FROM TANGENT GEOMETRY TO FINITE-STEP CURVATURE IN SAFE POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Constrained trust-region methods choose updates using first-order tangent models, but evaluate finite policy steps with nonlinear reward and cost surrogates. We characterize exactly what the tangent model determines and how curvature first alters that decision at finite step sizes. With m linearized constraints, the Fisher trust region's complete reward–constraint score image is preserved by $\mathcal{S}_m = \mathrm{range}(H^{-1}[g, b_1, \ldots, b_m])$, the unique smallest linear subspace with this property. Reward–Constraint Subspace (RCS) uses this tangent-exact domain to reoptimize projected reward and constraint curvature, then applies empirical-KL correction and nonlinear same-batch checks before committing an update. We also quantify the approximation from restricting updates to $\mathcal{S}_m$: along a regular small-radius branch, the full-space and restricted solutions share the same $O(\sqrt{\delta})$ tangent term and first differ by an $O(\delta)$ Fisher-orthogonal displacement governed by cross-domain active-Lagrangian curvature. The resulting reward-surrogate gap is $O(\delta^{3/2})$. Under weaker Fisher-relative remainder bounds, projection provides an $O(\delta)$ finite-surrogate comparator for arbitrary m. On Safety-Gymnasium, matched curvature diagnostics show improved finite-step model fidelity and update reorientation, while an independent nonlinear full-space reference recovers the predicted leakage magnitude, scaling, and direction.

1 INTRODUCTION

Safe reinforcement learning seeks high return while respecting cumulative cost budgets, commonly formalized as a constrained Markov decision process (CMDP) (Altman, 1999; García & Fernández, 2015). We focus on trust-region methods that place the safety constraint inside each policy update. For a single linearized constraint, CPO provides an analytic first-order KKT solution (Achiam et al., 2017). A committed neural-policy update, however, is evaluated at a finite displacement through nonlinear reward and cost surrogates. Curvature can therefore favor a different reward–cost trade-off from the tangent solution. Standard line search can shorten the step, but it cannot change its orientation.

This raises a natural transition question: *what does tangent-order constrained geometry determine exactly, and what changes first when the same decision is viewed at finite displacement?* A finite-step theory should distinguish structure that is already exact at first order from new information introduced by curvature.

With m linearized constraints, let $A = [g, b_1, \ldots, b_m]$ and $\mathcal{S}_m = \mathrm{range}(H^{-1}A)$. We show that Fisher projection onto $\mathcal{S}_m$ preserves the complete reward–constraint score image of the trust region, and that no smaller linear subspace does so. Thus $\mathcal{S}_m$ gives the exact decision domain of the tangent constrained problem, extending the usual KKT span characterization.

This reduction makes the finite-step question precise: *when a smooth constrained update moves beyond tangent order, what information can move the solution outside $\mathcal{S}_m$?* We show that the first departure is governed by cross-domain active Lagrangian curvature rather than arbitrary Hessian structure.

The tangent geometry then serves two complementary roles. First, $\mathcal{S}_m$ gives an exact low-dimensional representation of the first-order constrained decision, and we use it as the optimization domain of *Reward–Constraint Subspace* (RCS). Within this domain, RCS reoptimizes projected reward and

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.

constraint curvature, then applies empirical-KL correction and nonlinear same-batch checks before commitment. Second, we quantify the leading full-space correction omitted by the restriction to $\mathcal{S}_m$ and identify when the tangent domain remains second-order closed.

Our main contributions are:

- **An exact tangent decision domain.** We characterize $\mathcal{S}_m = \mathrm{range}(H^{-1}[g, b_1, \ldots, b_m])$ as the unique smallest linear subspace preserving the complete reward–constraint score image of the Fisher trust region. Its dimension is at most $m + 1$.
- **The leading finite-step correction outside the tangent domain.** Along a regular active branch, the full-space and $\mathcal{S}_m$-restricted solutions share the same $O(\sqrt{\delta})$ tangent term. They first differ by an $O(\delta)$ Fisher-orthogonal displacement determined by cross-domain active-Lagrangian curvature, producing an $O(\delta^{3/2})$ reward-loss surrogate gap. This gives a precise second-order closure condition and identifies cross-domain coupling—rather than global Hessian rank—as the relevant structural quantity. Under weaker remainder assumptions, Fisher projection also gives an $O(\delta)$ comparator for arbitrary m.
- **A curvature-corrected optimizer with matched mechanism tests.** RCS reoptimizes projected reward and constraint curvature in the score-preserving domain, then applies empirical-KL and nonlinear same-batch commitment checks. Matched curvature interventions test finite-step model fidelity and update reorientation, and an independently computed nonlinear full-space reference tests the predicted leakage magnitude, scaling, and direction.

2 BACKGROUND

We consider a constrained Markov decision process (CMDP) (Altman, 1999), defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \rho_0, r, \{\ell_i\}_{i=1}^m, \gamma, \{\bar{d}_i\}_{i=1}^m)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s'|s,a)$ is the transition kernel, and ρ_0 is the initial-state distribution. The function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the reward, while $\ell_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ specifies the ith safety-related cost. The discount factor is $\gamma \in (0, 1)$, and each $\bar{d}_i$ sets an upper budget on the corresponding expected cumulative cost. A stochastic policy $\pi_\theta(a|s)$ induces trajectories $\tau = (s_0, a_0, s_1, a_1, \ldots)$ according to $s_0 \sim \rho_0$, $a_t \sim \pi_\theta(\cdot|s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$.

For any signal $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define its discounted return under π_θ as

$$J_f(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right]. \tag{1}$$

The reward and cost returns are $J_R(\theta) \coloneqq J_r(\theta)$ and $J_{C_i}(\theta) \coloneqq J_{\ell_i}(\theta)$, respectively. Safe policy optimization seeks to maximize reward while satisfying all cost budgets,

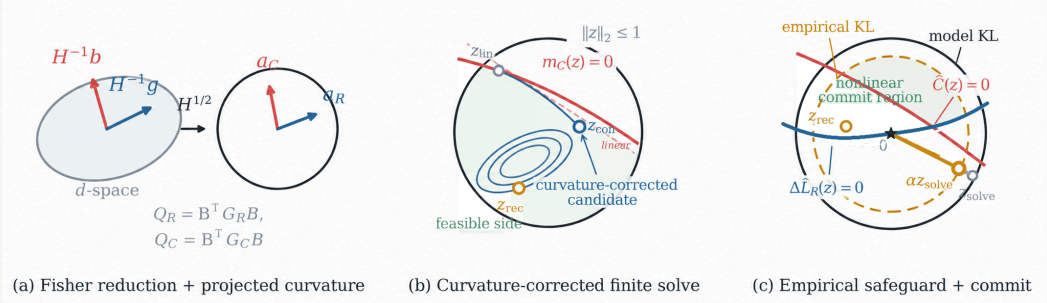
$$\max_\theta \; J_R(\theta) \qquad \text{s.t.} \qquad J_{C_i}(\theta) \leq \bar{d}_i, \quad i = 1, \ldots, m. \tag{2}$$

For the same signal f, let $V_f^\pi(s)$ and $Q_f^\pi(s,a)$ denote its state- and action-value functions, with advantage $A_f^\pi(s,a) = Q_f^\pi(s,a) - V_f^\pi(s)$. Reward quantities correspond to $f = r$, while the ith cost quantities correspond to $f = \ell_i$. These quantities provide the local reward and cost information used to construct policy-update surrogates.

At the current policy θ_k, let d denote a parameter displacement. We use a minimization convention: $\widehat{L}_R(d)$ is the reward-loss surrogate (the negative reward-advantage surrogate), and $\widehat{C}_i(d)$ is the residual surrogate for constraint i. Let $g = \nabla \widehat{L}_R(0)$, $b_i = \nabla \widehat{C}_i(0)$, and $c_i = \widehat{C}_i(0)$. Thus minimizing $g^\top d$ corresponds locally to maximizing reward. In the reported implementation, the offset is $c_{i,k} = (\widetilde{J}_{C_i,k} - \bar{d}_i)(1 - \gamma)$, where $\widetilde{J}_{C_i,k}$ is the conservative episodic-cost statistic detailed in Appendix B.

Let $H \succ 0$ denote the local Fisher information matrix. For sufficiently small policy displacements, $D_{\mathrm{KL}}(\pi_{\theta_k} \,\|\, \pi_{\theta_k + d}) \approx \frac{1}{2} d^\top H d$. Restricting the update to a Fisher trust region of radius δ gives

$$\min_d \; g^\top d \quad \text{s.t.} \quad c_i + b_i^\top d \leq 0, \quad i = 1, \ldots, m, \qquad \frac{1}{2} d^\top H d \leq \delta. \tag{3}$$



(a) Fisher reduction + projected curvature (b) Curvature-corrected finite solve (c) Empirical safeguard + commit

Figure 1: RCS update pipeline: Fisher-subspace reduction, projected-curvature reoptimization in the score-preserving tangent domain, followed by empirical-KL and nonlinear checks before commitment. Curvature can change the orientation of a finite step, not only its length.

This local model couples the reward and constraint gradients through the Fisher metric: g and $\{b_i\}_{i=1}^m$ determine the first-order reward–constraint trade-offs, while H determines the geometry and scale over which those trade-offs are realized.

3 Methodology

The methodology has three parts. Section 3.1 establishes the lossless Fisher decision reduction; Section 3.2 develops the curvature-corrected RCS update and finite-step commitment rule; and Section 3.3 characterizes the leading full-space departure and the associated comparator guarantees.

3.1 Lossless Fisher decision domain

For the local trust-region model in Eq. (3), define

$$A = [g, b_1, \dots, b_m], \qquad \mathcal{S}_m = \mathrm{range}(H^{-1}A). \tag{4}$$

Proposition 1 (Lossless Fisher decision reduction)**.** *For A and $\mathcal{S}_m$ in Eq. (4),*

$$\left\{A^\top d : \tfrac{1}{2} d^\top H d \le \delta\right\} = \left\{A^\top d : d \in \mathcal{S}_m,\ \tfrac{1}{2} d^\top H d \le \delta\right\}. \tag{5}$$

Moreover, with $r_A = \mathrm{rank}(H^{-1/2}A)$, $\dim(\mathcal{S}_m) = r_A \le m + 1$.

Corollary 1 (Unique smallest preserving subspace)**.** *For $\delta > 0$, a linear subspace V preserves the complete score image in Eq. (5) if and only if $\mathcal{S}_m \subseteq V$. Hence $\mathcal{S}_m$ is the unique smallest score-preserving linear subspace, and its dimension r_A is minimal.*

A Fisher-whitened projection proves both statements (Appendix A). The score vector records exactly the objective and constraint changes available to the tangent problem, so Eq. (5) identifies the part of the decision that is already exact at first order. Section 3.3 then asks what additional information allows a finite-step solution to leave this domain.

To make the construction concrete, we first present the RCS update for a single constraint, where the decision domain has dimension at most two. Let g be the reward-loss gradient, b the gradient of the local cost residual, c the current residual, and $H \succ 0$ the damped empirical Fisher operator. The first-order model is

$$\min_d\ g^\top d \quad \text{s.t.} \quad c + b^\top d \le 0, \qquad \frac{1}{2} d^\top H d \le \delta. \tag{6}$$

Its score-relevant subspace is $\mathcal{S}_1 = \mathrm{span}\{H^{-1}g, H^{-1}b\}$. Unless stated otherwise, we consider the generic case in which the reward and cost gradients are nonzero and noncollinear; Appendix A records the degenerate limits. The multi-constraint update follows the same construction with $\mathcal{S}_m$ and is instantiated in Section 4.3.

3.2 CURVATURE-CORRECTED RCS UPDATE AND FINITE-STEP COMMITMENT

Let $B \in \mathbb{R}^{p\times k}$, $k \le 2$, contain an H-orthonormal basis of $\mathcal{S}_1$, so that $B^\top HB = I_k$. Parameterize the step as

$$d = \sqrt{2\delta}\, Bz, \qquad \|z\|_2 \le 1. \tag{7}$$

Let $G_R = \nabla^2 \widehat{L}_R(0)$ and $G_C = \nabla^2 \widehat{C}(0)$ be the finite-batch reward-loss and cost-residual Hessians. RCS forms only the projected blocks,

$$Q_R = B^\top G_R B, \qquad Q_C = B^\top G_C B, \tag{8}$$

never forming a full Hessian. Thus the reduced quadratic model has dimension $k \le 2$ and requires at most k Hessian–vector products per enabled surrogate. With

$$a_R = \sqrt{2\delta}\, B^\top g, \qquad a_C = \sqrt{2\delta}\, B^\top b, \tag{9}$$

the local models are

$$m_R(z) = a_R^\top z + \delta z^\top Q_R z, \tag{10}$$

$$m_C(z) = c + a_C^\top z + \delta z^\top Q_C z. \tag{11}$$

Given a numerical feasibility tolerance $\tau_C \ge 0$, RCS solves

$$\min_{\|z\|_2\le 1} m_R(z) \quad \text{s.t.} \quad m_C(z) \le \tau_C. \tag{12}$$

If no acceptable constrained candidate survives finite-step reevaluation, RCS also computes a cost-restoration candidate

$$z_{\mathrm{rec}} \in \arg\min_{\|z\|_2\le 1} m_C(z).$$

Because Q_R and Q_C may be indefinite, the reduced QCQP is generally nonconvex. The implementation uses the small reduced dimension to run deterministic multi-start SLSQP, followed by explicit feasibility checks. Appendix B gives the initializations and tolerances.

