BEYOND THE TRUST RADIUS: OBJECTIVE-AWARE STEPS FOR POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A fixed trust radius controls policy movement without using the gap between the current loss and a certified lower bound. We connect these two scales through Chentsov–Polyak Safe Policy Optimization, which takes the smaller of a quadratic Fisher trust-region step and a smooth Polyak step. Given a certified objective lower bound and smoothness control, this is the largest step along the fixed natural-gradient direction consistent with both the lower-Taylor and movement bounds. Objective information can therefore tighten the permitted step within a fixed trust region. For constrained learning, a separate closed-form cap preserves cumulative-cost chance feasibility from a surrogate-feasible start under sub-Gaussian costs and certified cost and curvature bounds. Synthetic experiments show reduced sensitivity to oversized radii and report no chance-infeasible iterates among 20,000 tested with the safety cap. On three Safety-Gymnasium tasks, we use a running-best reward proxy for the objective bound, a Lagrangian fallback, and one shared radius. At one million environment steps, our method retains 93–97% of radius-selected trust-region policy-optimization-Lagrangian return on two navigation tasks, but 41% on pushing, with lower pushing cost; mean costs exceed the budget on all three tasks. This exposes a distinction for policy learning: controlling policy movement, scaling optimization steps, and certifying cost constraints require different information.

1 INTRODUCTION

Trust-region policy optimization controls the size of a policy update, but the trust radius alone does not express how much objective improvement remains. A fixed quadratic Fisher radius gives a natural-gradient step with scalar step size $\sqrt{2\delta/d}$, where $d = \nabla L^\top F^{-1} \nabla L$. As the gradient becomes small, this scalar grows; the allowed Fisher displacement remains $\sqrt{2\delta}$. An objective gap supplies a different scale. This paper studies the intersection of these two step limits along a fixed natural-gradient direction.

Chentsov–Polyak Safe Policy Optimization (CP-SPO) takes the smaller of a quadratic trust-region step and a smooth Polyak step. With a certified objective lower bound and a valid smoothness bound, the result is the largest step along a fixed natural-gradient direction whose lower Taylor model remains above the bound and whose quadratic Fisher displacement fits the radius. This is a characterization of two chosen local certificates. It does not establish optimality among policy updates, remove all tuning, or certify cost feasibility.

The distinction matters in constrained reinforcement learning (RL). TRPO (Schulman et al., 2015) controls policy movement; approaches such as CPO (Achiam et al., 2017), PCPO (Yang et al., 2020), and Lagrangian methods (Tessler et al., 2019; Stooke et al., 2020) add different mechanisms for controlling cost. We give a separate closed-form step cap that preserves a cumulative-cost chance constraint. Its guarantee requires a surrogate-feasible start, sub-Gaussian costs, and certified cost and curvature bounds. These requirements are additional to the objective-side certificate.

The contribution has three parts. First, we characterize the smooth min-form step and quantify its sensitivity to objective and curvature information. Second, we state conditional convergence and cost feasibility results with their distinct assumptions. Third, we isolate the practical cap in a matched RL ablation and test sensitivity to certified curvature and the empirical return target. The Fisher direction uses classical information geometry rather than a new metric-uniqueness result. On PointGoal1,

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.

the cap-only comparison improves final return by 8.6% and reduces cost-excess area by 5.7%. The three Safety-Gymnasium tasks also expose limits of the shared-radius approach. At one million environment steps, a shared-radius implementation retains 97% and 93% of the radius-selected TRPO-Lagrangian return on two navigation tasks, but 41% on pushing. Mean costs exceed the budget on all three. The real-RL results therefore evaluate a practical step heuristic, not the chance-feasibility theorem.

2 Setup and related work

Objective and constraint. For a CMDP with transition kernel P, reward r, cost c, discount $\gamma \in (0,1)$, and initial distribution ρ_0, let μ_{π_θ} be the trajectory law of a smooth policy $\pi_\theta(a \mid s)$. We minimize $L(\theta) = -J_r(\pi_\theta)$, where $J_r = \mathbb{E}_{\mu_{\pi_\theta}}[\sum_t \gamma^t r(s_t, a_t)]$. The certified constrained problem is

$$\min_{\theta\in\Theta} L(\theta) \quad \text{subject to} \quad \mathbb{P}_{\mu_{\pi_\theta}}\{Z(\theta) > C\} \le \varepsilon, \qquad Z(\theta) = \sum_{t=0}^{H} \gamma^t c(s_t, a_t). \tag{1}$$

This event concerns cumulative cost, not the event that a per-step constraint is ever violated. Mean-cost constraints and cumulative-cost chance constraints are also different; the Safety-Gymnasium experiments use an episodic mean-cost budget. We write L_{lb} for a certified lower bound over the region on which an objective theorem is applied, and $L^\star = \inf_\theta L(\theta)$ for the optimum. Their gap is not assumed zero except where explicitly required for convergence.

