(\theta_{\text{old}})$, the full quadratic Taylor model instead has unique maximizer $(F_c/\epsilon - \nabla_\theta^2 L_{\text{unclip}}(\theta_{\text{old}}))^{-1}g$.
The minimum weighted residual is

$$\mathcal{E}_{\tan} = \epsilon^2\left(\mathbb{E}_{\mu_{\text{old}}}[\widehat{A}^2/c] - g^\top F_c^\dagger g\right) \geq 0. \tag{A.54}$$

Suppose additionally that the normalized optimizer of Appendix A.5 is interior for $d^{\pi_{\theta_{\text{old}}}}$-a.e. state. The normalized target displacement is $\epsilon(\widehat{A} - \tau_s^*)/c$. For the same shared parameter vector Δ, its weighted normal equation is $F_c\Delta = \epsilon\mathbb{E}_{\mu_{\text{old}}}[(\widehat{A} - \tau_s^*)z] = \epsilon g$, since $\tau_s^*\mathbb{E}_{\pi_{\text{old}}}[z \mid s] = 0$. The exact orthogonal energy split is

$$\mathcal{E}_{\tan} = \epsilon^2\mathbb{E}_{s\sim d^{\pi_{\theta_{\text{old}}}}}\left[(\tau_s^*)^2\mathbb{E}_{\pi_{\text{old}}}\left[\frac{1}{c} \mid s\right]\right] + \epsilon^2\left(\mathbb{E}_{\mu_{\text{old}}}\left[\frac{(\widehat{A} - \tau_s^*)^2}{c}\right] - g^\top F_c^\dagger g\right). \tag{A.55}$$

The first nonnegative term is the cost of normalizing the independent target; the second is the representation error of the normalized target in the shared-parameter tangent space. When the nonnegativity projection is active, Eq. (A.45) remains valid, but the normalized target is no longer a linear centering correction and need not have the same tangent-space fit. The tangent calculation concerns a local quadratic model and supplies neither a hard ratio bound nor a guarantee of improved environment return.

B Algorithm Details

B.1 Hyperparameters

To ensure a fair comparison across policy-optimization methods, QARPO and all baseline algorithms use a unified training procedure and experimental configuration, without additional tuning for specific environments. Except for each algorithm's surrogate objective and QARPO's method-specific curvature parameter, all methods use the same data-sampling scheme, optimizer settings, and main training hyperparameters. Table B.1 summarizes the main training hyperparameters used in the Atari 2600 and MuJoCo experiments.

Table B.1: Training hyperparameter settings for the Atari 2600 and MuJoCo-v4 experiments.

Hyperparameters	Atari 2600	MuJoCo-v4
Number of workers	8	8
Rollout horizon	128	256
Learning rate	0.00025	0.0003
Learning rate decay	Linear	Linear
Optimizer	Adam	Adam
Total steps	10M	10M
Batch size	1024	2048
Update epochs	4	10
Number of minibatches	4	4
Minibatch size	256	512
GAE parameter	0.95	0.95
Discount factor	0.99	0.99
Value loss coefficient	0.5	0.5
Entropy loss coefficient	0.01	0.0
Probability ratio scale	0.2	0.2
QARPO curvature λ	0.5	0.5
PPO-Penalty target KL	0.02	0.01
PPO-RB rollback coefficient	0.01	0.3 / 0.02
TR-PPO KL threshold	0.001	0.035 / 0.05

PPO-Penalty initializes its KL penalty coefficient to 1 in both domains. For the MuJoCo PPO-RB rollback coefficient and TR-PPO KL threshold, slash-separated entries give the value for all other tasks first and Humanoid-v4 second. The Atari TR-PPO threshold starts at 0.001 and decays linearly during training, whereas the MuJoCo thresholds remain fixed.

B.2 Training Procedure

In practical training, QARPO follows the PPO training procedure. Within each minibatch, it additionally computes the reference scale and sample-wise quadratic curvature and constructs the corresponding policy surrogate objective. The complete procedure is shown in Algorithm 1.

Algorithm 1 Quadratic Advantage Response Policy Optimization (QARPO)

Require: Policy π_θ, value function V_φ, response scale ϵ, curvature parameter λ (default $\lambda = 1/2$), coefficients c_1, c_2, normalization constant $\eta > 0$
Ensure: Optimized policy π_{θ^*}