The reduced quadratic models choose candidate orientations, but they are still local approximations to a finite neural-policy displacement. RCS therefore reevaluates both constrained and restoration candidates with the original same-batch nonlinear surrogates before committing an update. For a solved subspace point z, instantiate $d = \sqrt{2\delta}Bz$ and recompute the same-batch likelihood ratios. Define

$$\widehat{A}_R(\theta) = \frac{1}{N}\sum_i \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_k}(a_i|s_i)}\widehat{A}_i^R, \quad \widehat{A}_C(\theta) = \frac{1}{N}\sum_i \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_k}(a_i|s_i)}\widehat{A}_i^C. \tag{13}$$

The finite-displacement reward-loss surrogate and cost residual used for commitment are

$$\widehat{L}_R(d) = -\widehat{A}_R(\theta_k + d), \qquad \widehat{C}(d) = c + \widehat{A}_C(\theta_k + d) - \widehat{A}_C(\theta_k). \tag{14}$$

At $d = 0$, their first derivatives give g and b, and their projected second derivatives give Q_R and Q_C.

The empirical batch KL is evaluated at the same instantiated policy. If a solved step exceeds the KL target, RCS bisects along that direction and keeps the largest empirical-KL-feasible radius. It then evaluates the KL-corrected constrained and restoration candidates using Eqs. (13)–(14). A candidate is accepted as finite-step feasible only if $\widehat{C}(d) \le \tau_C$ and the reward-loss surrogate does not deteriorate relative to the zero update. If no such candidate exists and the current residual satisfies $c > \tau_C$, RCS accepts only a candidate that strictly reduces the residual; otherwise the actor is left unchanged.

The commitment rule therefore separates proposal from acceptance: projected curvature determines candidate orientations in the lossless Fisher domain, while empirical-KL correction and nonlinear same-batch reevaluation screen them for finite-step model mismatch. These checks target update-level expected-cost control under the evaluation protocol. Setting either projected curvature block to zero yields the matched curvature ablations used in Section 4.4.

3.3 FINITE-STEP CURVATURE LEAKAGE AND COMPARATOR GUARANTEES

The first-order reduction is exact. We next ask what finite displacement adds beyond tangent order. Let $r = \sqrt{2\delta}$, $a = H^{-1/2}g$, $h_i = H^{-1/2}b_i$, and

$$\mathcal{U} = \mathrm{span}\{a, h_1, \dots, h_m\}, \qquad P = P_{\mathcal{U}}.$$

Let $G_R = \nabla^2 \widehat{L}_R(0)$ and $G_i = \nabla^2 \widehat{C}_i(0)$, with Fisher-whitened curvature

$$\widetilde{G}_R = H^{-1/2} G_R H^{-1/2}, \qquad \widetilde{G}_i = H^{-1/2} G_i H^{-1/2}.$$

To keep an active constrained trade-off nondegenerate as $r \downarrow 0$, write $c_i = r\bar{c}_i$ and parameterize $d = rH^{-1/2}u$. Assume the finite-batch surrogates are three-times continuously differentiable near the origin with locally bounded Fisher-relative third derivatives. Taylor expansion gives the scaled smooth problem

$$\begin{aligned} \min_{\|u\|_2 \le 1} \quad & F_r(u) := a^\top u + \frac{r}{2} u^\top \widetilde{G}_R u + r^2 R_R(r,u) \\ \text{s.t.} \quad & K_{i,r}(u) := \bar{c}_i + h_i^\top u + \frac{r}{2} u^\top \widetilde{G}_i u + r^2 R_i(r,u) \le 0, \quad i = 1, \dots, m, \end{aligned} \tag{15}$$

where the remainder functions and their required u-derivatives are locally bounded. Restricting u to $\mathcal{U}$ is equivalent to restricting the parameter step to $\mathcal{S}_m$. The quadratic Taylor model optimized by RCS is the special case $R_R = R_i = 0$.

Proposition 2 (Smooth finite-step curvature leakage)**.** *Assume the tangent problem at $r = 0$ has a unique solution u_0 with $\|u_0\|_2 = 1$, stable active set J, strictly positive active multipliers $\lambda_{0,j}$, trust-region multiplier $\nu_0 > 0$, and LICQ for $\{u_0, h_j : j \in J\}$. Let $u_F(r)$ be the regular local solution branch of Eq. (15) and $u_S(r)$ the corresponding branch restricted to $\mathcal{U}$. Define the active Lagrangian curvature and its Fisher-orthogonal leakage by*

$$\widetilde{G}_L := \widetilde{G}_R + \sum_{j \in J} \lambda_{0,j} \widetilde{G}_j, \qquad \ell_\perp := (I - P)\widetilde{G}_L u_0. \tag{16}$$

Then, as $r \downarrow 0$,

$$u_F(r) - u_S(r) = -\frac{r}{\nu_0} \ell_\perp + O(r^2), \tag{17}$$

and therefore, for $d_F = rH^{-1/2}u_F(r)$ and $d_S = rH^{-1/2}u_S(r)$,

$$d_F - d_S = -\frac{2\delta}{\nu_0} H^{-1/2} \ell_\perp + O(\delta^{3/2}). \tag{18}$$

Moreover, the finite-step reward-loss surrogate gap satisfies

$$\widehat{L}_R(d_S) - \widehat{L}_R(d_F) = \frac{(2\delta)^{3/2}}{2\nu_0} \|\ell_\perp\|_2^2 + O(\delta^2). \tag{19}$$

Equations (18)–(19) identify the part of curvature that matters at leading order. Relative to $\mathcal{U} \oplus \mathcal{U}^\perp$, only the cross-domain block $P_{\mathcal{U}^\perp} \widetilde{G}_L P_{\mathcal{U}}$ contributes to the leading escape, through $\ell_\perp = P_{\mathcal{U}^\perp} \widetilde{G}_L u_0$. Curvature confined to $\mathcal{U}^\perp$ has no leading-order effect, while even low-rank curvature can induce departure if it couples the tangent solution to $\mathcal{U}^\perp$. Thus the relevant structural quantity is cross-domain coupling, not global Hessian rank.

The condition $\ell_\perp = 0$ gives local *second-order closure* along u_0: the score-preserving domain remains correct through the first curvature correction. A stronger sufficient condition is local invariance, $\widetilde{G}_L \mathcal{U} \subseteq \mathcal{U}$, which removes the $O(\delta)$ leakage for every tangent direction in $\mathcal{U}$. When closure fails, Eq. (18) gives the first omitted direction and its scale. Since $\|d_F\|_H = r + O(r^2)$ on the regular full-radius branch,

$$\frac{\|d_F - d_S\|_H}{\|d_F\|_H} = \frac{\sqrt{2\delta}}{\nu_0} \|\ell_\perp\|_2 + O(\delta). \tag{20}$$

If $\ell_\perp = 0$, the displacement difference improves to $O(\delta^{3/2})$ and the reward-loss surrogate gap to $O(\delta^2)$.

Because the characterization is stated for nonlinear reward and constraint surrogates inside the local Fisher trust region, it describes the finite-step approximation introduced by restricting RCS to $\mathcal{S}_m$. The commitment correction in Section 3.2 handles the separate mismatch between the quadratic Fisher ball and realized empirical KL. Section 4.4 tests the nonlinear prediction on the original same-batch surrogates, while Appendix C.6 retains the quadratic specialization $R_R = R_i = 0$ as a complementary diagnostic.

The regular-branch result is sharper but local. For a broader comparison under weaker assumptions, valid for arbitrary candidates and arbitrary m, write

$$\widehat{L}_R(d) - \widehat{L}_R(0) = g^\top d + \varepsilon_R(d), \qquad \widehat{C}_i(d) = c_i + b_i^\top d + \varepsilon_{C_i}(d), \tag{21}$$

and assume on the Fisher ball

$$|\varepsilon_R(d)| \le \frac{\kappa_R}{2} d^\top H d, \qquad |\varepsilon_{C_i}(d)| \le \frac{\kappa_{C_i}}{2} d^\top H d. \tag{22}$$

Corollary 2 (General-m finite-step surrogate comparator)**.** *For every full-space d with $\|d\|_H^2 \le 2\delta$, its Fisher projection $\bar{d} \in \mathcal{S}_m$ preserves all linear reward and constraint scores, satisfies $\|\bar{d}\|_H \le \|d\|_H$, and obeys*

$$\widehat{L}_R(\bar{d}) \le \widehat{L}_R(d) + 2\kappa_R\delta, \qquad \widehat{C}_i(\bar{d}) \le \widehat{C}_i(d) + 2\kappa_{C_i}\delta \quad \forall i. \tag{23}$$

Proposition 2 and Corollary 2 are complementary. The proposition identifies the leading local escape direction and gives the sharper reward-loss gap scaling under regularity; the corollary provides an $O(\delta)$ comparison under weaker assumptions for arbitrary candidates and arbitrary m.

4 Experiments

We organize the experiments around three questions: **Q1 (outcome):** how RCS compares with established constrained policy-optimization methods in final-window return, cost, and cumulative training-time cost excess; **Q2 (extension):** whether the score-preserving Fisher geometry remains practical with multiple simultaneous constraints; and **Q3 (mechanism):** whether projected curvature improves finite-step surrogate fidelity and changes update orientation within $\mathcal{S}_m$, and whether independently computed full-space solutions match the predicted magnitude, scaling, and direction of the orthogonal leakage. Appendix C provides the exact protocols, aggregation rules, and numerical tables.

4.1 Evaluation design

Benchmarks and comparators. We evaluate RCS on three single-constraint Safety-Gymnasium tasks (Ji et al., 2023) spanning distinct reward–constraint interactions: `SafetyCarGoal1-v0` for goal-directed navigation, `SafetyCarPush1-v0` for object manipulation, and `SafetyHumanoidVelocity-v1` for velocity-constrained locomotion. For the multi-constraint experiment, we additionally use `SafetyCarGoal2-v0` with separate hazard and vase constraints. Environment descriptions and visualizations are provided in Appendix C.1.

We compare RCS with CPO (Achiam et al., 2017), C-TRPO (Milosevic et al., 2025), PCPO (Yang et al., 2020), CUP (Yang et al., 2022), FOCOPS (Zhang et al., 2020), P3O (Zhang et al., 2022), PPO-Lagrangian (PPO-Lag) (Ray et al., 2019), and MICE (Gao et al., 2025), all evaluated in the same archived environment/model/logging stack. Seeds, budgets, horizons, architectures, optimizer settings, and other implementation details are reported in Appendix C.2.

Metrics and scope. Cross-task comparisons report final-window return and cost together with cumulative mean-cost excess, separating final-window budget satisfaction from violations accumulated over training. These metrics assess expected-cost behavior under the evaluation protocol; they do not provide deployment-time or out-of-sample safety certification.

4.2 Cross-task return–safety trade-off

Figure 2 compares return, episodic cost, and cumulative mean-cost excess across the three tasks.

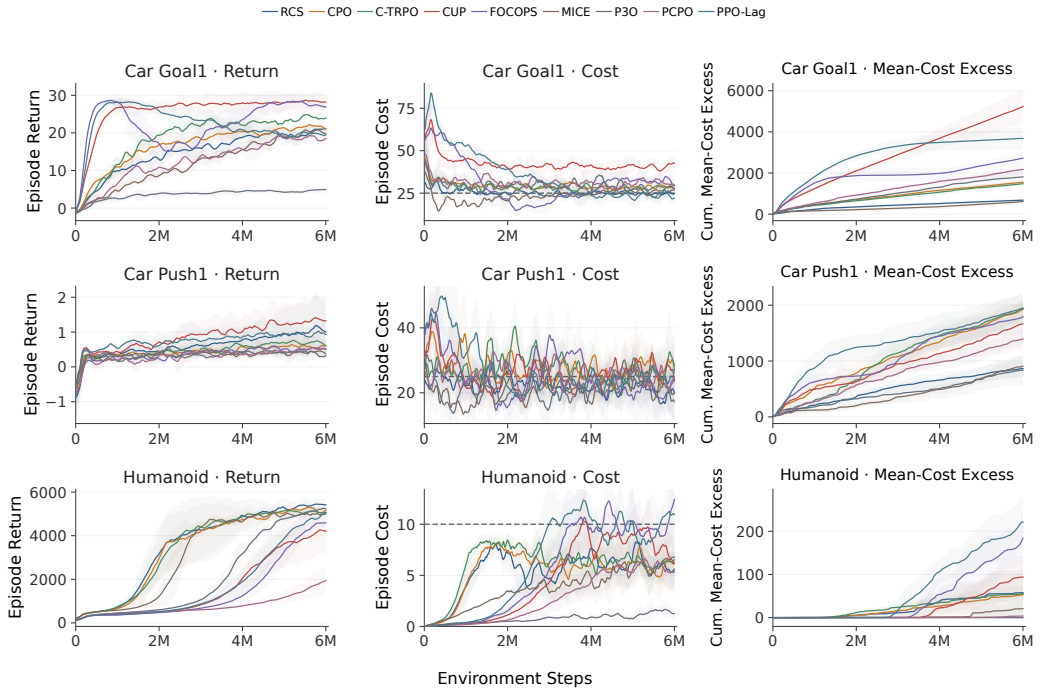


Figure 2: Cross-task return–safety trade-offs on SafetyCarGoal1-v0, SafetyCarPush1-v0, and SafetyHumanoidVelocity-v1. RCS combines competitive final-window return with mean costs near or below budget and relatively low cumulative mean-cost excess across the three constraint regimes.