Fisher metric. Let $s_\theta(a \mid s) = \nabla_\theta \log \pi_\theta(a \mid s)$ and d^{π_θ} denote the discounted state-visitation distribution.

$$F_s(\theta) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[s_\theta s_\theta^\top], \quad F(\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}}[F_s(\theta)], \quad d(\theta) = \nabla L^\top F^{-1} \nabla L. \tag{2}$$

Following TRPO, the state distribution is held fixed when differentiating the reference average KL. Under the model regularity in Appendix A.1, its Hessian at the current policy is $F(\theta)$. Thus $\frac{1}{2}\Delta\theta^\top F(\theta)\Delta\theta \le \delta$ is a *quadratic approximation* to an average-KL constraint, not an exact finite-step KL guarantee. Exact KL acceptance requires a separate check. With a fixed metric, solving the linearized objective along its natural-gradient direction gives

$$\theta^+ = \theta - \eta^{\mathrm{TR}} F^{-1}\nabla L, \qquad \eta^{\mathrm{TR}} = \sqrt{2\delta/d}. \tag{3}$$

Assumption 1 (Local regularity)**.** *The objective is C^2 and β-smooth in the chosen Euclidean coordinates on a neighbourhood containing every step segment under consideration. The objective bound L_{lb} is valid on this region, and $F(\theta) \succeq \mu_F I$ for $\mu_F > 0$. Cost regularity is imposed separately in Theorem 2.*

Trust-region methods and geometry. Natural policy gradients (Kakade, 2002) and TRPO (Schulman et al., 2015) already motivate Fisher preconditioning. CPO and PCPO incorporate local constraint approximations, while C-TRPO (Milosevic et al., 2025) changes the geometry to account for safety. Chentsov's metric characterization (Chentsov, 1972; Ay et al., 2015) is also an established ingredient. Our scalar cap operates on a fixed direction and can coexist with such choices, but is not a replacement for their constraint mechanisms.

Objective-based step sizes. The classical Polyak rule (Polyak, 1987) and its stochastic extension (Loizou et al., 2021) use an objective gap to adapt a step. Policy-gradient variants already use estimated target values (Li et al., 2024). Our focus is the intersection of a smooth lower-Taylor certificate and a quadratic movement certificate. Polyak steps already have a theoretical connection to tuned gradient clipping under convex generalized-smoothness assumptions (Takezawa et al., 2024). Adaptation from a valid lower bound (Hazan & Kakade, 2019) and lower-bound-dependent convergence neighborhoods (Orvieto et al., 2022) are also established. Our characterization does not establish a new general Polyak–clipping connection or certify a running-best target.

Chance-constrained optimization. Exponential-moment chance-constraint approximations (Nemirovski & Shapiro, 2007) and Gaussian chance constraints in power flow (Bienstock et al., 2014) are established precedents; our sub-Gaussian bound follows directly from the stated MGF assumption. Appendix A gives supporting derivations.

3 OBJECTIVE-AWARE STEPS INSIDE A TRUST REGION

Theorem 1 (Largest step satisfying the two local certificates)**.** *Fix a positive-definite preconditioner $M \succ 0$, a nonzero gradient $g = \nabla L(\theta)$, and write $d_M = g^\top M g$, $q_M = \|Mg\|_2^2$, and $\Delta = L(\theta) - L_{\mathrm{lb}} \geq 0$. Let the quadratic trust radius satisfy $\delta \geq 0$. Under Assumption 1, call $\eta \geq 0$ admissible if*

$$L(\theta) - \eta d_M - \tfrac{1}{2}\beta\eta^2 q_M \geq L_{\mathrm{lb}}, \qquad \tfrac{1}{2}\eta^2 d_M \leq \delta. \tag{4}$$

The largest admissible scalar is

$$\eta_M^\star = \min\{\eta_M^{\mathrm{smPol}}, \eta_M^{\mathrm{TR}}\}, \qquad \eta_M^{\mathrm{smPol}} = \frac{2\Delta}{d_M + \sqrt{d_M^2 + 2\beta q_M \Delta}}, \quad \eta_M^{\mathrm{TR}} = \sqrt{\frac{2\delta}{d_M}}. \tag{5}$$

When $\beta q_M = 0$, $\eta_M^{\mathrm{smPol}} = \Delta/d_M$. When $g = 0$, the direction is zero and we define the update to be zero rather than divide by d_M.

Proof. The first constraint is $(\beta q_M/2)\eta^2 + d_M\eta - \Delta \leq 0$, with nonnegative endpoint η_M^{smPol}. The second has endpoint η_M^{TR}. Both intervals contain zero, so their intersection ends at the smaller endpoint. □

For CP-SPO, set $M = F^{-1}$ and $q = \|F^{-1}g\|^2 = \tilde{d}$:

$$\eta_{\mathrm{FR}}^\star = \min\left\{\frac{2\Delta}{d + \sqrt{d^2 + 2\beta q\Delta}}, \sqrt{2\delta/d}\right\}, \qquad \theta^+ = \theta - \eta_{\mathrm{FR}}^\star F^{-1} g. \tag{6}$$

The objective branch tightens movement exactly when $\delta \geq \delta_{\mathrm{switch}} := d(\eta^{\mathrm{smPol}})^2/2$ at the current iterate. Above this local threshold the chosen scalar is independent of δ. Different radii can nevertheless produce different trajectories, and an overly small radius still limits progress. This local fact does not compare final performance with the best radius of another algorithm.

What the objective certificate means. A valid global lower bound already ensures $L \geq L_{\mathrm{lb}}$ for every policy. The additional restriction is that the *lower Taylor model* respect this bound. It is a conservative model-based step criterion, not a new proof that the true objective has a lower bound. Nor does it alone ensure descent: a loose bound makes the objective gap large. For example, with $L(\theta) = \theta^2/2$, $\theta = 1$, $F = \beta = 1$, $L_{\mathrm{lb}} = -100$, and $\delta = 100$, the certified root is $\sqrt{202} - 1 > 13$, which increases L after the update. A descent line search, or the additional cap $d/(\beta q)$, can enforce descent when the smoothness bound is valid. The latter changes the maximality criterion in Theorem 1.