while not converged **do**
  $\theta_{\text{old}} \leftarrow \theta$, $\varphi_0 \leftarrow \varphi$
  Collect trajectories $\mathcal{D} = \{(s_t, a_t, R_t)\}_{t=1}^{N}$ using $\pi_{\theta_{\text{old}}}$
  Estimate raw advantages $\widehat{A}_t^{\text{raw}}$ and value targets $\widehat{R}_t$ using V_{φ_0}
  for each optimization epoch **do**
    Shuffle $\mathcal{D}$ into minibatches
    for each minibatch $\mathcal{M}$, $m = |\mathcal{M}|$ **do**
      $\mu_{\mathcal{M}} \leftarrow m^{-1} \sum_{i \in \mathcal{M}} \widehat{A}_i^{\text{raw}}, \quad \sigma_{\mathcal{M}} \leftarrow \text{Std}_{i \in \mathcal{M}}(\widehat{A}_i^{\text{raw}})$
      $\widehat{A}_i^{\mathcal{M}} \leftarrow (\widehat{A}_i^{\text{raw}} - \mu_{\mathcal{M}})/(\sigma_{\mathcal{M}} + \eta)$ for $i \in \mathcal{M}$
      $\widehat{\kappa}_{A,\mathcal{M}} \leftarrow m^{-1} \sum_{i \in \mathcal{M}} |\widehat{A}_i^{\mathcal{M}}|$
      $c_i^{\mathcal{M},\lambda} \leftarrow \widehat{\kappa}_{A,\mathcal{M}} + \lambda \max\left[(|\widehat{A}_i^{\mathcal{M}}| - \widehat{\kappa}_{A,\mathcal{M}}), 0\right]$
      $r_i(\theta) \leftarrow \pi_\theta(a_i \mid s_i)/\pi_{\theta_{\text{old}}}(a_i \mid s_i)$
      $\mathcal{L}_p \leftarrow -m^{-1} \sum_{i \in \mathcal{M}} \left[r_i(\theta)\widehat{A}_i^{\mathcal{M}} - \frac{c_i^{\mathcal{M}}}{2\epsilon}(r_i(\theta) - 1)^2\right]$
      Compute policy entropy $\mathcal{L}_e$ and value loss $\mathcal{L}_v$
      $\mathcal{L} \leftarrow \mathcal{L}_p + c_1\mathcal{L}_v - c_2\mathcal{L}_e$
      Update θ and φ by gradient-based optimization
    end for
  end for
end while
return π_{θ^*}

C Additional Experimental Results

C.1 Algorithm Comparison

This section supplements the algorithm-comparison results in Section 6.1 by providing evaluation details, complete per-environment performance comparisons for all evaluated methods, and additional learning curves for QARPO, PPO-Clip, and PPO-Penalty on 6 MuJoCo continuous-control tasks and 12 Atari 2600 games. MuJoCo and Atari 2600 use 10 and 5 independent random seeds, respectively.

Probability of Improvement. Figure 4 in the main text further uses probability of improvement (POI) to measure QARPO's cross-task performance relative to each baseline. Following the definition of Agarwal et al. (2021b), $\Pr(\text{QARPO} > \text{Algorithm})$ is the probability that QARPO outperforms the corresponding baseline under the performance distribution across environments and random seeds. We compute QARPO's POI relative to each baseline separately for MuJoCo and Atari 2600 to avoid mixing the effects of different return structures and numbers of random seeds across task domains.

Aggregate Metrics. To reduce the impact of reward-scale differences across environments on cross-task statistics, we first normalize the rewards in each environment. For Atari 2600, we use the human-normalized score defined by Agarwal et al. (2021b), with random-policy and human reference scores taken from the Atari benchmark results reported by Mnih et al. (2015). For MuJoCo continuous-control tasks, we normalize using PPO-Clip as the reference algorithm. Following the evaluation protocol of Agarwal et al. (2021b), we then report four aggregate metrics: Median, Interquartile Mean (IQM), Mean, and Optimality Gap. The aggregate statistics for MuJoCo and Atari 2600 are presented in Figures 5 and Figures 6 in the main text. Specifically, the MuJoCo and Atari normalized scores are

$$\widetilde{R}_{e,j}^{a,\text{M}} = \frac{R_{e,j}^{a} - R_{e,\min}^{\text{PPO}}}{R_{e,\max}^{\text{PPO}} - R_{e,\min}^{\text{PPO}}}, \qquad \widetilde{R}_{e,j}^{a,\text{A}} = \frac{R_{e,j}^{a} - R_{e}^{\text{random}}}{R_{e}^{\text{human}} - R_{e}^{\text{random}}}. \tag{C.1}$$

Here, $R_{e,j}^{a}$ is the return score of algorithm a on environment e with seed j, and the superscripts M and A denote MuJoCo and Atari. The PPO-Clip reference values $R_{e,\min}^{\text{PPO}}$ and $R_{e,\max}^{\text{PPO}}$ are the minimum and maximum over all recorded training returns and seeds in environment e; R_{e}^{random} and R_{e}^{human} are the Atari random-policy and human reference scores. Normalized scores are not clipped to $[0, 1]$.

The aggregate metrics use these normalized scores; IQM averages the middle 50% of the pooled task–seed scores.

Per-Environment Learning Curves. In addition to cross-task aggregate statistics, Figures C.1 and C.2 provide learning curves for representative algorithms on 6 MuJoCo tasks and 12 Atari 2600 games, respectively, showing the training progress as a function of environment interaction steps. Each curve aggregates results from independent random seeds: solid lines denote mean returns, and shaded regions represent performance variability across random seeds. The final returns reported in Table C.1 use the same statistical convention as these learning curves.