Goal1: similar return with substantially less training-time overshoot. Several baselines reach higher nominal return but exceed the budget in mean final-window cost. Among high-return baselines with below-budget mean final-window cost, PPO-Lagrangian is the closest comparison: RCS achieves a similar final-window return (20.35 vs. 19.82) while reducing cumulative mean-cost excess from 3,680 to 680, or about 82%. This shows that similar final-window return and budget-satisfying mean cost can still accompany very different amounts of violation during training.

Push1: higher return without a looser mean cost. CUP reaches a higher nominal return but has mean final-window cost above budget. Among methods with below-budget mean final-window cost, RCS has the highest final-window return while keeping cumulative mean-cost excess near the low end of the group.

Humanoid: return after mean budget satisfaction. Most methods finish with mean cost below budget, so final-window budget satisfaction alone does not separate them well. Within this group, RCS has the highest final-window return while retaining substantial cost slack. P3O, by comparison, incurs no cumulative mean-cost excess but finishes at lower return. Here the main difference is how much reward is retained once mean cost is already low.

Across the three tasks, RCS achieves competitive final-window return with mean costs near or below the task budgets and relatively low cumulative mean-cost excess across different constraint regimes.

4.3 Multi-constraint geometry: an instantiation

We next instantiate the score-preserving construction with two simultaneous constraints on `SafetyCarGoal2-v0`, treating hazard and vase costs as distinct safety signals. RCS optimizes in the subspace $\mathcal{S}_2 = \mathrm{span}\{H^{-1}g, H^{-1}b_{\mathrm{hazards}}, H^{-1}b_{\mathrm{vases}}\}$, which has dimension at most three while preserving the first-order reward score and both constraint scores.

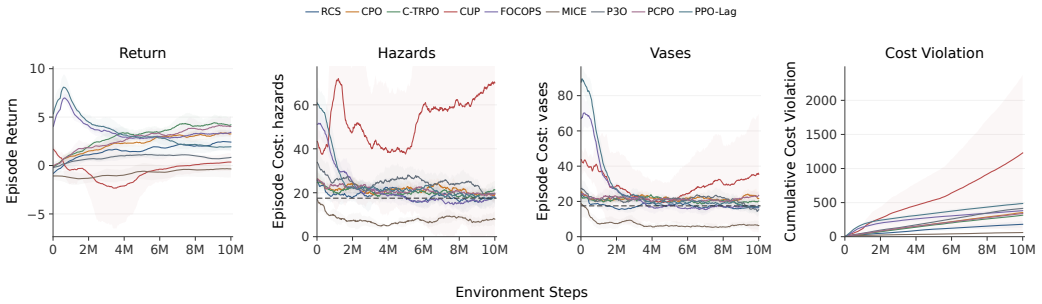


Figure 3: Multi-constraint return–safety trade-offs on SafetyCarGoal2-v0. RCS jointly handles hazard and vase constraints in the 3D score-preserving Fisher subspace, with competitive return and low cumulative component-wise mean-cost excess.

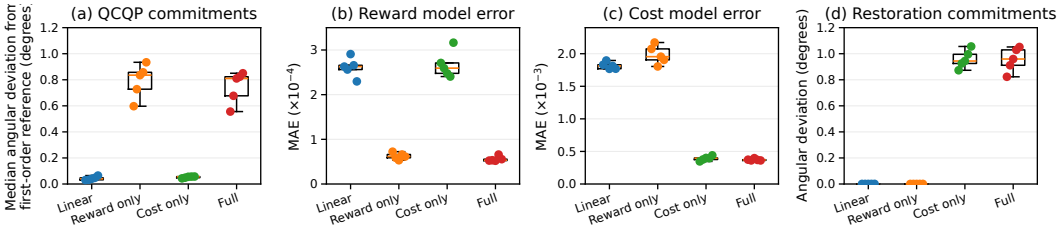


Figure 4: **Goal1 curvature mechanism under matched ablations.** Reward curvature improves reward-model fidelity and reorients QCQP commitments (a,b); cost curvature improves cost-model fidelity and reorients restoration commitments (c,d). Directional effects are roughly one degree and component-specific.

More generally, if $k = \dim(\mathcal{S}_m) \leq m + 1$, each enabled projected Hessian block requires at most k Hessian–vector products, and the reduced constrained problem has only k decision variables. The multi-constraint extension therefore scales with the number of reward and constraint directions that matter to the decision, rather than with the policy dimension.

Figure 3 shows the five-seed two-constraint results. RCS attains final-window return 2.44 ± 1.02, with final-window hazard cost 17.29 ± 4.17 and vase cost 16.55 ± 1.86 under limits of 17.5 for both constraints. Its cumulative component-wise mean-cost excess is $2{,}252 \pm 1{,}076$ for hazards and $1{,}416 \pm 146$ for vases. The methods occupy different points on the return–safety trade-off. FOCOPS reaches higher final-window return (3.35 ± 0.36) with near-budget final-window component costs, but has larger cumulative component-wise mean-cost excess. MICE keeps both component costs and cumulative excess lower, at a lower final-window return (-0.34 ± 1.00). RCS retains both constraint directions jointly in the reduced optimization rather than applying separate one-dimensional corrections.

Appendix C.3 reports the full numerical comparison and implementation details.

4.4 Mechanism test: model fidelity, reorientation, and leakage

The Goal1 factorial intervention isolates projected curvature. Full (Q_R, Q_C), Linear $(0, 0)$, Reward only $(Q_R, 0)$, and Cost only $(0, Q_C)$ use the same Fisher subspace, solver, empirical-KL correction, commitment rule, architecture, and five matched seeds.

Projected-curvature mechanism. At committed steps, Full reduces reward-model MAE from 2.61×10^{-4} (Linear) to 5.58×10^{-5} and cost-model MAE from 1.82×10^{-3} to 3.73×10^{-4}. The one-sided variants improve only the corresponding model (Appendix C.4). Figure 4 shows that these fidelity gains accompany directional changes: reward curvature reorients QCQP commitments, while cost curvature reorients restoration commitments. Relative to Linear, Full raises final-window return by 0.91, leaves final-window cost essentially unchanged (24.72 vs. 24.71), and reduces cumulative

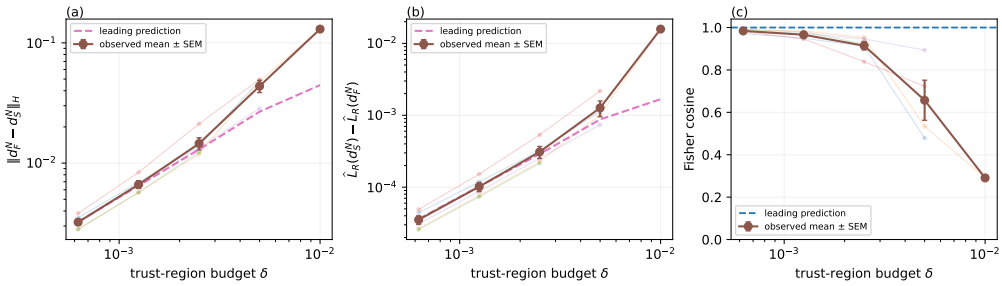


Figure 5: **Nonlinear full-space curvature leakage on SafetyCarGoal1-v0.** Observed versus leading-order displacement, nonlinear reward-loss surrogate gap, and Fisher alignment. Thin curves are certified seeds; thick curves are mean ± 1 SEM. The three smallest budgets (all five seeds certified) define the asymptotic fits; larger radii are stress tests.

mean-cost excess from 864 to 680. Both variants reject 0.6% of updates and require no empirical-KL shrinkage, so their acceptance and radial-correction rates are essentially the same.

Push1 Full-versus-Linear ablation. On SafetyCarPush1-v0, five matched seeds compare Full with a Linear variant that disables the projected reward and cost curvature blocks while leaving the rest of the update pipeline unchanged. Both variants attain final-window return 1.05; Full lowers final-window cost from 20.30 to 19.48 and cumulative mean-cost excess from $1{,}015$ to 865. Appendix C.5 reports the trajectories and final-window statistics.

Nonlinear full-space leakage diagnostic. We test Proposition 2 directly on the original same-batch likelihood-ratio surrogates. In the radius sweep, we co-scale the residual as $c_\delta = \eta\sqrt{2\delta s}$ with fixed $\eta = 0.6$, keeping the diagnostic on the regular active asymptotic family analyzed by the proposition. The independent matrix-free reference starts from the score-preserving subspace and expands only along nonlinear full-parameter KKT residuals, independently of the predicted leakage direction. All five seeds pass the local KKT checks at the three smallest budgets, which define the asymptotic fit window. The displacement exponent is 1.074 ± 0.043 SEM (prediction 1), and the reward-loss surrogate-gap exponent is 1.550 ± 0.047 (prediction $3/2$). At the smallest budget, the observed/predicted ratios are 1.000 ± 0.035 for displacement and 0.989 ± 0.053 for the gap, with Fisher cosine 0.984 ± 0.004. At larger radii, fewer seeds pass the certification checks; those points are included as stress tests. Appendix C.6 gives solver checks and the complementary quadratic reference.

5 Conclusion

We separate tangent geometry from finite-step curvature in constrained policy optimization. The Fisher score-preserving subspace $\mathcal{S}_m$ is exact for the complete first-order reward–constraint decision, while the leading full-space correction outside this domain is governed by cross-domain active Lagrangian curvature. Along a regular active branch, the correction is $O(\delta)$ in displacement and produces an $O(\delta^{3/2})$ reward-loss surrogate gap. Under weaker assumptions, Fisher projection gives a general-m $O(\delta)$ comparator.

RCS reoptimizes projected curvature in this tangent-exact domain and screens candidate steps with empirical-KL and nonlinear same-batch checks. Goal1 diagnostics connect curvature to model fidelity and reorientation, while the matched Push1 ablation preserves final-window return with lower final-window cost and cumulative mean-cost excess. An independent nonlinear full-space reference recovers the predicted small-radius leakage magnitude, scaling, and direction. Overall, the results combine an exact tangent-domain reduction with a quantified account of the structured finite-step discrepancy.

AI USE STATEMENT

Generative AI tools were used for language and structure editing, LaTeX assistance, literature retrieval and discovery, and critique of mathematical exposition. All AI-assisted edits and retrieved references were independently checked by the authors. Generative AI was not used to formulate mathematical claims or proofs, implement the method, analyze experimental data, or interpret the reported results. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix A gives the score-preserving Fisher reduction, the smooth finite-step curvature leakage perturbation proof, the general-m finite-step comparator, and the single-constraint KKT geometry and sensitivity derivations. Appendix B specifies the curvature-corrected Fisher-subspace actor update, projected-curvature switches, empirical-KL correction, commitment rule, and implementation settings. Appendix C.2 records the three-task single-constraint baseline tables, shared protocol, metric definitions, hardware convention, and visualization smoothing. Appendix C.3 reports the two-constraint Goal2 extension. Appendix C.4 reports the Goal1 factorial diagnostics. Appendix C.5 reports the matched Push1 Full-versus-Linear ablation. Appendix C.6 specifies the nonlinear and quadratic full-space references, numerical KKT checks, the one-direction enrichment test, and the fixed-η radius sweep. The supplementary material provides the experimental data underlying the reported results. Full training/evaluation code, configurations, and figure-generation scripts will be open-sourced upon acceptance.

REFERENCES

Abbas Abdolmaleki, Sandy H. Huang, Leonard Hasenclever, Michael Neunert, Francis Song, Martina Zambelli, Murilo F. Martins, Nicolas Heess, Raia Hadsell, and Martin Riedmiller. A distributional view on multi-objective policy optimization. In *International Conference on Machine Learning*, volume 119, pp. 11–22, 2020.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, volume 70, pp. 22–31, 2017.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Qinbo Bai, Amrit Singh Bedi, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 6737–6744, 2023.

Vivek S. Borkar. An actor-critic algorithm for Constrained Markov Decision Processes. *Systems & Control Letters*, 54(3):207–213, 2005.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Richard H. Byrd, Robert B. Schnabel, and Gerald A. Shultz. Approximate solution of the trust region problem by minimization over two-dimensional subspaces. *Mathematical Programming*, 40:247–263, 1988.

Juntao Dai, Yaodong Yang, Qian Zheng, and Gang Pan. Safe reinforcement learning using finite-horizon gradient-based estimation. In *International Conference on Machine Learning*, volume 235, pp. 9872–9903, 2024.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R. Jovanović. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 3304–3312, 2021.

Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Basar, and Mihailo R. Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for Constrained MDPs. *Journal of Machine Learning Research*, 26(256):1–76, 2025.

Shiqing Gao, Jiaxin Ding, Luoyi Fu, and Xinbing Wang. Controlling underestimation bias in constrained reinforcement learning for safe exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18290–18316. PMLR, 2025.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

Ather Gattami, Qinbo Bai, and Vaneet Aggarwal. Reinforcement learning for Constrained Markov Decision Processes. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 2656–2664, 2021.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Juntao Dai, and Yaodong Yang. Safety-gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pp. 267–274, 2002.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1531–1538, 2001.

Alexander Mattick, Dominik Seuß, and Christopher Mutschler. SafeMPO: Constrained reinforcement learning with probabilistic incremental improvement. In *International Conference on Learning Representations*, 2026.

Nikola Milosevic, Johannes Müller, and Nico Scherf. Embedding safety into RL: A new take on trust region methods. In *International Conference on Machine Learning*, volume 267, pp. 44199–44224, 2025.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. ReLOAD: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained MDPs. In *International Conference on Machine Learning*, volume 202, pp. 25303–25336, 2023.