3.1 SENSITIVITY TO INFORMATION SUPPLIED TO THE STEP

Curvature. For $\Delta > 0$ and $k \geq 1$, the same root obeys

$$\frac{1}{\sqrt{k}} \leq \frac{\eta^{\mathrm{smPol}}(k\beta)}{\eta^{\mathrm{smPol}}(\beta)} \leq 1. \tag{7}$$

This follows directly by bounding $\sqrt{d^2 + 2k\beta q\Delta}$ between $\sqrt{d^2 + 2\beta q\Delta}$ and $\sqrt{k}\sqrt{d^2 + 2\beta q\Delta}$. Overestimation preserves the certificate but shortens the step. Underestimation need not preserve it: using $\widehat{\beta} < \beta$ can lower the valid Taylor bound by $(\beta - \widehat{\beta})q\eta^2/2$. The scalar β is therefore required problem information for the certified smooth rule.

Objective target. Implicit differentiation of $d\eta + (\beta q/2)\eta^2 = \Delta$ gives $\partial\eta/\partial\Delta = 1/(d + \beta q\eta) > 0$. A more conservative lower bound increases the cap and can make the trust-region branch bind more often. A target that is too optimistic about the attained loss can shrink or zero the empirical gap. Neither case is covered by a rate theorem that assumes $L_{\mathrm{lb}} = L^\star$.

An analytic bound is available, for example, from bounded nonnegative rewards: $L_{\mathrm{lb}} = -R_{\max}/(1-\gamma)$ for infinite-horizon discounted return, though it can be very loose. A running-best reward estimate plus a sampling margin bounds rewards of *visited* policies under suitable sampling assumptions; it does not upper-bound $\sup_\theta J_r(\pi_\theta)$. Theorem 5 in the appendix makes this distinction explicit. Such an estimate is a practical proxy, not a certified objective lower bound.

3.2 Convergence scope and the role of Fisher geometry

For exact gradients and $L_{\mathrm{lb}} = L^\star$, write $\Delta_t = L(\theta_t) - L^\star$, $\chi_t = \beta q_t \Delta_t / d_t^2$, and $a_t = 2/(1 + \sqrt{1 + 2\chi_t})$. On a Polyak-binding iteration, $\eta_t = a_t \Delta_t / d_t$, and smoothness gives the explicit bound

$$\Delta_{t+1} \le \left(1 - a_t + \tfrac{1}{2}\chi_t a_t^2\right) \Delta_t = 2(1 - a_t)\Delta_t. \tag{8}$$

If $\chi_t \le \bar{\chi} < 4$ and the Fisher-norm PL inequality $d_t \ge 2\mu_{\mathrm{PL}}\Delta_t$ hold at every nonstationary iterate, the exact-oracle method converges geometrically for every fixed $\delta > 0$, across both branches and their transition band. The following explicit form makes the radius dependence visible. For $0 < \bar{\chi} < 4$ and $\Delta_0 > 0$, define

$$\begin{aligned} c &= \frac{2}{1 + \sqrt{1 + 2\bar{\chi}}}, \qquad h = \frac{3 - \sqrt{1 + 2\bar{\chi}}}{2}, \\ \underline{a} &= \min\left\{c, 2\sqrt{\mu_{\mathrm{PL}}\delta/\Delta_0}\right\}, \qquad \rho = 1 - h\underline{a} \in (0, 1). \end{aligned} \tag{9}$$

Then $\Delta_t \le \rho^t \Delta_0$. In particular, entry into the sufficient objective-binding region $\Delta_t \le 4\mu_{\mathrm{PL}}\delta$ occurs no later than

$$\left\lceil \frac{\max\{0, \log(\Delta_0/(4\mu_{\mathrm{PL}}\delta))\}}{\log(1/\rho)} \right\rceil. \tag{10}$$

Thereafter the factor improves to $2(1 - c) < 1$, independently of the radius. To see why progress also holds before entry, put $b_t = \eta_t d_t / \Delta_t \le a_t$. The root identity implies $1 - \chi_t b_t/2 \ge h > 0$, so Taylor descent gives $\Delta_{t+1} \le (1 - h b_t)\Delta_t$. This first bounds all gaps by Δ_0; PL then gives $b_t \ge \underline{a}$, yielding the rate. Theorem 9 supplies the full proof and the separate sufficient trust-binding region. These are upper bounds on entry, not a prediction of the actual first switch. Appendix A.8 gives a convex variable-metric rate with controlled Fisher drift. These are conditional results, not claims that neural policies satisfy PL, that a practical target converges to the optimum, or that adding a cost cap preserves the same rates. A cost cap can stop progress at a constraint boundary.

Chentsov's classical theorem characterizes Fisher–Rao up to scale within a family of ambient statistical metrics invariant under sufficient Markov transformations. Under that ambient construction, pullback to each state-conditional policy and fixed-reference averaging give cF. This motivates the Fisher direction (Theorem 4 in the appendix); it does not select the objective bound, smoothness model, or finite update. In particular, $\beta\|F^{-1}g\|_2^2$ uses Euclidean coordinates and an additive step is not invariant under a nonlinear reparametrization. We make no coordinate-invariance claim for the complete CP-SPO algorithm.

4 A separate cap for cumulative-cost chance feasibility

Theorem 2 (Feasibility from a certified feasible start)**.** *Let $0 < \varepsilon < 1$. Suppose $Z(\theta) - \mathbb{E}Z(\theta)$ is sub-Gaussian with a certified proxy $\sigma_c^2(\theta)$, where $\sigma_c \ge 0$, and let $z_\varepsilon = \sqrt{2\log(1/\varepsilon)}$. Define*

$$\tilde{c}(\theta) = \mathbb{E}Z(\theta) + z_\varepsilon \sigma_c(\theta) - C. \tag{11}$$

Assume $\tilde{c}(\theta_0) \le 0$ and a certified bound $\|\nabla^2 \tilde{c}\| \le L_{\tilde{c}}$ on every step segment. For any proposed direction v, let $s = \nabla\tilde{c}^\top v$ and $q = \|v\|^2$. At a feasible iterate with $L_{\tilde{c}} q > 0$, set

$$\eta^{\mathrm{safe}} = \frac{s + \sqrt{s^2 - 2L_{\tilde{c}} q\, \tilde{c}(\theta)}}{L_{\tilde{c}} q}. \tag{12}$$