Table C.1: Performance comparison of QARPO and baseline methods on 6 MuJoCo continuous-control tasks and 12 Atari 2600 games. Atari tasks are selected with reference to prior subset-evaluation research Aitchison et al. (2023). MuJoCo and Atari use 10 and 5 independent random seeds. Each reported value is obtained by averaging returns over the last 10% of training steps for each random seed and then averaging across seeds. Bold numbers indicate the highest return, while underlined numbers indicate the second-highest return.

Environment	PPO-Clip	PPO-Penalty	PPO-RB	TR-PPO	SPO	QARPO
Ant-v4	5,129.3	4,305.0	5,109.0	4,943.1	4,582.8	**5,389.6**
HalfCheetah-v4	3,807.9	2,731.6	2,574.2	3,160.6	3,581.1	**4,542.5**
Hopper-v4	1,256.8	1,477.0	1,248.1	1,366.0	1,545.3	**1,663.6**
Humanoid-v4	827.8	1,828.2	889.9	779.0	2,887.2	**3,172.9**
HumanoidStandup-v4	134,046.3	160,911.5	175,119.2	146,118.5	153,820.4	**176,303.8**
Walker2d-v4	2,963.3	3,128.2	**3,764.0**	3,260.7	3,156.2	3,603.0
Amidar	661.9	279.3	706.2	519.4	543.6	**777.8**
BattleZone	15,160.1	2,823.7	17,588.4	20,110.5	25,383.6	**29,061.5**
Bowling	43.7	26.9	52.2	54.9	55.9	**66.2**
Breakout	408.9	24.8	401.0	360.0	382.1	**429.4**
DoubleDunk	-1.7	-1.8	-1.9	-2.2	-2.5	**-1.3**
Frostbite	287.4	270.8	288.3	285.9	**312.5**	302.3
KungFuMaster	26,502.1	22,762.8	28,993.7	23,931.5	29,697.6	**35,184.1**
NameThisGame	5,282.7	5,089.9	5,885.0	5,767.9	5,385.6	**6,118.7**
Phoenix	11,147.2	4,251.1	10,518.4	9,864.7	12,317.6	**14,290.3**
Pong	20.3	18.7	20.4	20.7	20.4	**20.9**
Qbert	16,854.7	2,183.5	17,624.4	14,813.2	14,878.0	**17,754.2**
Riverraid	9,146.4	7,647.0	8,856.9	7,928.0	8,340.9	**11,097.5**

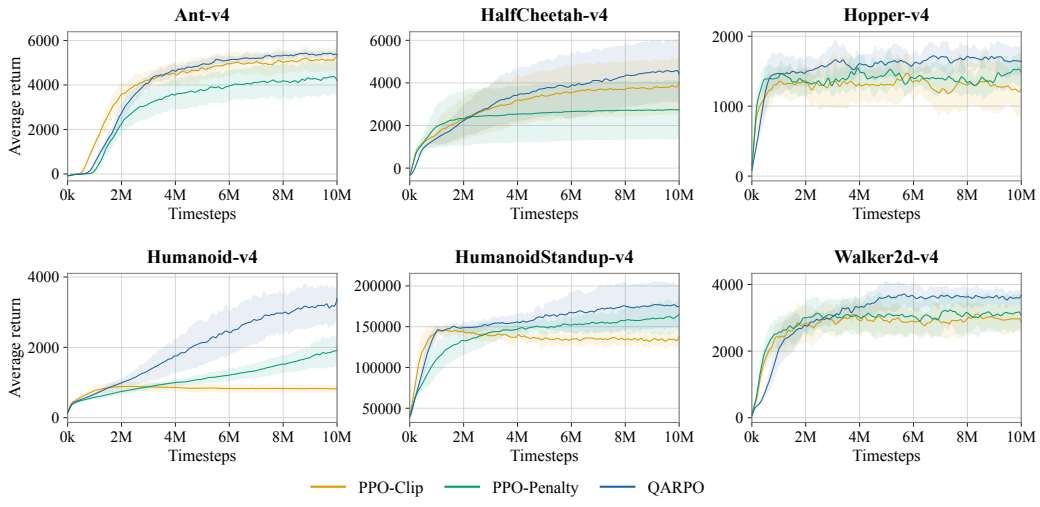


Figure C.1: Complete learning curves of the algorithms on 6 MuJoCo continuous-control tasks. Solid lines show mean returns over 10 independent random-seed runs, and shaded regions indicate performance variability across random seeds.