Santiago Paternain, Luiz F. O. Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7555–7565, 2019.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *International Conference on Machine Learning*, volume 28, pp. 307–315, 2013.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. 2019.

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

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, volume 119, pp. 9133–9143, 2020.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, volume 139, pp. 11480–11491, 2021.

Long Yang, Jiaming Ji, Juntao Dai, Linrui Zhang, Binbin Zhou, Pengfei Li, Yaodong Yang, and Gang Pan. Constrained update projection approach to safe policy optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2020.

Yujie Yang, Jinglin Teh, Ziyu Lin, Kaicheng Yu, Tao Zhang, and Shengbo Li. Breaking safety paradox with feasible dual policy iteration. In *International Conference on Learning Representations*, 2026.

Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Xueqian Wang, Bo Yuan, and Dacheng Tao. Penalized proximal policy optimization for safe reinforcement learning. In *International Joint Conference on Artificial Intelligence*, pp. 3744–3750, 2022.

Yiming Zhang, Quan Vuong, and Keith W. Ross. First order constrained optimization in policy space. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15338–15349, 2020.

A Geometric Derivations

A.1 Lossless Fisher decision reduction

Proof of Proposition 1 and Corollary 1. Let $M = H^{-1/2}A$, $z = H^{1/2}d$, and let P_m project orthogonally onto $\mathcal{U}_m = \mathrm{range}(M)$. Since $M^\top(I - P_m) = 0$ and P_m is an orthogonal projector, the parameter-space projection $\Pi_m = H^{-1/2}P_m H^{1/2}$ satisfies

$$A^\top \Pi_m d = M^\top P_m z = M^\top z = A^\top d, \qquad \|\Pi_m d\|_H^2 = \|P_m z\|_2^2 \le \|z\|_2^2 = \|d\|_H^2. \tag{A.1}$$

Moreover, $\mathrm{range}(\Pi_m) = H^{-1/2}\mathcal{U}_m = \mathcal{S}_m$. Thus every point in the Fisher ball has a score-equivalent representative in $\mathcal{S}_m$ with no larger norm, while the reverse score-image inclusion follows from $\mathcal{S}_m \subseteq \mathbb{R}^p$. This proves Eq. (5), including the existence of a reduced-domain optimizer of the tangent problem. Invertibility of H gives

$$\dim(\mathcal{S}_m) = \mathrm{rank}(H^{-1}A) = \mathrm{rank}(H^{-1/2}A) = r_A \le m + 1.$$

For minimality, suppose a linear subspace V preserves that score image for $\delta > 0$, and write P_V^H for its H-orthogonal projector. The support functions of the full and V-restricted images must agree for every $w \in \mathbb{R}^{m+1}$:

$$\sqrt{2\delta}\,\|H^{-1}Aw\|_H = \max_{\|d\|_H^2 \le 2\delta} w^\top A^\top d = \max_{\substack{d \in V \\ \|d\|_H^2 \le 2\delta}} w^\top A^\top d = \sqrt{2\delta}\,\|P_V^H H^{-1}Aw\|_H. \tag{A.2}$$

Equality of the norms of a vector and its orthogonal projection forces $H^{-1}Aw \in V$. As this holds for all w, $\mathcal{S}_m \subseteq V$. Conversely, any V containing $\mathcal{S}_m$ preserves the full score image by the projection identity above. Hence $\mathcal{S}_m$ is the unique smallest preserving linear subspace.

A.2 Proof of the curvature leakage expansion

Proof of Proposition 2. The C^3 assumption gives Eq. (15) with locally bounded remainder derivatives. Fix the stable active set $J = \{j_1, \dots, j_k\}$ and let $\mu = (\lambda_{j_1}, \dots, \lambda_{j_k}, \nu)^\top$. The full-branch KKT equations, with the ball constraint normalized as $\frac{1}{2}(\|u\|_2^2 - 1) = 0$, are

$$\begin{aligned} 0 &= \nabla_u F_r(u) + \sum_{j \in J} \lambda_j \nabla_u K_{j,r}(u) + \nu u, \\ 0 &= K_{j,r}(u) \quad (j \in J), \qquad 0 = \tfrac{1}{2}(\|u\|_2^2 - 1). \end{aligned} \tag{A.3}$$

At $r = 0$, stationarity reads $a + \sum_{j \in J} \lambda_{0,j} h_j + \nu_0 u_0 = 0$, so $u_0 \in \mathcal{U}$. Introduce

$$T = [h_{j_1} \quad \cdots \quad h_{j_k} \quad u_0], \qquad \beta = \begin{pmatrix} \frac{1}{2}u_0^\top \widetilde{G}_{j_1} u_0 \\ \vdots \\ \frac{1}{2}u_0^\top \widetilde{G}_{j_k} u_0 \\ 0 \end{pmatrix}.$$

LICQ makes T full column rank. The Jacobian of the active KKT equations with respect to (u, μ) at $r = 0$ is

$$\mathcal{K}_0 = \begin{pmatrix} \nu_0 I & T \\ T^\top & 0 \end{pmatrix}.$$

It is nonsingular: if $\nu_0 v + Tw = 0$ and $T^\top v = 0$, taking the inner product with v gives $\nu_0 \|v\|_2^2 = 0$, whence $v = 0$ and then $w = 0$. The implicit-function theorem thus yields smooth full and $\mathcal{U}$-restricted branches with the common tangent point u_0.

Write a dot for $d/dr|_{r=0}$. Differentiating the KKT equations removes every $r^2 R$ term and gives the sensitivity system

$$\begin{pmatrix} \nu_0 I & T \\ T^\top & 0 \end{pmatrix} \begin{pmatrix} \dot{u}_F \\ \dot{\mu}_F \end{pmatrix} = -\begin{pmatrix} \widetilde{G}_L u_0 \\ \beta \end{pmatrix}, \qquad \widetilde{G}_L = \widetilde{G}_R + \sum_{j \in J} \lambda_{0,j} \widetilde{G}_j. \tag{A.4}$$

All columns of T lie in $\mathcal{U}$. Applying $I - P$ to the first block row, and P to the same row together with the second block row, yields respectively

$$(I - P)\dot{u}_F = -\nu_0^{-1}(I - P)\widetilde{G}_L u_0 = -\nu_0^{-1}\ell_\perp, \tag{A.5}$$

$$P\dot{u}_F = \dot{u}_S. \tag{A.6}$$

For the second identity, the projected equations are precisely the restricted-branch sensitivity system; its solution is unique by the same nonsingularity argument on $\mathcal{U}$. Consequently,

$$\nu_0\big(u_F(r) - u_S(r)\big) = -r\ell_\perp + O(r^2). \tag{A.7}$$

This proves Eq. (17); multiplying by $rH^{-1/2}$ and using $r^2 = 2\delta$ proves Eq. (18).

For the value gap, set $\Delta_r = u_S(r) - u_F(r) = r\nu_0^{-1}\ell_\perp + O(r^2)$ and define the full-branch Lagrangian with its multipliers held fixed,

$$\mathcal{L}_r(u) = F_r(u) + \sum_{j\in J} \lambda_{F,j}(r)K_{j,r}(u) + \frac{\nu_F(r)}{2}(\|u\|_2^2 - 1).$$

Both branches satisfy the active cost equalities and the ball equality. Hence their objective difference equals their Lagrangian difference, while full-branch stationarity and the scaled Taylor expansion give

$$\begin{aligned} F_r(u_S) - F_r(u_F) &= \mathcal{L}_r(u_F + \Delta_r) - \mathcal{L}_r(u_F) \\ &= \tfrac{1}{2}\Delta_r^\top\big[\nu_0 I + O(r)\big]\Delta_r + O(\|\Delta_r\|_2^3) \\ &= \frac{r^2}{2\nu_0}\|\ell_\perp\|_2^2 + O(r^3). \end{aligned} \tag{A.8}$$

Finally, $\widehat{L}_R(rH^{-1/2}u) - \widehat{L}_R(0) = rF_r(u)$; multiplication by $r = \sqrt{2\delta}$ gives Eq. (19).

Proof of Corollary 2. Take $\bar{d} = \Pi_m d$. The projection identities above preserve every linear score and satisfy $\|\bar{d}\|_H \le \|d\|_H$. Under Eq. (22), each remainder at d and $\bar{d}$ has magnitude at most its corresponding $\kappa\delta$, so their difference is at most $2\kappa\delta$. Applying this separately to reward and every constraint gives Eq. (23).

For the single-constraint case, $m = 1$, write $h = h_1$ and $P = P_1$. The whitened model is the linear objective $a^\top z$ with affine residual $c + h^\top z$ inside the ball $\|z\|_2^2 \le 2\delta$. The general projection above reduces z to $\mathrm{span}\{a, h\}$ without changing either linear score. Mapping back gives

$$d^\star \in \mathrm{span}\{H^{-1}g, H^{-1}b\}. \tag{A.9}$$

This is the reward–cost plane used by the remainder of the single-constraint analysis. The statement guarantees a score-equivalent representative and the existence of an optimizer in the span; finite-displacement effects are handled by the comparator analysis below.

A.3 Ordering of the KKT family

For a single constraint define

$$q = g^\top H^{-1}g, \qquad r_b = b^\top H^{-1}g, \qquad s = b^\top H^{-1}b, \qquad \chi = \frac{r_b}{\sqrt{qs}} \in (-1, 1),$$

and let $\mu = \lambda\sqrt{s/q}$, $\hat{g} = H^{-1/2}g/\sqrt{q}$, and $\hat{b} = H^{-1/2}b/\sqrt{s}$. The full-radius stationary family is

$$d_\mu = -\sqrt{2\delta}\,\frac{H^{-1/2}(\hat{g} + \mu\hat{b})}{\sqrt{1 + 2\chi\mu + \mu^2}}, \qquad \mu \in [0, \infty]. \tag{A.10}$$

Below we use r_b for the cross statistic to avoid confusion with the trust-region radius $r = \sqrt{2\delta}$ used in Proposition 2.

Introduce multipliers $\lambda \geq 0$ and $\nu \geq 0$ for the affine and trust-region constraints. Stationarity of the local Lagrangian gives

$$g + \lambda b + \nu H d = 0, \qquad d = -\frac{1}{\nu} H^{-1}(g + \lambda b). \tag{A.11}$$

For a boundary solution, $d^\top H d = 2\delta$. Substituting the stationary direction and using the definitions of q, r_b, s yields

$$\begin{aligned} 2\delta &= \frac{1}{\nu^2}(g + \lambda b)^\top H^{-1}(g + \lambda b) \\ &= \frac{1}{\nu^2}\big(q + 2\lambda r_b + \lambda^2 s\big). \end{aligned} \tag{A.12}$$

Now write $r_b = \chi\sqrt{qs}$ and $\lambda = \mu\sqrt{q/s}$. Then

$$q + 2\lambda r_b + \lambda^2 s = q\big(1 + 2\chi\mu + \mu^2\big) = qD(\mu)^2, \tag{A.13}$$

where $D(\mu) = (1 + 2\chi\mu + \mu^2)^{1/2}$. These identities recover the normalized family in Eq. (A.10).

To determine its order, first compute the two linear scores before differentiating:

$$\begin{aligned} g^\top d_\mu &= -\frac{\sqrt{2\delta}}{\sqrt{q}\, D(\mu)} g^\top H^{-1}\left(g + \mu\sqrt{\frac{q}{s}}\, b\right) \\ &= -\frac{\sqrt{2\delta}}{\sqrt{q}\, D(\mu)}\left(q + \mu\sqrt{\frac{q}{s}}\, r_b\right) = -\sqrt{2\delta q}\,\frac{1 + \chi\mu}{D(\mu)}, \\ b^\top d_\mu &= -\frac{\sqrt{2\delta}}{\sqrt{q}\, D(\mu)} b^\top H^{-1}\left(g + \mu\sqrt{\frac{q}{s}}\, b\right) \\ &= -\frac{\sqrt{2\delta}}{\sqrt{q}\, D(\mu)}\left(r_b + \mu\sqrt{\frac{q}{s}}\, s\right) = -\sqrt{2\delta s}\,\frac{\chi + \mu}{D(\mu)}. \end{aligned} \tag{A.14}$$

Set $A_\chi = (1 + \chi\mu)/D$ and $B_\chi = (\chi + \mu)/D$. Since $D'(\mu) = (\chi + \mu)/D(\mu)$, the quotient rule gives the equality chains

$$\begin{aligned} A'_\chi(\mu) &= \frac{\chi D - (1 + \chi\mu)D'}{D^2} = \frac{\chi D^2 - (1 + \chi\mu)(\chi + \mu)}{D^3} \\ &= -\frac{\mu(1 - \chi^2)}{D^3} \leq 0, \\ B'_\chi(\mu) &= \frac{D - (\chi + \mu)D'}{D^2} = \frac{D^2 - (\chi + \mu)^2}{D^3} \\ &= \frac{1 - \chi^2}{D^3} > 0. \end{aligned} \tag{A.15}$$

Thus increasing μ weakens first-order reward improvement and strengthens first-order cost reduction, proving the Pareto ordering. Since $B'_\chi(\mu) > 0$, the chain rule also gives

$$\frac{dA_\chi}{dB_\chi} = \frac{A'_\chi}{B'_\chi} = -\mu, \qquad \frac{d^2 A_\chi}{dB_\chi^2} = -\frac{1}{B'_\chi(\mu)} < 0,$$

so μ is the normalized marginal reward price along a strictly concave local frontier.