When $L_{\tilde{c}} q = 0$, use $\eta^{\mathrm{safe}} = +\infty$ if $s \ge 0$, and $\eta^{\mathrm{safe}} = \tilde{c}/s$ if $s < 0$. If $v = 0$, take no step. Any update $\theta^+ = \theta - \eta v$ with $0 \le \eta \le \eta^{\mathrm{safe}}$ preserves $\tilde{c} \le 0$. Therefore wrapping successive candidate steps by this cap gives

$$\mathbb{P}_{\mu_{\pi_{\theta_t}}}\{Z(\theta_t) > C\} \le \varepsilon \quad \textit{for every iterate } t. \tag{13}$$

Proof. Chernoff's bound gives the chance constraint from $\tilde{c} \le 0$; the zero-variance case is deterministic for the strict event in (13). Taylor's bound gives $\tilde{c}(\theta - \eta v) \le \tilde{c}(\theta) - \eta s + L_{\tilde{c}} q \eta^2/2$. Its nonpositive interval containing zero ends at (12); the affine case gives the stated conventions. Induction from the feasible start completes the proof. □

Component	Information supplied	Statement supported
Movement	Current Fisher metric and radius	Quadratic movement bound; measured finite-step KL is checked separately.
Objective	Gradient, valid lower bound, and smoothness bound	Lower-Taylor admissibility along the fixed direction; descent needs its stated additional conditions.
Practical target	Estimated return, running best, margin, and EWMA	Empirical candidate length; the exact specialization has separate stationarity assumptions.
Chance constraint	Feasible tightened cost, sub-Gaussian proxy, gradient, and curvature bound	Per-policy trajectory-cost feasibility after applying the cost cap.

Table 1: Inputs and scope of the step components. A practical reward target does not supply the information required by the certified objective or cost statements.

The probability here is over a new trajectory at a fixed policy. An ε bound for each iterate does not imply that the probability of any violation over an entire training run is at most ε. The theorem also does not turn point estimates of mean cost, its gradient, or curvature into confidence bounds. Statistical certification requires simultaneous valid bounds and an explicit failure budget. An infeasible initial policy needs a separate restoration method; setting the step to zero or using a Lagrangian fallback is not such a guarantee.

Information requirements and composition. The movement, objective, and cost conditions answer different questions about one proposed update. The Fisher metric and radius specify a local movement budget. Objective information determines whether the remaining gap calls for a shorter step within that budget. Cost information determines whether a candidate preserves the tightened constraint. Table 1 separates these inputs from the empirical target used when a certified objective bound is unavailable.

Each certified scalar condition defines an admissible interval starting at zero. Their intersection can therefore be implemented by taking the smallest endpoint, provided all certificates refer to the same current policy and proposed direction. This explains the modular role of the cost cap: it can further restrict an objective-scaled candidate without changing the objective root. If a controller replaces the direction, the directional certificates must be evaluated for that replacement before invoking a joint guarantee. The practical controller in Algorithm 1 is evaluated empirically under its own protocol.

5 Practical implementation

The reported RL experiments use the first-order objective cap $\eta_t^{\mathrm{P}} = (\widehat{L}_t - \widetilde{L}_t)_+/\widehat{d}_t$, a running-best reward proxy $\widetilde{L}_t$, and a Lagrangian fallback. They do not use a certified L_{lb}, the finite-curvature root, or the chance-safety cap. The primary implementation caps the reward candidate before cost control; standalone branch labels describe the later control stage. Algorithm 1 gives this ordering. The matched cap-only ablation instead holds an augmented reward–cost controller fixed and toggles only the cap. In loss convention its direction uses $g_L + \lambda \nabla J_C$, and subtracting that natural gradient is equivalent to ascending reward minus the cost penalty. The detailed estimator settings and experiment-specific limitations are in Appendix B.1.

Proposition 1 analyzes an exact, unsmoothed reward-only specialization of the stated running-best target. For a smooth lower-bounded loss, bounded metrics $\mu_F I \preceq F_t \preceq M_F I$, and true-loss Armijo acceptance, consider the exact target

$$J_t^+ = \max_{s \le t} J_s + m_t, \qquad m_t = \alpha_{\mathrm{margin}} \sqrt{\log(t+e)/(t+1)}.$$

Accepting the first Armijo-passing trial, or retaining the policy when all trials fail, makes the true return nondecreasing. Thus the target gap is exactly m_t. With Armijo coefficient $c_A \in (0,1)$, let

$$D_T = \max\left\{ \frac{\beta m_T}{2(1-c_A)\mu_F}, \frac{m_T^2}{2\delta} \right\}. \tag{14}$$

Algorithm 1: Practical CP-SPO reward candidate and subsequent cost control

Input: radius δ, empirical reward target $\widehat{J}^{+}$, mean-cost budget C, controller state

Estimate reward $\widehat{J}$, loss gradient $\widehat{g}_L$, cost statistics, and positive-definite damped Fisher $\widehat{F}$;

Compute $v \approx \widehat{F}^{-1}\widehat{g}_L$ by conjugate gradients; $d \leftarrow \widehat{g}_L^{\top} v$;

if $d \leq 0$ **then**

  retain θ and stop this iteration;

Update the empirical reward target using the specified EWMA and margin;

$\eta^{\mathrm{P}} \leftarrow [\widehat{J}^{+} - \widehat{J}]_{+}/d$; $\eta^{\mathrm{TR}} \leftarrow \sqrt{2\delta/d}$;