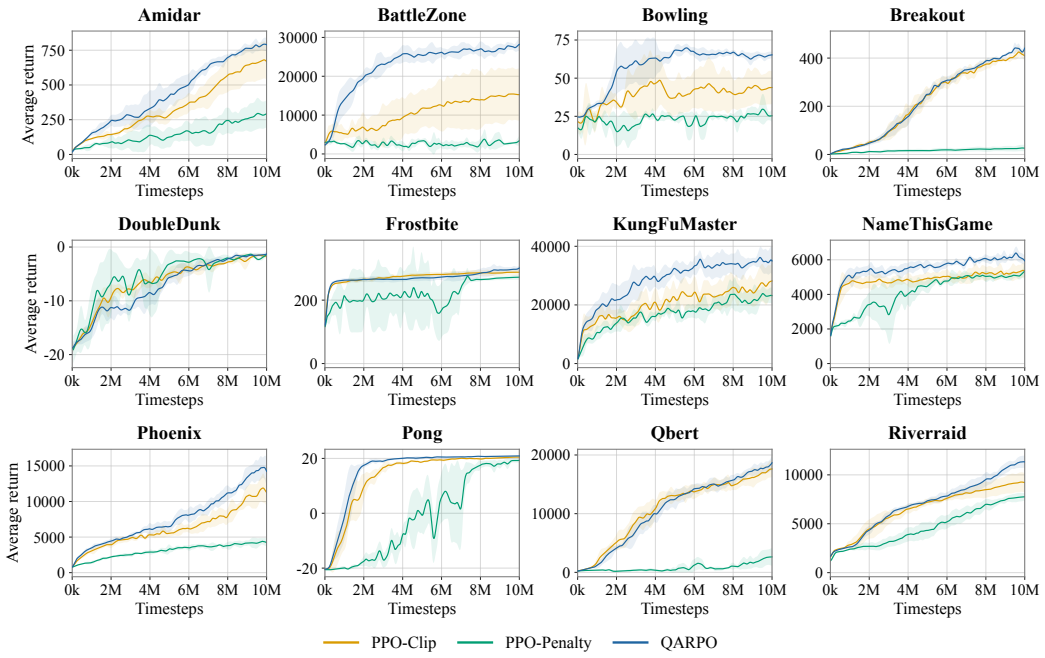


Figure C.2: Complete learning curves of the algorithms on 12 Atari 2600 games. Solid lines show mean returns over five independent random-seed runs, and shaded regions indicate performance variability across random seeds.

C.2 Ablation Studies

This section supplements the ablation results in Section 6.2. Except for the ablated factor, all experiments follow the same training configurations and evaluation protocol as the algorithm comparison. We use 5 and 3 random seeds for MuJoCo and Atari 2600, respectively. Complete results are presented below.

Curvature Structures. We compare the effects of different curvature-allocation strategies on QARPO's performance to evaluate the role of advantage-dependent curvature design in regulating responses across advantage-magnitude regions. Complete results are shown in Figures C.3 and C.4.

Table C.2 summarizes the curvature-allocation functions and their independent scalar target responses, where $c_i = \kappa_A \phi(\alpha_i)$ and $|r_i^\circ - 1| = \epsilon q(\alpha_i)$. For Constant Response and Lower Bound Curvature, the response formulas apply to $\alpha > 0$; at $\alpha = 0$, both the advantage and curvature vanish, so these cases fall outside the $c_i > 0$ assumption and the scalar objective does not select a unique target.

Table C.2: Curvature structures and corresponding independent scalar target responses. QARPO uses $0 < \lambda < 1$, with $\lambda = 1/2$ by default.

Scheme	$\phi(\alpha)$	$q(\alpha)$
Constant Response	α	1
Shared Curvature	1	α
Upper Bound Curvature	$\max(1, \alpha)$	$\min(\alpha, 1)$
Lower Bound Curvature	$\min(1, \alpha)$	$\max(1, \alpha)$
QARPO	$1 + \lambda \max(\alpha - 1, 0)$	$\alpha / \phi(\alpha)$

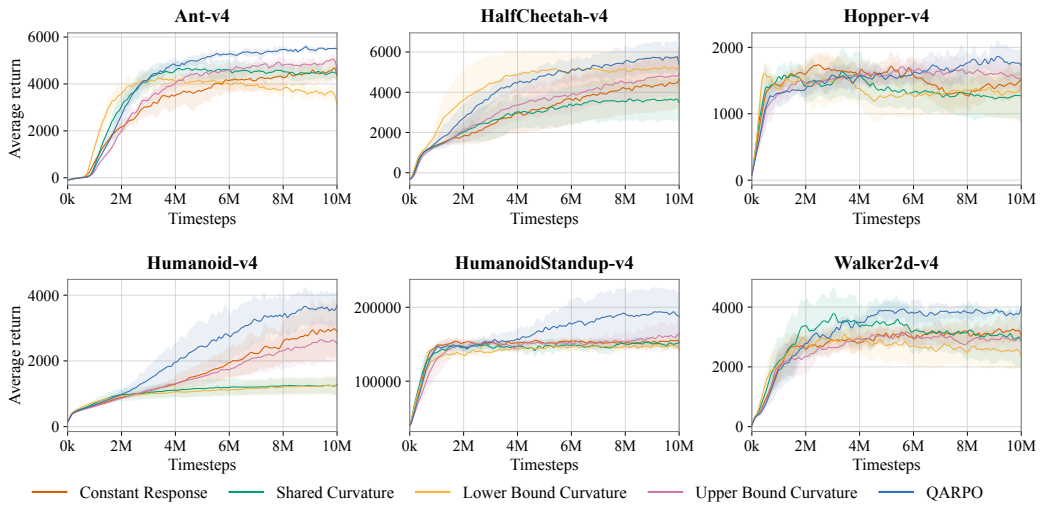


Figure C.3: Training curves for the curvature-structure ablation on 6 MuJoCo continuous-control tasks. Solid lines denote mean returns, and shaded regions indicate variability over random seeds.