For completeness, the angular order is the same. In the basis $e_1 = \hat{g}$ and $\hat{b} = \chi e_1 + \sqrt{1 - \chi^2} e_2$, write $x(\mu) = 1 + \chi\mu$ and $y(\mu) = \mu\sqrt{1 - \chi^2}$. Differentiating $\varphi = \operatorname{atan2}(y, x)$ gives

$$\begin{aligned} \varphi'(\mu) &= \frac{xy' - yx'}{x^2 + y^2} = \frac{(1 + \chi\mu)\sqrt{1 - \chi^2} - \chi\mu\sqrt{1 - \chi^2}}{D(\mu)^2} \\ &= \frac{\sqrt{1 - \chi^2}}{D(\mu)^2} > 0, \end{aligned} \tag{A.16}$$

so the scalar and geometric orders coincide.

A.4 Derivation of the analytic anchor

We first determine when the affine constraint intersects the trust-region ball. For every z with $\|z\|_2 \le \sqrt{2\delta}$, Cauchy–Schwarz gives

$$h^\top z \ge -\|h\|_2\|z\|_2 \ge -\sqrt{2\delta}\,\|h\|_2 = -\sqrt{2\delta s}. \tag{A.17}$$

Both inequalities are equalities at $z_\infty = -\sqrt{2\delta}\,h/\|h\|_2$. Hence

$$\min_{\|z\|_2\le\sqrt{2\delta}} (c + h^\top z) = c - \sqrt{2\delta s} \le 0 \quad\Longleftrightarrow\quad \eta = \frac{c}{\sqrt{2\delta s}} \le 1. \tag{A.18}$$

If $\eta > 1$, the local problem is infeasible, although d_∞ remains the direction of maximum first-order cost reduction.

Next consider the unconstrained reward solution $d_0 = -\sqrt{2\delta/q}\,H^{-1}g$. Its residual is

$$\begin{aligned} c + b^\top d_0 &= c - \sqrt{\frac{2\delta}{q}}\, b^\top H^{-1} g = c - \sqrt{\frac{2\delta}{q}}\, r_b \\ &= \sqrt{2\delta s}\left(\frac{c}{\sqrt{2\delta s}} - \frac{r_b}{\sqrt{qs}}\right) = \sqrt{2\delta s}(\eta - \chi). \end{aligned} \tag{A.19}$$

Therefore d_0 is feasible exactly when $\eta \le \chi$. In this regime it already minimizes the reward objective over the trust-region ball, so adding a satisfied cost constraint cannot change the optimum and $\mu^\star = 0$.

Suppose now that $\chi < \eta < 1$. The unconstrained reward solution is infeasible, while the preceding Cauchy–Schwarz construction provides a strictly feasible point; Slater's condition therefore holds (Boyd & Vandenberghe, 2004). Neither constraint can be inactive: an inactive cost constraint would return d_0, and an inactive trust region would require $g + \lambda b = 0$, excluded by $|\chi| < 1$. Thus the optimum lies on the boundary family and satisfies

$$B_\chi(\mu^\star) = \frac{\chi + \mu^\star}{\sqrt{1 + 2\chi\mu^\star + (\mu^\star)^2}} = \eta. \tag{A.20}$$

The strict monotonicity of B_χ guarantees uniqueness. To solve the active-boundary equation, observe that

$$\begin{aligned} 1 + 2\chi\mu + \mu^2 &= (\chi + \mu)^2 + (1 - \chi^2), \\ \frac{\chi + \mu}{\sqrt{(\chi + \mu)^2 + (1 - \chi^2)}} &= \eta \\ \Longrightarrow\quad (1 - \eta^2)(\chi + \mu)^2 &= \eta^2(1 - \chi^2). \end{aligned} \tag{A.21}$$

The unsquared equation fixes the sign of $\chi + \mu$ to that of η, so the admissible root is

$$\begin{aligned} \chi + \mu^\star &= \eta\sqrt{\frac{1 - \chi^2}{1 - \eta^2}}, \\ \mu^\star &= -\chi + \eta\sqrt{\frac{1 - \chi^2}{1 - \eta^2}}. \end{aligned} \tag{A.22}$$

Combining the regimes yields

$$\mu^\star(\chi, \eta) = \begin{cases} 0, & \eta \le \chi, \\ -\chi + \eta\sqrt{\dfrac{1 - \chi^2}{1 - \eta^2}}, & \chi < \eta < 1, \\ \infty, & \eta \ge 1 \text{ (recovery anchor).} \end{cases} \tag{A.23}$$

At $\eta = 1$, the Cauchy–Schwarz feasibility bound is tight; for $\eta > 1$ the linearized constraint is infeasible, but the same limiting direction remains the cost-recovery limit of maximum first-order cost reduction. Finally, $q = 0$, $s = 0$, and $|\chi| = 1$ reduce the construction to its one-dimensional limits: respectively a flat reward model, a constraint with no first-order directional variation, and collinear whitened gradients.

A.5 Scale invariance of the normalized coordinates

Consider a change of reward and cost units, $g' = \alpha g$ and $(b', c') = (\beta b, \beta c)$ with $\alpha, \beta > 0$. The quadratic statistics transform as

$$q' = (g')^\top H^{-1} g' = \alpha^2 q, \qquad r'_b = (b')^\top H^{-1} g' = \alpha\beta r_b, \qquad s' = (b')^\top H^{-1} b' = \beta^2 s. \tag{A.24}$$

It follows directly that

$$\begin{aligned} \chi' &= \frac{r'_b}{\sqrt{q's'}} = \frac{\alpha\beta r_b}{\sqrt{\alpha^2 q\, \beta^2 s}} = \frac{r_b}{\sqrt{qs}} = \chi, \\ \eta' &= \frac{c'}{\sqrt{2\delta s'}} = \frac{\beta c}{\sqrt{2\delta\, \beta^2 s}} = \frac{c}{\sqrt{2\delta s}} = \eta. \end{aligned} \tag{A.25}$$

To preserve the stationary orientation, the rescaled multiplier must satisfy

$$g' + \lambda' b' = \alpha g + \lambda' \beta b = \alpha(g + \lambda b), \qquad \lambda' = \frac{\alpha}{\beta}\lambda. \tag{A.26}$$

The normalized coordinate is therefore unchanged:

$$\mu' = \lambda' \sqrt{\frac{s'}{q'}} = \frac{\alpha}{\beta}\lambda \sqrt{\frac{\beta^2 s}{\alpha^2 q}} = \lambda \sqrt{\frac{s}{q}} = \mu. \tag{A.27}$$

Thus (χ, η, μ) depend on the local geometry but not on the reporting units of reward or cost.

A.6 Finite-step ambiguity bounds and pressure sensitivity

Write the finite-batch reward-loss increment and cost residual as in Eq. (21). Under Eq. (22), a candidate with radial fraction ρ obeys

$$\frac{1}{2} d^\top H d = \delta\rho^2 \quad \Longrightarrow \quad |\varepsilon_R(d)| \le \kappa_R \delta \rho^2, \qquad |\varepsilon_C(d)| \le \kappa_C \delta \rho^2. \tag{A.28}$$

For feasibility, let $m_C(d) = c + b^\top d$. If $m_C(d) < -\kappa_C \delta\rho^2$, then

$$\widehat{C}(d) = m_C(d) + \varepsilon_C(d) \le m_C(d) + \kappa_C \delta\rho^2 < 0, \tag{A.29}$$

whereas $m_C(d) > \kappa_C \delta\rho^2$ implies $\widehat{C}(d) > 0$ by the symmetric lower bound. Hence finite and linear feasibility can disagree only inside the stated $\kappa_C \delta\rho^2$ margin.

For reward ranking, define the finite increment $\Delta_R(d) = \widehat{L}_R(d) - \widehat{L}_R(0)$. For any two candidates,

$$\begin{aligned} &\Delta_R(d_i) - \Delta_R(d_j) = g^\top (d_i - d_j) + \varepsilon_R(d_i) - \varepsilon_R(d_j), \\ &|\varepsilon_R(d_i) - \varepsilon_R(d_j)| \le \kappa_R \delta (\rho_i^2 + \rho_j^2). \end{aligned} \tag{A.30}$$

Therefore the sign of the finite reward difference must match the linear ordering whenever the latter gap exceeds the right-hand side.

Finally consider the active full-radius KKT path. Define the normalized correction shape

$$e_C(\mu) = \frac{\varepsilon_C(d_\mu)}{\sqrt{2\delta s}},$$

and introduce a scalar perturbation magnitude t. The perturbed boundary is the root of

$$F(\mu, t) = \eta - B_\chi(\mu) + t\, e_C(\mu) = 0, \qquad \widetilde{\mu}(0) = \mu^\star. \tag{A.31}$$

Since $B'_\chi(\mu^\star) > 0$ in the generic active regime, the implicit-function theorem gives a differentiable local root. Differentiating $F(\widetilde{\mu}(t), t) = 0$ and evaluating at $t = 0$ yields

$$\begin{aligned} 0 &= -B'_\chi(\mu^\star)\, \left.\frac{d\widetilde{\mu}}{dt}\right|_{t=0} + e_C(\mu^\star), \\ \left.\frac{d\widetilde{\mu}}{dt}\right|_{t=0} &= \frac{e_C(\mu^\star)}{B'_\chi(\mu^\star)}. \end{aligned} \tag{A.32}$$

At the active solution, $B_\chi(\mu^\star) = \eta$ and $D(\mu)^2 = (\chi + \mu)^2 + (1 - \chi^2)$ imply

$$D(\mu^\star)^2 = \frac{1 - \chi^2}{1 - \eta^2}. \tag{A.33}$$

Substituting this identity into $B'_\chi(\mu) = (1 - \chi^2)/D(\mu)^3$ yields

$$\frac{1}{B'_\chi(\mu^\star)} = \frac{\sqrt{1 - \chi^2}}{(1 - \eta^2)^{3/2}}, \tag{A.34}$$

This is the dual-coordinate sensitivity. Combining the angular derivative above with $D(\mu^\star)^2 = (1 - \chi^2)/(1 - \eta^2)$ gives the corresponding physical angle sensitivity,

$$\left.\frac{d\widetilde{\varphi}}{dt}\right|_{t=0} = \varphi'(\mu^\star) \left.\frac{d\widetilde{\mu}}{dt}\right|_{t=0} = \frac{e_C(\mu^\star)}{\sqrt{1 - \eta^2}}, \tag{A.35}$$

which diverges at rate $O((1 - \eta^2)^{-1/2})$. For a non-infinitesimal correction with $t = 1$ and a uniform bound $|e_C(\mu)| \le \epsilon$ over a monotone segment, the root identity $B_\chi(\widetilde{\mu}) = \eta + e_C(\widetilde{\mu})$ gives the exact bracket

$$B_\chi^{-1}(\eta - \epsilon) \le \widetilde{\mu} \le B_\chi^{-1}(\eta + \epsilon), \tag{A.36}$$

whenever both arguments lie in the range of that segment. This bracket makes explicit that the re-evaluation step resolves a perturbation of the same ordered KKT coordinate rather than an unrelated search direction.

B Implementation Details

B.1 Training procedure

Algorithm 1 summarizes the end-to-end RCS training loop. The pseudocode keeps only operations that define the method: on-policy estimation, Fisher-subspace construction, low-dimensional constrained optimization, empirical-KL correction, and nonlinear same-batch commitment. Numerical stopping tolerances and other solver engineering constants not listed below are fixed in the implementation used for the reported runs.

Algorithm 1 Reward–Constraint Subspace (RCS) actor update

Require: Actor π_{θ_0}, critics V_{ϕ_R}, V_{ϕ_C}, cost budget $\bar{d}$, KL radius δ
Ensure: Updated actor π_{θ_K}

1: **for** $k = 0, 1, \ldots, K - 1$ **do**
2: Collect $\mathcal{D}_k \sim \pi_{\theta_k}$ and estimate $\widehat{A}_R, \widehat{A}_C$; fit both critics
3: Compute local statistics (g, b, c_k, H)
4: Form $B = \mathrm{orth}_H(H^{-1}g, H^{-1}b)$ and the projected models (m_R, m_C)
5: Solve for the constrained point z_{con} and restoration point z_{rec} in the unit ball
6: Map $z_{\mathrm{con}}, z_{\mathrm{rec}}$ to parameter-space candidates and enforce empirical-KL feasibility
7: Re-evaluate nonlinear same-batch reward and cost scores on $\mathcal{D}_k$
8: Choose $d_k^\star$ by Eq. (B.4) and set $\theta_{k+1} = \theta_k + d_k^\star$
9: **end for**

Low-dimensional QCQP solver. The reduced QCQP may be indefinite and therefore nonconvex. Leveraging $k \le 2$, the implementation uses SciPy SLSQP for both the constrained and restoration problems with a small fixed first-order start set: the origin and normalized reward-descent, cost-descent, and exploratory opposite-cost directions ($-a_R$, $-a_C$, and a_C when nonzero), each scaled to radius 0.95 and deduplicated. These are numerical seeds, not a discretization of the first-order KKT frontier. The analytic KKT anchor in Eq. (A.23) exactly identifies the solution of the linearized single-constraint tangent problem, but once Q_R or Q_C is retained it is generally neither stationary nor otherwise privileged for the curvature-corrected QCQP. Including it would therefore add one more local-search initialization without providing a coverage or optimality guarantee. We keep the solver specification fixed and make no global-solver claim. The $+a_C$ seed is an exploratory boundary

seed, not a KKT stationary direction. For the constrained solve, a returned point is retained only if $\|z\|_2^2 \le 1 + 10^{-7}$ and $m_C(z) \le \tau_C + 10^{-7}$; among feasible returns, the point with lowest modeled reward loss is selected. The restoration solve analogously selects the valid point with lowest modeled cost. SLSQP uses at most 100 iterations and function tolerance 10^{-10}. This is a deterministic local search with explicit candidate validation. Empirical-KL overshoot is then corrected by 30 iterations of one-dimensional bisection along the solved direction.