$\eta \leftarrow \min\{\eta^{\mathrm{P}}, \eta^{\mathrm{TR}}\}$;

Pass $\theta - \eta v$ to the experiment's mean-cost controller, updating its dual state as specified by that protocol;

Apply surrogate and measured-KL backtracking checks to the controlled candidate; retain θ if rejected;

Whenever $c_A T m_{2T} > L(\theta_0) - L_{\inf}$, $\min_{T \le t < 2T} d_t < D_T$ and hence $\min_{T \le t < 2T} \|\nabla L(\theta_t)\|^2 < M_F D_T = O(\sqrt{\log T / T})$. Otherwise every iterate in that window would accept its first candidate and decrease loss by at least $c_A m_t$, exceeding the total loss available above $L_{\inf}$. The argument does not need $L^{\star}$ or an infinite backtracking budget. This is a stationarity result for that idealization; it does not certify the noisy target, empirical acceptance tests, or post-cap cost controller used in the RL runs.

Which step is being measured? The practical update has three distinguishable stages: forming an objective-capped candidate, applying the cost controller, and accepting or shortening the controlled proposal. A candidate ratio measures the first stage; an accepted displacement measures the result of all three. A later controller-branch label consequently does not determine whether the earlier objective cap was active. Conversely, an active cap need not change the accepted update if subsequent control or backtracking removes the difference. Recording both stages makes selective activation interpretable without treating every capped proposal as a distinct policy update.

The target settings also have different operational roles. The EWMA changes the temporal weighting of the return estimates; the margin changes the separation between the running best and the target. At fixed current return and direction, increasing the target enlarges the first-order cap. Fisher damping and the solve tolerance instead affect the direction and its normalization. The reported target components and trainer-specific margin coefficients are retained in Appendix B.1; the EWMA comparison changes only its reported coefficient. These choices belong to the implemented procedure even when the objective cap, rather than the radius, limits a candidate.

Remaining settings and computational cost. The reported shared radius is $\delta = 0.01$; return-target smoothing uses an EWMA coefficient 0.9, which is *not* the objective smoothness constant β. Target margins, Fisher damping, dual steps, and line-search tolerances remain operational choices. For P parameters and K conjugate-gradient iterations, both CP-SPO and its TRPO-Lagrangian backbone use $O(P)$ update-vector memory and $O(KC_{\mathrm{FVP}})$ Fisher-solve work. The first-order cap adds scalar arithmetic after the shared solve; the smooth root additionally uses $q = \|v\|^2$, one $O(P)$ reduction. Computing a *certified* curvature or cost bound can add problem-specific work not included in this scalar-overhead accounting.

The configured comparison uses 4,096 environment samples per update, one reward and one cost gradient, one natural-gradient solve with at most ten Fisher-vector products, at most ten backtracking trials, and twenty value-update iterations. These are matched configured budgets, not measured equal FLOPs. Reported median seconds per thousand steps are $2.75/2.75$ (CP-SPO/TRPO-Lagrangian) on PG1, $2.96/2.90$ on CG1, and $2.81/2.80$ on PP1. These small-policy timings do not establish scaling to large language models or high-dimensional control.

PointGoal1: cap on versus cap off	Relative change
Mean return over the final 50k steps	$+8.6\%$
Return area under the training curve	$+8.1\%$
Cumulative cost-excess area	-5.7%
Across-seed return standard deviation	$+10\%$

Table 2: Matched cap-only ablation with three paired seeds and 500k steps. Direction, controller, target, radius, and acceptance checks are held fixed. These descriptive changes are not significance tests.

6 Experiments and evidence limits

We distinguish synthetic mechanism tests from practical RL evaluation. The synthetic experiments isolate the roles of the objective cap and cost certificate; the RL experiments evaluate the practical proxy variant. Appendix B gives the configurations, additional baselines, and protocol-specific results.

What each comparison isolates. The matched cap-only study asks whether enabling the cap changes learning under a fixed augmented controller. Its paired design keeps the backbone and configured budget common, so the contrast concerns that intervention within this protocol. It does not distinguish objective-dependent shortening from a different rule that shortens comparable proposals. That distinction requires an explicit generic shortening control with its selection procedure and accepted-step behavior reported.

The radius-selected comparison asks a different practical question: how the shared CP-SPO configuration compares with a baseline selected from the stated radius sweep. The separate seven-method suite compares its own common configuration and evaluation window. Neither substitutes for the within-controller ablation. Finally, the finite-curvature study directly evaluates a model-certificate residual, while the synthetic cost study evaluates policy-level chance infeasibility. Keeping these endpoints distinct connects each experiment to the component it tests. For policy learning, return and cost are read together: lower cost can accompany lower return, and reduced cost excess can remain above the budget. The paired cap comparison and the cross-task performance table therefore provide complementary evidence.

Controlled cap-only comparison. Objective-gap clipping improves return and reduces cumulative cost excess under a fixed controller on PointGoal1, even though only a small fraction of accepted updates differ. Table 2 is the direct test of this selective effect; it is a one-task result, not evidence that objective-aware clipping dominates other ways of shortening a step.

Selective activity under cost control. At 500k steps, reported phase-wise Polyak activation is 0–7%; the Lagrangian fallback accounts for 51–93.7% of updates across tasks and phases (Table 10 in the appendix). Thus much of training uses the Lagrangian controller. These branch counts do not measure whether the cap changed a candidate before that controller. Applied steps differed on 7.0% of accepted controller-active updates; the median of the three seed-level median candidate ratios was 0.83. The across-seed return-standard-deviation ratio was 1.10, so this ablation does not support a variance-reduction claim. It does show that selective activation can change performance on this task.

Author draft: supplied summaries; measurement and protocol confirmation pending.