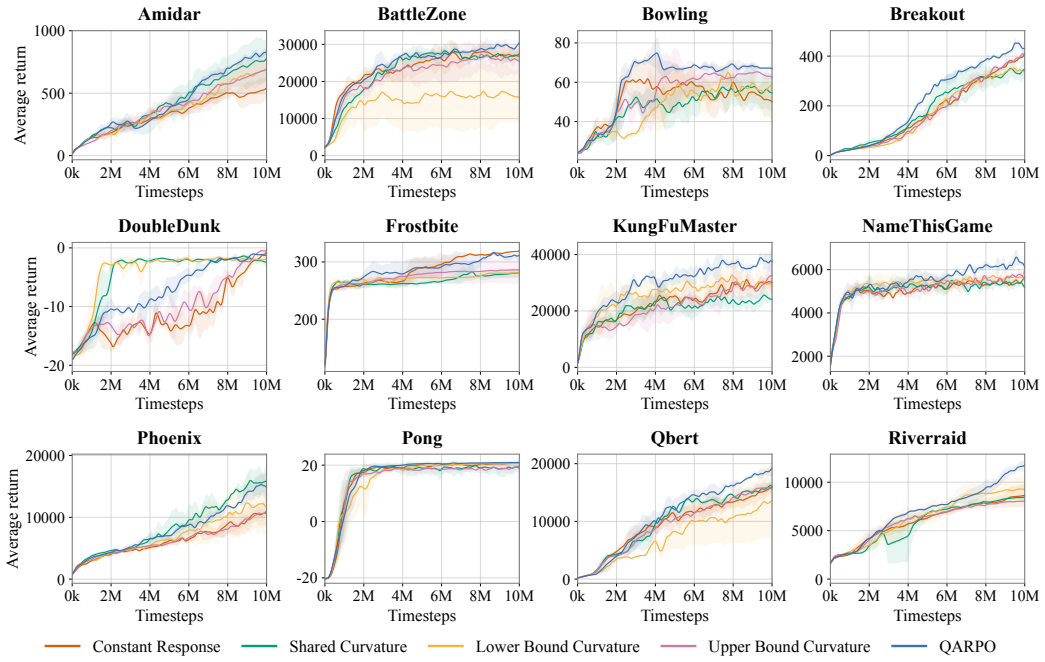


Figure C.4: Training curves for the curvature-structure ablation on 12 Atari 2600 games. Solid lines show mean returns, and shaded regions indicate variability across random seeds.

Hyperparameter Sensitivity. We analyze the sensitivity of QARPO to two key hyperparameters that control curvature allocation and response constraints, respectively. For the curvature adjustment parameter λ_{QARPO}, we evaluate $\{0.25, 0.5, 0.75\}$ to examine the effect of curvature strength. For the response-constraint scale ϵ, we test different settings to evaluate the stability under varying constraint strengths. Complete results are shown in Figures C.5, C.6, C.7, and C.8.

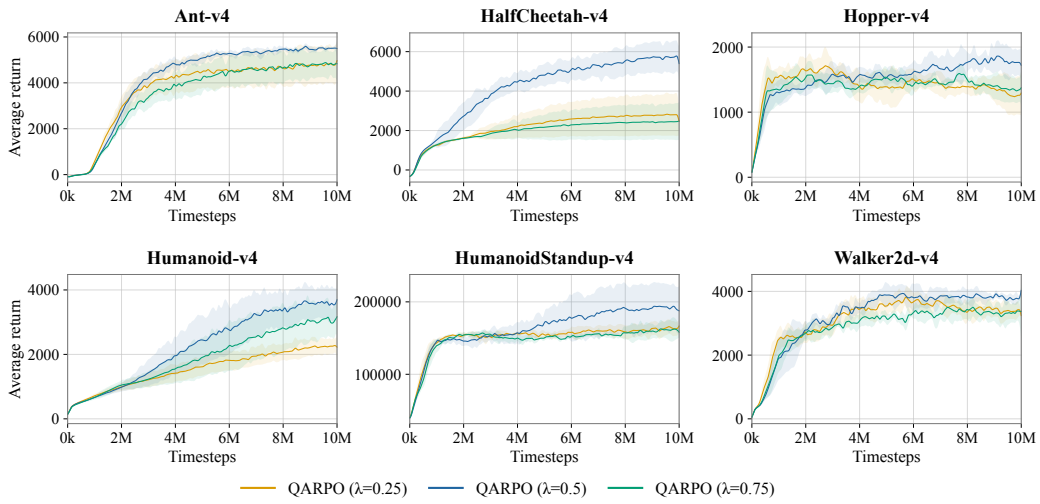


Figure C.5: Sensitivity analysis of the curvature parameter λ_{QARPO} on 6 MuJoCo continuous-control tasks. We evaluate the effect of different curvature strengths with $\lambda_{\mathrm{QARPO}} \in \{0.25, 0.5, 0.75\}$. Solid lines show mean returns, and shaded regions indicate variability across random seeds.