The cost residual is written for a generic task-dependent budget $\bar{d}$ as

$$c_k = (\widetilde{J}_{C,k} - \bar{d})(1 - \gamma). \tag{B.1}$$

The reported configuration computes $\widetilde{J}_{C,k}$ from fast and slow windows of $W_f = 10$ and $W_s = 50$ episodic costs. For $w \in \{f, s\}$ let μ_w be the window mean and let $u_w = \min\{3, \mathrm{sd}_w / \sqrt{n_w}\}$ when $n_w \ge 5$ (and $u_w = 0$ otherwise). With $M_k = \max(\mu_f, \mu_s)$, the uncertainty gate is

$$\omega_k = \mathrm{clip}\left(\frac{M_k - (\bar{d} - 5)}{3}, 0, 1\right), \qquad \widetilde{J}_{C,k} = \max\{\mu_f + \omega_k u_f, \mu_s + \omega_k u_s\}.$$

Thus the uncertainty reserve turns on between five and two cost units below the hard budget and is fully active near the boundary. The reported runs use zero additional safety buffer. The multi-constraint implementation applies the same construction componentwise.

Let $\mathcal{C}_k$ denote the constrained and restoration candidates after empirical-KL correction, and define

$$\mathcal{F}_k = \{d' \in \mathcal{C}_k : \widehat{C}(d') \le \tau_C,\ \widehat{L}_R(d') \le \widehat{L}_R(0)\}, \tag{B.2}$$

$$\mathcal{R}_k = \{d' \in \mathcal{C}_k : \widehat{C}(d') < c_k\}. \tag{B.3}$$

The nonlinear same-batch commitment used in the implementation is the lexicographic rule

$$d_k^\star = \begin{cases} \arg\min\limits_{d' \in \mathcal{F}_k} \big(\widehat{L}_R(d'), \widehat{C}(d'), \widehat{D}_{\mathrm{KL}}(d')\big), & \mathcal{F}_k \neq \varnothing, \\ \arg\min\limits_{d' \in \mathcal{R}_k} \big(\widehat{C}(d'), \widehat{L}_R(d'), \widehat{D}_{\mathrm{KL}}(d')\big), & \mathcal{F}_k = \varnothing,\ c_k > \tau_C,\ \mathcal{R}_k \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{B.4}$$

Thus a step is accepted as finite-step feasible only when it does not deteriorate the reward surrogate. If the current policy is unsafe and no such step exists, RCS accepts only a strict reduction of the raw cost residual; otherwise the actor is left unchanged. This selector is applied after empirical-KL correction and is separate from the local geometric model.

B.2 Key experimental settings

Table 1 reports the RCS-specific settings used in the experiments. Shared seeds, environment-specific training and cost budgets, and aggregation rules are defined once in Table 2. Reward and cost advantages use generalized advantage estimation (GAE) (Schulman et al., 2016). Low-level logging options and parallelization details are intentionally omitted.

The curvature modes differ only in the retained projected Hessian blocks; the Fisher subspace, same-batch selector, KL correction, and optimizer are unchanged. The Goal1 RCS baseline entry reuses the Full runs from the matched curvature intervention. The two-constraint Goal2 extension uses the direct $m = 2$ construction in Appendix C.3.

C Additional Experimental Details

C.1 Safety-Gymnasium environments

Safety-Gymnasium (Ji et al., 2023) is a benchmark suite for safe reinforcement learning in which an agent maximizes task reward while respecting explicit safety costs. We use the standard, unmodified Safety-Gymnasium environments shown in Figure 6. The selected tasks cover qualitatively different reward–constraint interactions: goal-directed navigation, object manipulation, and velocity-constrained locomotion. The multi-constraint experiment additionally uses a navigation task with two distinct safety signals.

Table 1: Key RCS settings used by the reported RCS experiments.

Parameter	Setting
Policy network	Gaussian MLP $[64, 64]$, Tanh; learned state-independent log standard deviation
Reward/cost critics	MLP $[64, 64]$, Tanh
Discount / GAE	$\gamma = 0.99$, $\lambda_R = \lambda_C = 0.95$
Critic optimization	Adam, learning rate 10^{-3}; batch size 128; 10 update epochs
Target KL	$\delta = 0.01$
Fisher damping	0.1
Conjugate-gradient solve	15 iterations
Low-dimensional QCQP solver	deterministic multi-start SLSQP; max 100 iterations; $\texttt{ftol} = 10^{-10}$
Empirical-KL projection	30 bisection iterations along the solved direction
Cost feasibility tolerance	$\tau_C = 10^{-3}$
Cost-control statistic	fast/slow windows $10/50$; SEM coefficient 1; max bonus 3; gate margins $5/2$; zero safety buffer
Curvature modes	Full, Linear, Reward only, and Cost only

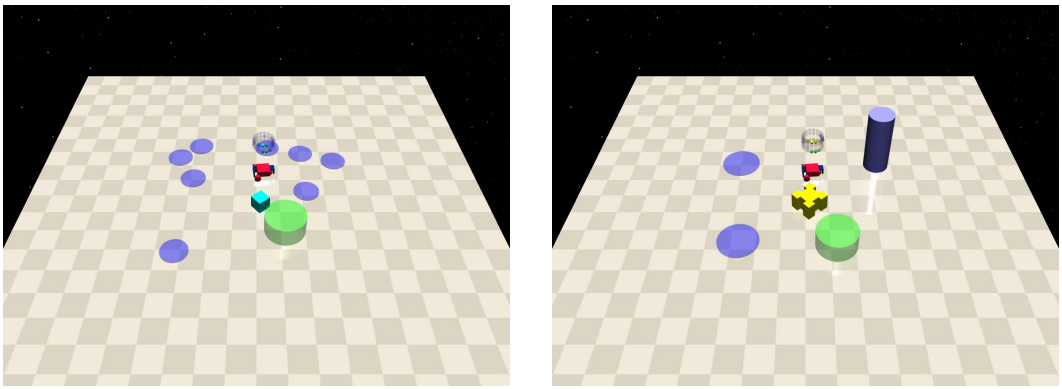
SafetyCarGoal1-v0. The Car agent must navigate toward a goal while avoiding hazardous regions. Entering hazards generates the safety cost, whereas progress toward and arrival at the goal determine task reward. We use this task for the baseline comparison and for the matched curvature ablation. The episodic cost budget in our experiments is $\bar{d} = 25$.

SafetyCarPush1-v0. The Car agent must manipulate a movable box by pushing it toward a target goal while avoiding hazards. Reward reflects progress in bringing the agent toward the box and the box toward the goal, whereas hazard interactions generate safety cost. This task therefore couples safe navigation with object manipulation. The episodic cost budget is $\bar{d} = 25$.

SafetyHumanoidVelocity-v1. This environment uses a MuJoCo Humanoid locomotion task with an explicit velocity constraint. The agent is rewarded for locomotion, while a unit safety cost is generated at a time step when its velocity exceeds the environment-defined threshold. Consequently, unlike the navigation tasks above, the safety signal is induced by the agent's motion rather than by contact with spatial hazards. The episodic cost budget in our experiments is $\bar{d} = 10$.

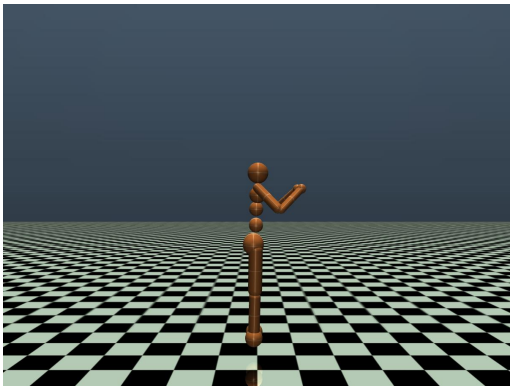
SafetyCarGoal2-v0. We use `SafetyCarGoal2-v0` only for the multi-constraint instantiation in Section 4.3. Compared with Goal1, Goal2 contains both hazards and safety-relevant vases. We retain the hazard-related and vase-related cost signals as separate constraints, yielding the three-dimensional score-relevant domain $\mathrm{span}\{H^{-1}g, H^{-1}b_{\mathrm{hazards}}, H^{-1}b_{\mathrm{vases}}\}$. Both constraints use an episodic budget of 17.5 in the reported experiment.

Environment fidelity. We do not modify the task definitions, reward functions, safety semantics, or environment layouts of Safety-Gymnasium tasks. Our changes are confined to the policy- optimization procedure and the treatment of the two cost channels in the Goal2 multi-constraint experiment. Training horizons and implementation settings are reported in Table 1 and Table 2.

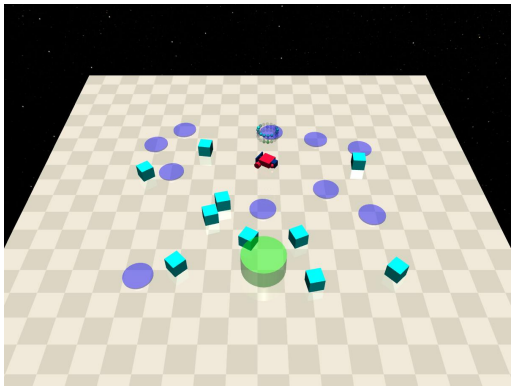


(a) `SafetyCarGoal1-v0`

(b) `SafetyCarPush1-v0`



(c) `SafetyHumanoidVelocity-v1`



(d) `SafetyCarGoal2-v0`

Figure 6: Safety-Gymnasium environments used in our experiments. The first three environments constitute the single-constraint cross-method comparison, while `SafetyCarGoal2-v0` is used for the two-constraint extension. Official environment visualizations are from Safety-Gymnasium (Ji et al., 2023).

C.2 Shared protocol, baseline results, and metric definitions

Table 2 defines the shared seed set, task budgets, final-window aggregation, and reporting conventions for all experiments in this paper, including Goal2 and the matched diagnostics. Unless stated otherwise, outcome tables report the mean $\pm$ s.d. across these runs. The cross-method comparison uses the logger fields available for every completed run: epoch-level `Metrics/EpRet` and `Metrics/EpCost`. All nine methods use the same supplied implementation framework within each task; method-specific optimizer settings follow the archived implementations.

Visualization smoothing. Learning curves are smoothed *only for display*. The plotted return, cost, and violation trajectories use a centered rolling mean with the same window for every method within a task. With the plotting settings used for this paper, the three 300-epoch single-constraint tasks and their curvature-ablation figures use a 9-epoch window; the 500-epoch two-constraint Goal2 figures use a 39-epoch window. The underlying CSV logs, final-window means and standard deviations, budget-exceedance rates, cumulative violation statistics, and all table entries are computed from the unsmoothed data. Smoothing therefore affects visual readability only, not any reported numerical comparison.

Using the shared final-window definition, for each task with budget $\bar{d}_{\text{task}}$, the *budget-exceeding-epoch rate* is the fraction of those final-window epochs whose logged mean episodic cost exceeds $\bar{d}_{\text{task}}$. The learning-curve figures additionally show cumulative mean-cost excess, defined as $\sum_t \max(\texttt{EpCost}_t - \bar{d}_{\text{task}}, 0)$ over the full run. Both are epoch-level logger diagnostics. Exact episode-level violation rate and cumulative positive violation are reported for the matched RCS

Table 2: Shared evaluation and reporting protocol across the experiments. RCS-specific optimization settings appear in Table 1.

Setting	Value
Methods	RCS, CPO, C-TRPO, CUP, FOCOPS, MICE, P3O, PCPO, PPO-Lagrangian
Random seeds	$\{0, 2, 4, 6, 12\}$
Training budget	6M steps for single-constraint tasks; 10M for Goal2; 20,000 steps per epoch
Cost budget	25 (Goal1, Push1); 10 (Humanoid); 17.5 per component (Goal2)
Policy architecture	Gaussian MLP with hidden layers $[64, 64]$
Mini-batch size	128
Shared infrastructure	Same environment, model, rollout-buffer, and logging stack
Compute device	Push1 runs use CUDA; Goal1, Humanoid, and Goal2 use CPU. Runtime comparisons are interpreted only within a task/hardware configuration.
Final-window aggregation	Last 10% of epochs (30 for single-constraint tasks; 50 for Goal2)
Baseline learning-curve summary	Seed mean with 95% confidence interval
Curvature-ablation curve summary	Seed mean with ± 1 s.d. shading
Runtime summary	Wall-clock seconds per epoch, averaged over post-warmup epochs and then over seeds

Table 3: SafetyCarGoal1-v0 baseline comparison. Bold marks the best numerical value in each column; † marks mean final-window cost above budget. The last column is wall-clock time per epoch.