Objective-aware clipping versus generic shortening. This comparison asks whether objective-dependent shortening improves learning beyond a generic shortening rule under the same augmented reward–cost controller. The reference setup is the matched PointGoal1 study in Table 2: three paired seeds, 500k environment steps, and final-50k return. Its reported cap-on/off changes are already given there. The separately supplied generic-control summaries appear in Table 3. Shared implementation settings and the stated reward target are in Appendix B.1. Appendix **??** records only the new control, any protocol differences, and its shortening diagnostics.

Metric	Supplied preliminary comparison
Final-50k return	20.5 ± 1.3 (cap); 20.7 ± 1.3 (generic); difference -0.2
Cost-excess area	-5.2% relative to generic control

Table 3: Author-supplied preliminary generic-control summaries. Return entries are mean $\pm$ across-seed SD; the difference is objective cap minus generic control. The cost-excess-area reduction is 5.2% against generic shortening, distinct from the previously reported 5.7% against cap-off. The reference budget is 25 per episode. Measurement confirmation, actual protocol, and paired uncertainty remain to be completed in Appendix **??**; these summaries are not significance tests.

	PG1	CG1	PP1
TRPO-Lag return ($\delta = 0.05$)	24.65 ± 0.04	28.19 ± 1.19	1.38 ± 0.51
CP-SPO return ($\delta = 0.01$)	23.86 ± 0.49	26.13 ± 1.61	0.56 ± 0.05
CP-SPO / TRPO-Lag return	96.8%	92.7%	40.6%
TRPO-Lag mean cost	53.9	70.5	62.3
CP-SPO mean cost	53.2	69.0	41.7
Cost budget	25	25	25

Table 4: Reported 1M-step comparison, three seeds; returns are mean $\pm$ standard deviation. The shared-radius method retains 92.7–96.8% of navigation return and 40.6% of pushing return relative to the selected baseline. All six mean costs exceed the budget; costs are reported as means.

The supplied point estimates describe a return–cost tradeoff: the cap has 0.2 lower mean return (approximately 1%) and 5.2% lower cost-excess area than generic shortening. Paired uncertainty and shortening diagnostics are needed to assess this difference and its attribution. These summaries do not establish statistical significance, dominance, or chance feasibility.

Real-RL evaluation. We evaluate PointGoal1 (PG1), CarGoal1 (CG1), and PointPush1 (PP1) in Safety-Gymnasium (Ji et al., 2023), with episodic mean-cost budget 25. The reported policy/value networks have two 64-unit tanh hidden layers. CP-SPO uses one shared $\delta = 0.01$. The 500k-step TRPO-Lagrangian sweep uses $\delta \in \{0.001, 0.01, 0.05, 0.1\}$; its selected 0.05 configuration is extended to 1M steps. This is a radius-selected baseline, not a new complete 1M radius sweep. Table 4 gives the three-seed extension, averaging the final 50k steps within each run.

The navigation ratios indicate a useful operating point for the shared radius, while PP1 shows that this choice can sacrifice substantial return. Its lower mean cost (41.7 versus 62.3) is still above budget and does not establish deployable safety. A separate PP1 seven-method study at default settings reports CP-SPO return 0.58 ± 0.02 and cost 58.2 ± 12.0, with TRPO-PID return 0.66 ± 0.29. That study uses a different reported variant and averaging window (Table 15 in the appendix); its $0.66/0.58 \approx 1.14$ ratio is not the ratio against the radius-selected baseline in Table 4. We do not combine these protocols into one performance claim.

Sensitivity to return smoothing. On PG1 with 500k steps and three matched seeds, changing the return EWMA coefficient from 0.9 to 0.5 changed return, cost-excess area, and return standard deviation by -10.4%, $+11.8\%$, and $+12.3\%$, respectively. At 0.99 the changes were -4.1%, $+7.2\%$, and $+13.0\%$. Thus sharing a radius does not eliminate sensitivity to the objective target; this limited study favors the prespecified 0.9 setting, without establishing a universal choice.

Synthetic step and radius tests. The two-dimensional oracle quadratic (E1, 15 initializations) reports CP-SPO final gap $\leq 10^{-10}$ versus 2.4×10^{-2} for its fixed-radius TRPO comparator. The noisy four-dimensional quadratic (E2, 20 seeds) reports a robustness ratio 1.04 for CP-SPO versus 47.31 for TRPO, computed over the oversized-radius subset $\delta \geq 0.1$ of the full $[0.01, 1]$ sweep. These are the reported first-order-cap experiments, not direct tests of the finite-curvature certificate. The 50-dimensional phase study (E4) reports contraction 0.78 ± 0.03 and a crossing at 12 ± 1.5 iterations. The conditional phase theorem does not predict an exact crossing time or a universal activation percentage, so these are mechanism diagnostics rather than confirmations of those stronger claims. A separate finite-curvature study uses the exact smooth root on the E1 quadratic: seven smoothness multipliers, five radii, and 15 deterministic starts give 525 runs and 31,500 updates. Every valid

overestimate preserved the lower-Taylor certificate (minimum residual -1.55×10^{-15}, tolerance -10^{-10}). Multipliers 1, 2, and 4 reached gap 10^{-10} for all starts at $\delta \in \{0.1, 0.3, 1.0\}$ in median ranges 30–32, 37–39, and 48–50 iterations. A $10\times$ estimate was slower, with median terminal gaps 3.16×10^{-9}–4.53×10^{-9} after 60 iterations; a $0.5\times$ underestimate violated the certificate on 37% of updates at $\delta = 0.3$. This directly tests the one-sided sensitivity in (7).