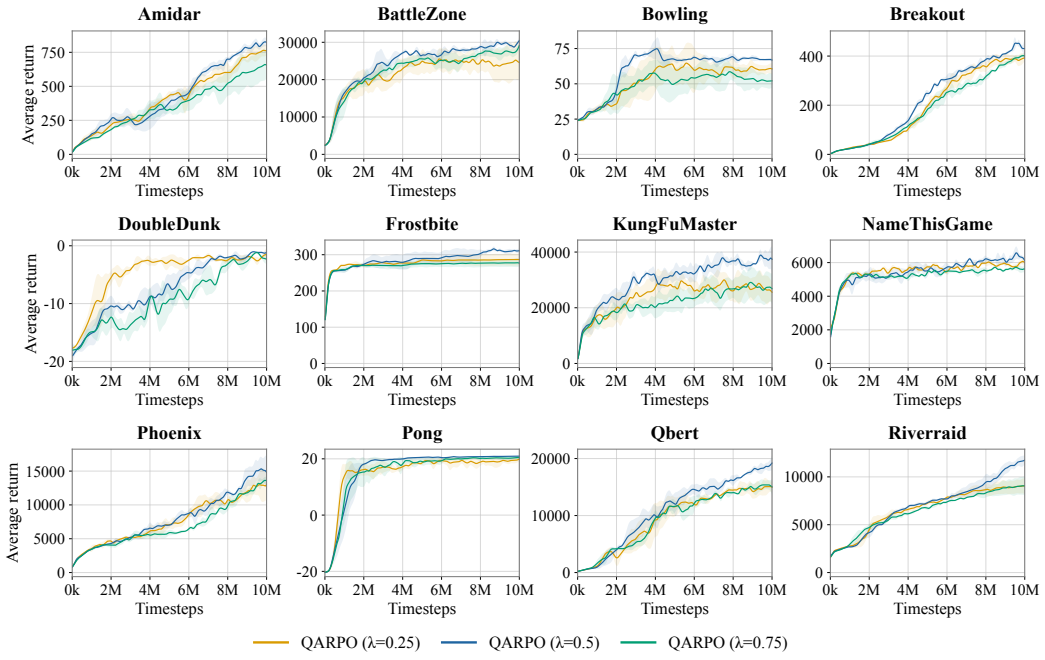


Figure C.6: Sensitivity analysis of the curvature parameter λ_{QARPO} on 12 Atari 2600 games. We evaluate QARPO under different curvature strengths with $\lambda_{\mathrm{QARPO}} \in \{0.25, 0.5, 0.75\}$. Solid lines show mean returns, and shaded regions indicate variability across random seeds.

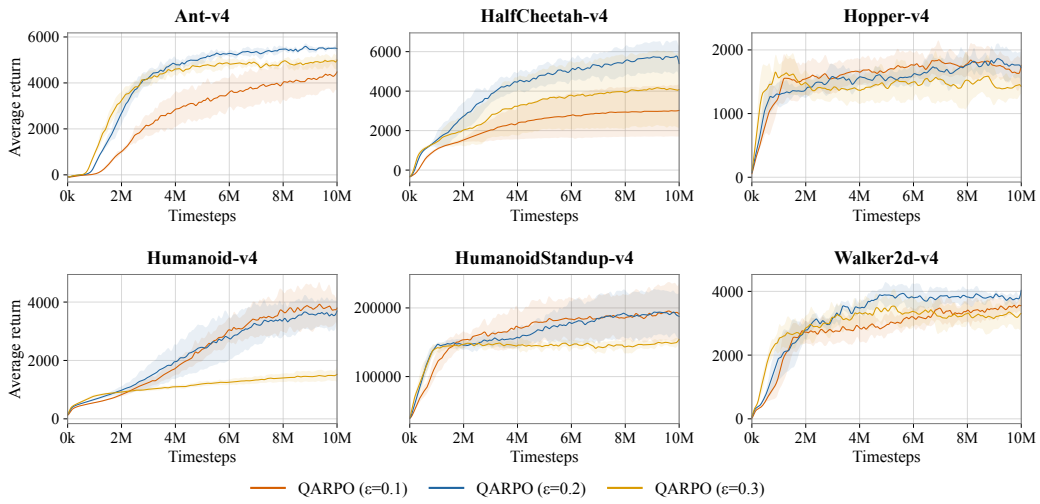


Figure C.7: Sensitivity analysis of the response-constraint scale parameter ϵ on 6 MuJoCo continuous-control tasks. We evaluate QARPO under different constraint scales with $\epsilon \in \{0.1, 0.2, 0.3\}$. Solid lines show mean returns, and shaded regions indicate variability across random seeds.