Method	Return ↑	Cost ↓	Exceed. epochs ↓	Cum. cost excess ↓	Time/epoch ↓
RCS (ours)	20.35 ± 1.49	24.72 ± 1.17	$42.7 \pm 19.1\%$	680 ± 83	7.25 ± 0.42
CPO	21.71 ± 0.29	$28.58 \pm 1.00^{\dagger}$	$77.3 \pm 6.4\%$	$1{,}552 \pm 83$	6.63 ± 0.27
C-TRPO	24.23 ± 1.25	$28.70 \pm 1.18^{\dagger}$	$78.0 \pm 3.8\%$	$1{,}481 \pm 207$	6.96 ± 0.49
CUP	$\mathbf{28.43 \pm 2.56}$	$40.79 \pm 4.15^{\dagger}$	$99.3 \pm 1.5\%$	$5{,}231 \pm 932$	8.03 ± 0.10
FOCOPS	27.42 ± 1.09	$31.93 \pm 4.55^{\dagger}$	$86.7 \pm 16.5\%$	$2{,}720 \pm 364$	$\mathbf{6.60 \pm 0.23}$
MICE	20.41 ± 1.11	$27.29 \pm 1.20^{\dagger}$	$67.3 \pm 12.3\%$	$\mathbf{616 \pm 148}$	8.52 ± 0.27
P3O	4.72 ± 0.92	24.96 ± 2.13	$46.0 \pm 11.9\%$	$1{,}813 \pm 366$	14.50 ± 0.65
PCPO	18.62 ± 1.11	$31.44 \pm 2.76^{\dagger}$	$87.3 \pm 4.9\%$	$2{,}175 \pm 314$	6.78 ± 0.28
PPO-Lagrangian	19.82 ± 1.21	$\mathbf{23.56 \pm 4.36}$	$\mathbf{36.0 \pm 26.8\%}$	$3{,}680 \pm 599$	8.01 ± 0.09

curvature study, where per-episode counts are available. Keeping these definitions separate avoids mixing incompatible safety statistics.

C.3 Two-constraint Goal2 extension

To instantiate the general m-constraint construction beyond the one-cost setting, we use SafetyCarGoal2-v0 with separate hazard and vase constraints. The RCS extension forms $\mathcal{S}_2 = \mathrm{span}\{H^{-1}g, H^{-1}b_{\mathrm{hazards}}, H^{-1}b_{\mathrm{vases}}\}$, so the reduced problem has dimension at most three. It projects reward curvature and one cost-curvature block per constraint into this shared basis and imposes both quadratic cost models in the reduced QCQP.

When restoration is needed, the implementation normalizes each modeled residual by $\sigma_i = \max\{\sqrt{2\delta\, b_i^{\top} H^{-1} b_i}, 10^{-6}\}$ and solves the smooth epigraph problem

$$\min_{\|z\|_2 \le 1,\, t} \; t \quad \text{s.t.} \quad \frac{m_{C_i}(z) - \tau_i}{\sigma_i} \le t, \qquad i = 1, \ldots, m.$$

Deterministic starts include the origin, reward descent $-a_R$, both signs $\pm a_{C_i}$ for every constraint, and–when constraints are currently violated–a normalized violation-weighted recovery direction

Table 4: SafetyCarPush1-v0 baseline comparison. Bold marks the best numerical value in each column; † marks mean final-window cost above budget.

Method	Return ↑	Cost ↓	Exceed. epochs ↓	Cum. cost excess ↓	Time/epoch ↓
RCS (ours)	1.05±0.97	19.48±2.55	**23.3±13.1%**	865±155	8.01±0.03
CPO	0.58±0.16	27.32±4.09†	56.7±12.7%	1,923±176	9.11±1.14
C-TRPO	0.68±0.16	27.21±6.28†	48.7±14.1%	1,947±317	**7.12±0.14**
CUP	**1.34±0.75**	27.77±4.73†	53.3±20.0%	1,664±462	19.13±5.63
FOCOPS	0.53±0.16	20.58±7.91	31.3±26.3%	1,796±320	7.96±0.15
MICE	0.42±0.08	21.06±7.02	28.7±23.0%	907±215	10.97±1.39
P3O	0.39±0.05	20.71±3.19	32.0±10.7%	**835±305**	24.41±2.45
PCPO	0.51±0.19	23.95±3.83	39.3±12.8%	1,394±219	9.62±1.08
PPO-Lagrangian	0.93±0.36	25.98±4.37†	53.3±18.6%	1,939±282	10.81±0.59

Table 5: SafetyHumanoidVelocity-v1 baseline comparison. Bold marks the best numerical value in each column; † marks mean final-window cost above budget. The last column is wall-clock time per epoch.

Method	Return ↑	Cost ↓	Exceed. epochs ↓	Cum. cost excess ↓	Time/epoch ↓
RCS (ours)	**5,411.90±312.22**	5.72±1.16	6.0±7.2%	57±59	5.71±0.22
CPO	5,189.38±214.28	5.72±2.52	12.7±16.4%	54±66	5.21±0.08
C-TRPO	5,068.76±152.49	5.99±1.17	8.7±6.9%	55±23	5.14±0.02
CUP	4,058.31±1,200.03	7.40±2.47	22.0±24.3%	94±56	12.88±0.83
FOCOPS	4,413.55±287.11	9.85±1.42	39.3±21.3%	184±69	**2.74±0.06**
MICE	4,982.67±174.99	6.08±1.32	10.0±10.0%	21±20	6.25±0.16
P3O	5,143.82±191.09	**1.52±0.48**	**0.0±0.0%**	**0±0**	4.90±0.13
PCPO	1,695.06±587.92	5.69±1.60	4.0±4.3%	4±3	5.22±0.01
PPO-Lagrangian	4,783.11±166.02	10.07±1.68†	45.3±21.7%	221±69	2.82±0.05

$-\sum_i v_i a_{C_i}$, where $v_i = [(c_i - \tau_i)/\sigma_i]_+$. This specifies the higher-dimensional extension of the single-constraint multi-start rule.

The compared multi-constraint baselines are isolated engineering extensions of the supplied scalar-cost implementations, not claims about the exact algorithms in the original publications. CPO-M solves one trust-region problem with both linearized component constraints; PCPO-M projects the reward TRPO step onto their intersection in Fisher geometry; PPO-Lag-M and FOCOPS-M maintain one multiplier per component; CUP-M uses the sum of independently multiplier-weighted cost advantages in its projection stage; P3O-M sums separate positive-part penalties; C-TRPO-M sums feasible-regime barrier terms and uses a normalized mixture of violated component advantages for recovery; and MICE-M treats any positive component cost as unsafe and adds its shared intrinsic safety signal to each component before the multi-constraint CPO update.

Table 6 reports the two-constraint comparison under the shared protocol. RCS obtains mean final-window return 2.44±1.02, mean final-window hazards cost 17.29±4.17, and mean final-window vases cost 16.55±1.86, with both mean final-window component costs below their 17.5 limits. FOCOPS reaches higher mean final-window return with larger cumulative component-wise mean-cost excess, whereas MICE reduces both component costs and cumulative component-wise mean-cost excess further at lower mean final-window return. The results characterize the return–safety trade-off obtained by optimizing both constraints in one shared subspace.

C.4 Curvature-ablation diagnostics

The factorial intervention matches Full, Linear, Reward only, and Cost only on the shared runs. Figure 7 provides the corresponding one-row, four-panel outcome trajectories, while Figure 4 in the main paper reports the directional mechanism diagnostic. Table 7 uses the shared final-window definition and exact episode-level cumulative positive violation over the complete run; all entries

Table 6: SafetyCarGoal2-v0 two-constraint comparison. Red values with † mark a mean final-window component cost above its limit. Unlike the other timing columns, time/epoch uses the measured seed-0 wall-clock time after warmup.

Method	Return ↑	Hazards cost ↓	Vases cost ↓	Cum. hazard excess ↓	Cum. vase excess ↓	Time/epoch ↓
RCS (ours)	2.44±1.02	17.29±4.17	16.55±1.86	2,252±1,076	1,416±146	11.86
CPO	3.31±0.38	19.29±2.48†	23.40±2.94†	3,804±373	3,476±323	11.22
C-TRPO	**4.29±0.93**	19.77±2.62†	19.66±1.37†	3,308±369	3,191±111	**9.61**
CUP	0.29±1.00	68.22±78.62†	33.85±33.44†	19,439±22,693	7,727±7,358	21.86
FOCOPS	3.35±0.36	16.83±0.61	17.19±1.93	3,919±591	4,735±506	11.00
MICE	−0.34±1.00	**7.33±2.99**	**6.51±2.23**	**751±167**	**387±141**	11.58
P3O	0.74±0.13	19.58±1.56†	16.18±2.07	5,248±754	3,401±420	22.52
PCPO	4.09±0.99	19.07±1.87†	22.40±2.16†	3,620±387	3,387±193	10.41
PPO-Lagrangian	1.91±0.39	17.20±4.13	17.14±3.86	5,057±632	6,047±796	18.87

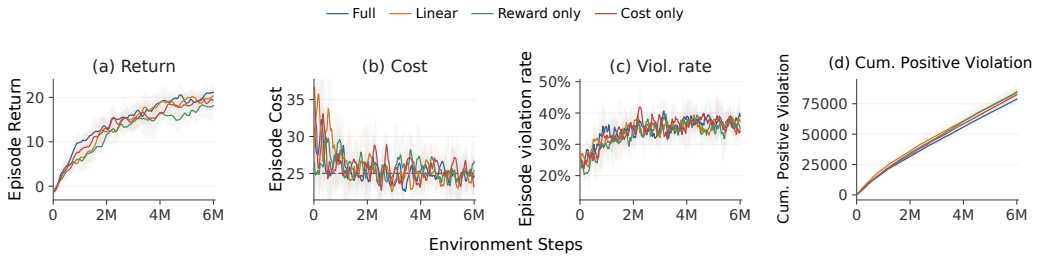


Figure 7: Matched curvature-ablation outcomes on SafetyCarGoal1-v0. The four variants differ only in the retained projected curvature blocks; curves show return, cost, episode violation rate, and cumulative positive violation.

come from unsmoothed logs. Relative to Linear, Full increases mean final-window return from 19.44 to 20.35, leaves mean final-window cost essentially unchanged (24.71 vs. 24.72), and reduces cumulative positive violation from 84,055 to 78,774. Together with the finite-step model-error and orientation diagnostics, these outcome metrics summarize the matched curvature intervention.

Table 7: Matched curvature ablation on SafetyCarGoal1-v0. Variants differ only in retained curvature blocks. "Cum. mean-cost excess" is epoch-level; "Cum. viol." is episode-level. † marks mean final-window cost above budget; bold marks the best mean in each column.

Variant	Return ↑	Cost ↓	Viol. rate ↓	Cum. viol. ↓	Cum. mean-cost excess ↓
Full (Q_R, Q_C)	**20.35±1.49**	24.72±1.17	37.4±1.5%	**78,774±4,512**	**680±83**
Linear $(0, 0)$	19.44±1.34	24.71±1.65	37.1±2.6%	84,055±6,582	864±188
Reward only $(Q_R, 0)$	17.99±2.84	25.12±2.15†	36.0±1.5%	84,866±5,143	823±144
Cost only $(0, Q_C)$	19.27±0.90	**24.64±1.28**	**35.6±3.3%**	82,443±3,858	858±146

Using the cross-method epoch-level safety metric, Full also has lower cumulative mean-cost excess than Linear (680 vs. 864). Thus the episode-level cumulative violation and the baseline-compatible epoch-level excess agree in direction, both indicating lower training-time violation for Full under the shared protocol.

Across all 6,000 logged actor updates, accepted steps retain unit reduced radius to numerical tolerance; the largest recorded empirical KL is 0.00817, below the target 0.01. Hence empirical-KL bisection does not activate in these Goal1 runs. Full and Linear also have very similar QCQP/restoration/rejection frequencies. Roughly 70% of committed updates use the restoration candidate, making restoration a substantial component of the observed training-time safety behavior.

Table 9 tests whether enabling each curvature block improves its corresponding finite-displacement quadratic model. The error is evaluated at the final committed, empirical-KL-corrected displacement

Table 8: Goal1 commitment diagnostics. Rates are run-level percentages over 300 actor updates. "KL-shrunk" counts accepted updates whose final reduced radius is below its solved radius after empirical-KL projection.

Variant	QCQP commit ↑	Restoration commit	Rejected ↓	KL-shrunk accepted ↓
Full (Q_R, Q_C)	$28.3 \pm 1.7\%$	$71.1 \pm 2.3\%$	$0.6 \pm 0.8\%$	$0.0 \pm 0.0\%$
Linear $(0, 0)$	$26.6 \pm 1.6\%$	$72.8 \pm 1.7\%$	$0.6 \pm 0.4\%$	$0.0 \pm 0.0\%$
Reward only $(Q_R, 0)$	$25.0 \pm 1.3\%$	$74.7 \pm 1.5\%$	$0.3 \pm 0.3\%$	$0.0 \pm 0.0\%$
Cost only $(0, Q_C)$	$27.5 \pm 1.7\%$	$72.1 \pm 1.7\%$	$0.4 \pm 0.1\%$	$0.0 \pm 0.0\%$

$d_k^\star$ and its subspace coordinate $z_k^\star$ (both set to zero for a rejected update):

$$E_{R,k} = \left| \widehat{L}_R(d_k^\star) - \widehat{L}_R(0) - m_R(z_k^\star) \right|, \qquad E_{C,k} = \left| [\widehat{C}(d_k^\star) - c_k] - [m_C(z_k^\star) - c_k] \right|.$$

Each run contributes $T^{-1} \sum_{k=1}^{T} E_{R,k}$ and $T^{-1} \sum_{k=1}^{T} E_{C,k}$ over its $T = 300$ actor updates; Table 9 aggregates these per-run errors.