Synthetic cost cap. The E5 report records zero chance-infeasible iterates among 20,000 tested with the cap, versus $6,412$ without it (4-D Gaussian cost model, $\varepsilon = 0.05$, ten seeds, 2,000 updates per seed). An iterate is counted as infeasible when its trajectory-cost violation probability exceeds ε; this is distinct from counting violating episodes. The case-complete follow-up, including the affine-curvature and boundary branches of the cap, also reports zero among 20,000 iterates, at a 5.9% final-reward penalty. This is a separate follow-up protocol. These controlled tests support the cost-cap mechanism; they do not certify the practical RL runs.

Reading the theoretical and practical evidence together. The local theory explains which information can justify shortening a proposed step. The experiments ask whether this shortening changes learning once estimation, cost control, and acceptance are present. These questions have different units of analysis: a synthetic residual is checked at an update, a return curve summarizes a training run, and across-seed variation describes repeated runs. An update-level certificate does not establish a statistical performance advantage; better final return does not validate every hypothesis of the certificate. Keeping these levels explicit connects the theory to the experiments without transferring a guarantee between them.

Implications for choosing a policy update. The matched cap-only result motivates evaluating objective scaling as a separately testable part of the controller. A shared radius fixes a configured movement budget, while the objective target and controller can still change accepted movement. The practical operating point must therefore be assessed through return and cost jointly at the stated interaction budget. In the supplied navigation runs, the common setting retains much of the selected baseline's return; pushing exposes a substantial return penalty alongside lower cost. These outcomes do not specify an optimal return–cost preference. Applying the method under a chance-cost requirement would additionally need the certified cost inputs and feasible initialization of Theorem 2. Those conditions are distinct from the empirical mean-cost controller evaluated here.

Limitations. Three small-policy tasks and three-seed primary comparisons do not establish broad RL robustness. The certified smooth rule and practical proxy variant require distinct evaluations; the neural-policy PL and chance-certificate assumptions are unverified here. The current study does not establish a universal radius, a bound relative to the best TRPO trajectory, or chance-feasible learning on Safety-Gymnasium.

7 Conclusion

Objective information can tighten a step within a fixed quadratic trust region. CP-SPO characterizes this interaction through a closed-form minimum and gives a separate cost-feasibility cap under additional assumptions. The matched PointGoal1 ablation shows that selective use of the practical cap can affect learning. The sizeable pushing return gap and persistent cost-budget excess delimit the current evidence; broader evaluation and certified real-RL cost control remain open.

REPRODUCIBILITY STATEMENT

Proofs appear in Appendix A; configurations, seed counts, training horizons, averaging windows, and additional results appear in Appendix B. We used Lean 4.33.1/mathlib to check the proofs (kernel-checked support for the scalar certificates, finite-simplex Chentsov theorem, KL/Fisher identities, adaptive confidence bounds, and changing-metric convergence).

AI USE STATEMENT

Generative AI was used for language editing, text polishing, literature research. No new RL training results were generated as part of this assistance. AI also assisted with Lean formalization; the submitted proof files are checked by Lean's kernel and include an axiom-dependency audit.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *ICML*, 2017.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. AMS & Oxford University Press, 2000.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. Information geometry and sufficient statistics. *Probability Theory and Related Fields*, 162:327–364, 2015.

Martin Bauer, Martins Bruveris, and Peter W. Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

Daniel Bienstock, Michael Chertkov, and Sean Harnett. Chance-constrained optimal power flow: Risk-aware network control under uncertainty. *SIAM Review*, 56(3):461–495, 2014. doi: 10.1137/130910312.

Nikolai N. Chentsov. *Statistical Decision Rules and Optimal Inference*. Nauka, Moscow, 1972. In Russian. English translation: Translations of Mathematical Monographs, vol. 53, American Mathematical Society, 1982.

Elad Hazan and Sham Kakade. Revisiting the Polyak step size. *arXiv preprint arXiv:1905.00313*, 2019. URL https://arxiv.org/abs/1905.00313. Revised August 2022.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety-gymnasium: A unified safe reinforcement learning benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Sham Kakade. A natural policy gradient. In *NeurIPS*, 2002.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–łojasiewicz condition. In *ECML-PKDD*, 2016.

Yunxiang Li, Rui Yuan, Chen Fan, Mark Schmidt, Samuel Horváth, Robert M. Gower, and Martin Takáč. Enhancing policy gradient with the Polyak step-size adaption. *arXiv:2404.07525*, 2024.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *AISTATS*, 2021.

Nikola Milosevic, Johannes Müller, and Nico Scherf. Embedding safety into RL: A new take on trust region methods. In *ICML*, 2025. arXiv:2411.02957.

Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007. doi: 10.1137/050622328.

Francesco Orabona and Ryan D'Orazio. New perspectives on the Polyak step-size: Surrogate functions and negative results. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5196. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e45879046fd900c2536e419e361c94c0-Abstract-Conference.html.

Antonio Orvieto, Simon Lacoste-Julien, and Nicolas Loizou. Dynamics of SGD with stochastic Polyak stepsizes: Truly adaptive variants and convergence to exact solution. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software Inc., 1987.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *ICML*, 2015.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *ICML*, 2020.

Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Parameter-free clipped gradient descent meets Polyak. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4ebba705ffdee81e0a638c99fe066ce2-Abstract-Conference.html.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *ICLR*, 2019.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. Projection-based constrained policy optimization. In *ICLR*, 2020.

A THEORETICAL SUPPORT

This appendix supplies the local step derivation, the classical ambient Fisher characterization, conditional convergence results, and explicit oracle-error transfer bounds. The pointwise radius-insensitivity result does not compare final loss to a radius-selected competing algorithm. The objective and movement results are separate from the cost-feasibility theorem in the main text. The geometric discussion is limited to the classical ambient Fisher characterization.