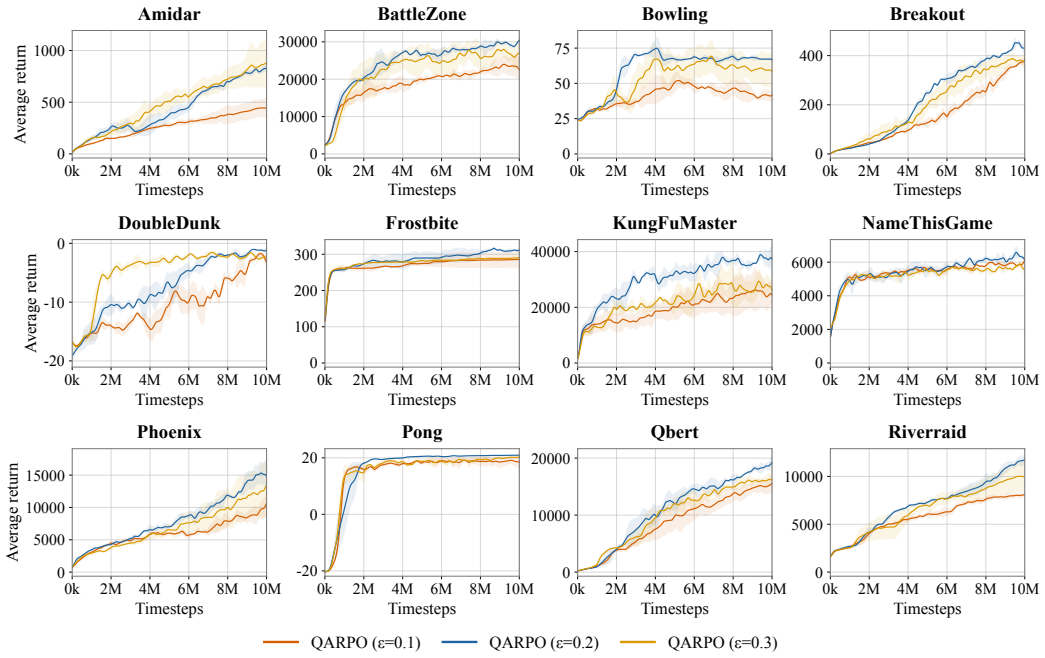


Figure C.8: Sensitivity analysis of the response-constraint scale parameter ϵ on 12 Atari 2600 games. We evaluate QARPO under different constraint scales with $\epsilon \in \{0.1, 0.2, 0.3\}$. Solid lines show mean returns, and shaded regions indicate variability across random seeds.

Curvature Functional Forms. We compare QARPO's piecewise-affine curvature with a globally affine advantage-dependent variant. Both designs retain the quadratic restoration objective and differ only in their curvature allocation. Figures C.10 and C.11 report the corresponding learning curves.

Specifically, let the relative advantage magnitude be $\alpha_i = |\widehat{A}_i|/\kappa_A$, where $\kappa_A = \mathbb{E}_{(s,a)\sim\mu_{\text{old}}}[|\widehat{A}|]$ is the old-policy reference scale in Eq. (8). Default QARPO uses piecewise curvature according to the advantage region:

$$c_i(\lambda) = \kappa_A + \lambda \max[(|\widehat{A}_i| - \kappa_A), 0]. \tag{C.2}$$

Here, λ controls curvature growth in the high-advantage region. For $|\widehat{A}_i| < \kappa_A$, the curvature equals κ_A; for $|\widehat{A}_i| > \kappa_A$, it increases affinely with advantage magnitude. The comparison variant instead uses a single affine curvature function across all advantage magnitudes:

$$c_i^{\text{global}} = \frac{1}{2}(|\widehat{A}_i| + \kappa_A). \tag{C.3}$$

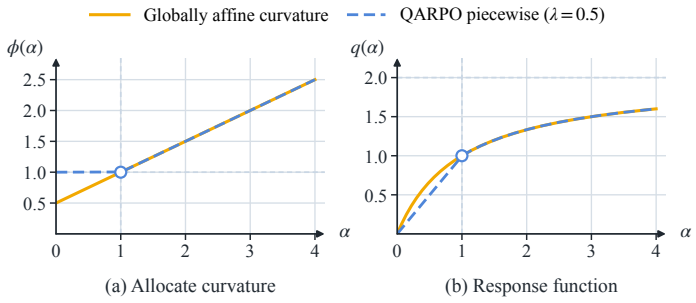
Under globally affine curvature, the corresponding quadratic objective is

$$L_{\text{global}}(\theta) = \mathbb{E}_{(s_i,a_i)\sim\mu_{\text{old}}}\left[r_i(\theta)\widehat{A}_i - \frac{(|\widehat{A}_i| + \kappa_A)}{4\epsilon}(r_i(\theta) - 1)^2\right]. \tag{C.4}$$

By the first-order optimality condition of the quadratic objective, the independent scalar target displacement under globally affine curvature can be expressed as

$$r_i^\circ = 1 + \epsilon\frac{\widehat{A}_i}{c_i^{\text{global}}}, \qquad |r_i^\circ - 1| = \epsilon\frac{|\widehat{A}_i|}{\frac{1}{2}(|\widehat{A}_i| + \kappa_A)} = \epsilon\frac{2\alpha_i}{1+\alpha_i}. \tag{C.5}$$

This independent scalar target response increases with advantage magnitude and approaches 2ϵ. At the default $\lambda = 0.5$, QARPO has the same target response for $\alpha_i \geq 1$, but a smaller target displacement for $0 < \alpha_i < 1$: $\epsilon\alpha_i < 2\epsilon\alpha_i/(1+\alpha_i)$. The comparison therefore isolates the effect of the weak-advantage curvature floor on learning performance.