Table 9: Finite-displacement model MAE in the curvature ablation. Bold marks the lowest mean.

Variant	Reward MAE ↓	Cost MAE ↓
Full (Q_R, Q_C)	$\mathbf{5.58 \times 10^{-5}} \pm 5.82 \times 10^{-6}$	$\mathbf{3.73 \times 10^{-4}} \pm 1.45 \times 10^{-5}$
Linear $(0, 0)$	$2.61 \times 10^{-4} \pm 2.19 \times 10^{-5}$	$1.82 \times 10^{-3} \pm 5.28 \times 10^{-5}$
Reward only $(Q_R, 0)$	$6.24 \times 10^{-5} \pm 7.29 \times 10^{-6}$	$1.98 \times 10^{-3} \pm 1.44 \times 10^{-4}$
Cost only $(0, Q_C)$	$2.67 \times 10^{-4} \pm 2.99 \times 10^{-5}$	$3.92 \times 10^{-4} \pm 3.43 \times 10^{-5}$

C.5 PUSH1 FULL-VERSUS-LINEAR CURVATURE ABLATION

We run matched Full and Linear variants on SafetyCarPush1-v0 with the same five seeds, training budget, Fisher subspace, solver, empirical-KL correction, and commitment rule. Linear sets both projected curvature blocks to zero while leaving the rest of the update unchanged. Table 10 reports unsmoothed final-window statistics, and Figure 8 shows the corresponding trajectories.

Table 10: Matched Full-versus-Linear ablation on SafetyCarPush1-v0. Values are mean ± s.d. over five seeds. Final-window statistics use the shared last-10% aggregation; cumulative mean-cost excess is computed from unsmoothed epoch-level logs over the complete run.

Variant	Return ↑	Final-window cost ↓	Cum. mean-cost excess ↓
Full (Q_R, Q_C)	1.05 ± 0.97	19.48 ± 2.55	865 ± 155
Linear $(0, 0)$	1.05 ± 0.44	20.30 ± 3.76	$1{,}015 \pm 360$

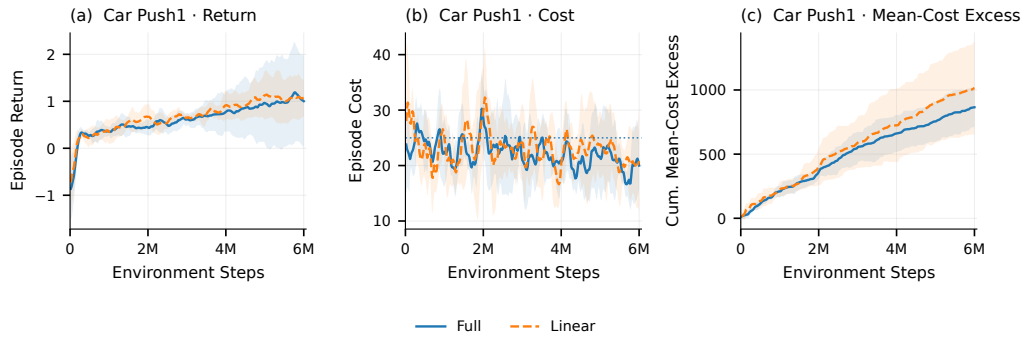


Figure 8: Matched Full-versus-Linear ablation on SafetyCarPush1-v0. Curves show seed mean with ± 1 s.d. shading; the dotted horizontal line marks the cost budget. Full retains projected reward and cost curvature, while Linear sets both projected curvature blocks to zero.

C.6 Independent full-space curvature leakage diagnostic

The diagnostic reuses the final Goal1 Full checkpoints (epoch 299) from the shared runs. For each checkpoint we collect one fresh on-policy batch and reconstruct the local reward gradient g, cost gradient b, damped Fisher operator H, and reward/cost Hessian–vector-product operators using the same definitions as RCS. Let $q = g^\top H^{-1} g$, $s = b^\top H^{-1} b$, and $\chi = b^\top H^{-1} g / \sqrt{qs}$. For each trust-region budget we fix $\eta = 0.6$ and set $c_\delta = \eta\sqrt{2\delta s}$. Holding η fixed matches the asymptotic family of Proposition 2; keeping the literal training residual fixed while changing δ would define a different local family.

Nonlinear same-batch reference. For the main diagnostic, both the restricted solution d_S^N and the full-space local reference d_F^N optimize the original likelihood-ratio reward-loss and cost-residual surrogates inside the same damped-Fisher quadratic trust region. The full-space basis starts from $\mathcal{S}_1$ and is expanded only with the current nonlinear full-parameter KKT stationarity residual $H^{-1} r_{\mathrm{KKT}}$; the predicted leakage vector is never supplied. At each expansion, the nonlinear constrained problem is re-solved in the enlarged Fisher-orthonormal basis. We certify a local reference only when normalized stationarity is at most 3×10^{-3}, the active-cost and trust-region residuals are at numerical tolerance, and the fitted cost and trust-region multipliers are positive. These checks establish a numerically regular local KKT point, not global optimality of the nonconvex surrogate problem.

At $\delta \in \{6.25 \times 10^{-4}, 1.25 \times 10^{-3}, 2.5 \times 10^{-3}\}$, all five seeds pass these checks. Seedwise log–log fits on exactly this uniformly certified window give 1.074 ± 0.043 SEM for $\|d_F^N - d_S^N\|_H$ and 1.550 ± 0.047 SEM for $\widehat{L}_R(d_S^N) - \widehat{L}_R(d_F^N)$. At the smallest budget, the observed displacement is 1.000 ± 0.035 SEM times its leading prediction, the nonlinear reward-loss surrogate gap is 0.989 ± 0.053 times its leading prediction, and the Fisher cosine with $-(2\delta/\nu_0) H^{-1/2} \ell_\perp$ is 0.984 ± 0.004. Certification drops to four seeds at $\delta = 5 \times 10^{-3}$ and one seed at 10^{-2}; those larger-radius points are retained as a stress test but excluded from the asymptotic fit. The complete records are provided in `data/nonlinear_diagnostics.csv` and `data/nonlinear_reference_trace.csv`.

Quadratic specialization and one-direction enrichment. From the active tangent solution and its multipliers we form the active Lagrangian curvature and Fisher-orthogonal leakage $\ell_\perp$ in Eq. (16). We Fisher-orthogonalize $v_{\mathrm{leak}} = H^{-1/2} \ell_\perp$ against $\mathcal{S}_1$ and define the enriched diagnostic domain $\mathcal{S}_1^+ = \mathcal{S}_1 + \mathrm{span}\{v_{\mathrm{leak}}\}$. We solve the same constrained quadratic local model in $\mathcal{S}_1$ and $\mathcal{S}_1^+$, obtaining d_S and d_{S^+} for $\delta \in \{6.25 \times 10^{-4}, 1.25 \times 10^{-3}, 2.5 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$. This targeted enrichment asks whether the single leading direction predicted by Proposition 2 is sufficient to recover the omitted finite-step correction.

As a complementary specialization with tighter numerical certification across the full radius sweep, we compute a matrix-free full-space quadratic reference d_F. The solver begins from the solution in the score-preserving subspace and expands its search space using full-space KKT stationarity residuals; again, the predicted leakage direction is not supplied as a basis vector or initialization. Across the five

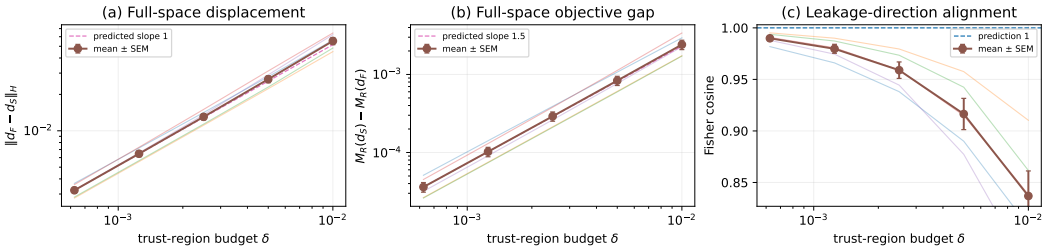


Figure 9: Complementary quadratic full-space leakage diagnostic. The matrix-free reference does not receive the predicted leakage direction. All 25 seed–budget quadratic references pass the prescribed KKT-residual checks, and the full sweep closely follows the predicted displacement and objective-gap scalings.

seeds and five budgets, all 25 quadratic references pass the prescribed KKT-residual checks; the final stationarity residual ranges from 1.24×10^{-4} to 9.90×10^{-4}, while the active-cost and trust-region residuals are at numerical precision. These checks are numerical first-order stationarity checks, not a global-optimality certificate for an indefinite quadratic model.

For the quadratic reference, seedwise log–log fits give a displacement exponent of 1.024 ± 0.015 SEM for $\|d_F - d_S\|_H$ and a reward-model-gap exponent of 1.516 ± 0.016 SEM for $M_R(d_S) - M_R(d_F)$. At $\delta = 6.25 \times 10^{-4}$, the observed displacement magnitude is 0.995 times the leading prediction on average, the objective gap is 1.005 times the leading prediction, and the Fisher cosine with the predicted direction is 0.990 ± 0.002 SEM. At $\delta = 10^{-2}$, these ratios are 1.070 and 1.058, while the cosine decreases to 0.837 ± 0.024.

How much of the quadratic full-space improvement is one-dimensional? Define the captured fraction

$$\rho_{\text{capture}} = \frac{M_R(d_S) - M_R(d_{S^+})}{M_R(d_S) - M_R(d_F)}.$$

At the smallest trust-region budget, the predicted one-dimensional enrichment captures $98.5\% \pm 0.3\%$ SEM of the full-space reward-model improvement. The captured fraction is $97.1\% \pm 0.6\%$, $94.5\% \pm 1.1\%$, $89.9\% \pm 1.8\%$, and $82.7\% \pm 2.5\%$ as δ increases through the remaining four budgets. Thus the theoretically identified direction explains nearly all of the first omitted quadratic correction in the asymptotic regime, while higher-order directions account for a growing share at larger radius.

The complete quadratic per-seed diagnostics and residual-expansion trace are included under `data/` as `fullspace_diagnostics.csv` and `reference_trace.csv`. The quadratic solve is a numerical local-model reference rather than a claim of global optimization for the original nonlinear policy surrogate.

D Related Work

Constrained reinforcement learning. CMDPs formulate safe control as reward maximization under expected cumulative constraints (Altman, 1999). A broad family of methods learns the reward–cost trade-off across policy updates through actor–critic, primal–dual, penalty, or optimistic formulations (Borkar, 2005; Paternain et al., 2019; Gattami et al., 2021; Ding et al., 2021; Tessler et al., 2019; Stooke et al., 2020; Zhang et al., 2022; Xu et al., 2021; Bai et al., 2023; Ding et al., 2025; Moskovitz et al., 2023). Recent work also improves the learned safety signal itself: MICE addresses cost underestimation during safe exploration, while FDPI targets the scarcity of constraint-violating samples used to learn feasibility (Gao et al., 2025; Yang et al., 2026). A complementary line places the constraint directly inside each policy update. CPO solves a constrained trust-region surrogate, while later projection, first-order policy-space, and trust-region-geometry methods vary how the safe update is represented, estimated, or fitted (Achiam et al., 2017; Yang et al., 2020; Zhang et al., 2020; Yang et al., 2022; Dai et al., 2024; Milosevic et al., 2025; Mattick et al., 2026). RCS follows this update-level perspective.

The comparator suite reflects this diversity, covering constrained trust-region methods, projection and first-order policy-space corrections, penalty or dual-relaxation methods, and safety-signal correction approaches. These baselines span the main constraint-handling mechanisms most directly related to RCS.

Information-geometric constrained updates. Natural-gradient and relative-entropy methods provide the Fisher geometry used to control policy displacement (Amari, 1998; Kakade, 2001; Peters et al., 2010), while conservative policy iteration, safe policy iteration, and TRPO connect local update control to approximate improvement (Kakade & Langford, 2002; Pirotta et al., 2013; Schulman et al., 2015). RCS uses the same local metric to identify, before adding finite-step curvature, the smallest linear domain that preserves all first-order reward–constraint information relevant to the constrained update.

Tangent solution versus finite-step departure. In the single-constraint CPO subproblem, KKT stationarity places a regular tangent optimizer in $\mathrm{span}\{H^{-1}g, H^{-1}b\}$ (Achiam et al., 2017). Proposition 1 extends this geometric view by identifying the complete first-order score-preserving domain. Proposition 2 then characterizes the transition beyond tangent order for smooth finite surrogates: the first omitted correction is the Fisher-orthogonal action of the active Lagrangian curvature, giving an $O(\delta)$ displacement correction and an $O(\delta^{3/2})$ reward-loss surrogate gap.

Reduced-order and multi-objective perspectives. Classical trust-region optimization also uses low-dimensional search spaces, including two-dimensional subspace methods that minimize a quadratic model over a span built from gradient or Newton information (Byrd et al., 1988). RCS differs in how the domain is chosen: $\mathcal{S}_m$ follows from exact reward–constraint score preservation under the Fisher metric before curvature is evaluated, rather than from a spectral or Newton approximation of the Hessian. Multi-objective policy optimization addresses a different setting, where weighted scalarizations or Pareto sets encode persistent preferences (Abdolmaleki et al., 2020); here the reward and constraint semantics remain fixed. The reduction removes dependence on policy dimension, although a large number of constraints can still make the reduced nonconvex QCQP difficult. Active-set identification or constraint pruning are natural directions in that regime.