(a) Allocate curvature (b) Response function

Figure C.9: Comparison of QARPO's piecewise-affine curvature and the globally affine variant. At $\lambda = 0.5$, the designs coincide for $\alpha \geq 1$ and differ for $0 \leq \alpha < 1$. (a) Curvature-allocation functions. (b) Corresponding independent scalar target response functions.

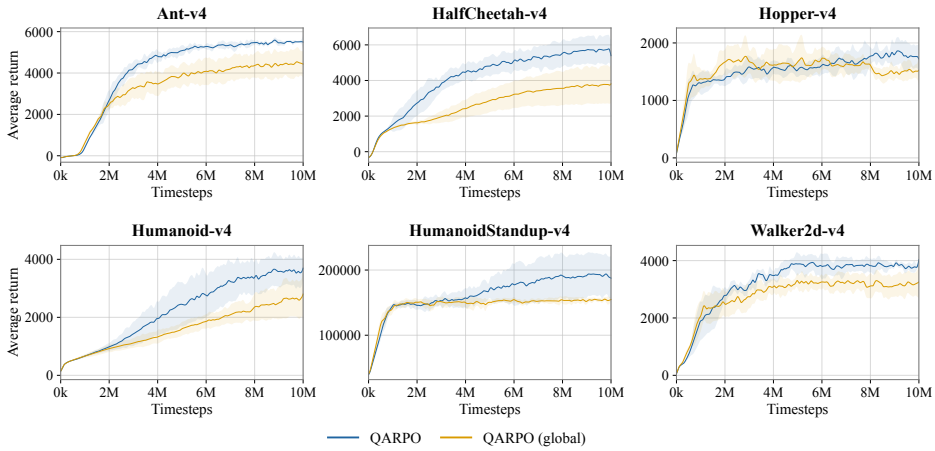


Figure C.10: Ablation of the globally affine curvature form on 6 MuJoCo continuous-control tasks. Solid lines show mean returns, and shaded regions indicate variability across random seeds.

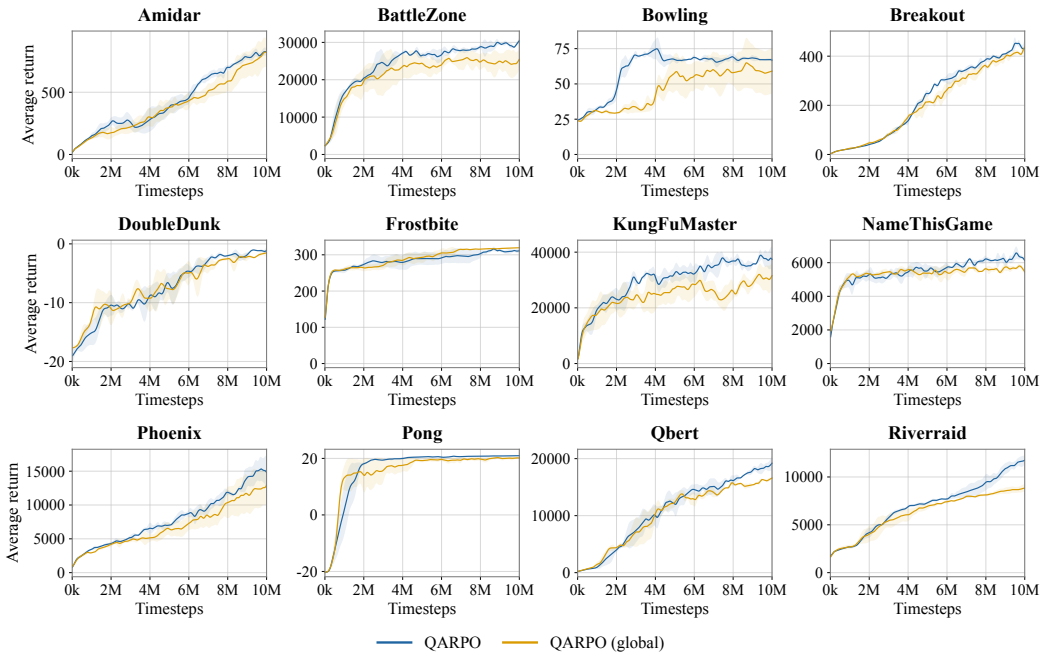


Figure C.11: Ablation of the globally affine curvature form on 12 Atari 2600 games. We compare QARPO with piecewise-affine and globally affine curvature designs. Solid lines show mean returns, and shaded regions indicate variability across random seeds.