A.1 SETTING: FULL STATEMENTS DEFERRED FROM SEC. 2

CMDP details. The CMDP $\mathcal{M}=(\mathcal{S},\mathcal{A},P,r,c,\gamma,\rho_0)$ has state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P:\mathcal{S}\times\mathcal{A}\to\mathcal{P}(\mathcal{S})$, scalar reward $r:\mathcal{S}\times\mathcal{A}\to[0,R_{\max}]$, scalar cost $c:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$, discount $\gamma\in(0,1)$, initial-state distribution ρ_0. A stochastic policy $\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})$ induces the trajectory probability measure μ_π on $(s_0,a_0,s_1,a_1,\ldots)$ via $s_0\sim\rho_0$, $a_t\sim\pi(\cdot|s_t)$, $s_{t+1}\sim P(\cdot|s_t,a_t)$. The discounted expected reward and cost are $J_r(\pi)=\mathbb{E}_{\mu_\pi}\big[\sum_t\gamma^t r(s_t,a_t)\big]$, $J_c(\pi)=\mathbb{E}_{\mu_\pi}\big[\sum_t\gamma^t c(s_t,a_t)\big]$. The cumulative-cost chance constraint is enforced through a certified upper-tail surrogate, as specified in Theorem 2. An episode-wise event $\{c(s_t,a_t)>0 \text{ for some } t\le H\}$ is a different constraint: for finite H, one sufficient surrogate is $\mathbb{E}[\sum_{t\le H}h(c(s_t,a_t))]\le\varepsilon$, where $h(u)\ge\mathbf{1}\{u>0\}$. A bound on discounted cumulative cost does not in general imply this episode-wise guarantee. For bounded nonnegative rewards the analytic objective lower bound $L_{\mathrm{lb}}=-R_{\max}/(1-\gamma)$ is valid globally. It need not equal the optimal loss $L^\star$. A running-best reward estimate is only a diagnostic and does not replace a global lower-bound certificate.

Parametric policies and trust-region update. Consider a dominated parametric family $\{\pi_\theta\}_{\theta\in\Theta}$, $\Theta\subset\mathbb{R}^n$ open, with common positive action support, C^2 log densities, and finite score second moments. For the local KL–Fisher identification, assume differentiation under the action and frozen-reference state integrals is valid through second order, and the reference-average KL is C^2 locally. These are model regularity assumptions; see Amari & Nagaoka (2000, Secs. 2.1–2.2) for background. The matrix in (2) averages, under the fixed reference state distribution, the pull-backs of the state-conditional Fisher metrics (Ay et al., 2015). The TRPO step $\eta^{\mathrm{TR}}=\sqrt{2\delta/d(\theta)}$ comes from solving $\frac{1}{2}\eta^2 d(\theta)\le\delta$ for the largest admissible η (Schulman et al., 2015).

A.2 CHENTSOV'S THEOREM: FULL STATEMENT AND PARAMETRIC PULL-BACK

For a finite outcome space, write $\mathcal{P}_+(\mathcal{X})$ for the interior probability simplex. In score coordinates its tangent vectors are mean-zero functions, and $g_P^{\mathrm{FR}}(u,v)=\mathbb{E}_P[uv]$.

Theorem 3 (Chentsov uniqueness on finite simplexes)**.** *Consider a smooth family of Riemannian metrics, one on each finite interior probability simplex, invariant under sufficient Markov embeddings between these simplexes. Then the family equals $c\,g^{\mathrm{FR}}$ for one positive constant c. Consequently, its pull-back to a regular parametric model is c times that model's Fisher information matrix (Chentsov, 1972; Ay et al., 2015).*

This is an ambient statistical invariance statement. Coordinate covariance on a single parameter space does not select Fisher. On general outcome spaces, Ay et al. (2015) characterize compatible local, continuous quadratic-form fields across their class of integrable statistical models under sufficient-statistic invariance. Smoothness and identifiability of one finite-dimensional policy alone do not establish those hypotheses. The separate result of Bauer et al. (2016) characterizes smooth weak Riemannian metrics on the full space of smooth positive probability densities, invariant under diffeomorphisms of a connected compact manifold without boundary of dimension at least two. Neither result asserts invariance of a finite coordinate update with a Euclidean smoothness bound.

A.3 NOTATION TABLE

Policy geometry	
$F(\theta)$	Fisher information matrix of parametric π_θ
CP-SPO step quantities	
$d := \nabla L^\top F^{-1} \nabla L$	Fisher gradient norm squared
$\tilde{d} := \nabla L^\top F^{-2} \nabla L$	Auxiliary form, with $\tilde{d}/d \le 1/\mu$ via $F \succeq \mu I$
δ	Quadratic Fisher movement radius
$\eta^{\mathrm{smPol}}, \eta^{\mathrm{TR}}$	Smooth Polyak certificate (15) and TR step $\sqrt{2\delta/d}$
$\eta^\star_{\mathrm{FR}} := \min(\eta^{\mathrm{smPol}}, \eta^{\mathrm{TR}})$	CP-SPO step (Theorem 1)
L_{lb}	Certified lower bound used by Thm. 1
$\widehat{J}_t^+$	Best-seen reward bound under the concentration assumptions (Theorem 5)
ε	Chance-constraint violation budget
C, τ, λ_t, ρ	Cost limit, Lagrangian threshold, multiplier, dual-step
Convergence-rate quantities	
$\Phi_t := \frac{1}{2}\|\theta_t - \theta^\star\|^2_{F_t}$	Variable-Fisher Lyapunov value
$\Gamma_T := \prod_t (1+\nu_t)$	Fisher metric-drift product in Theorem 6
$\Delta_t := L(\theta_t) - L^\star$	Suboptimality at iteration t
μ, β	Fisher PD lower bound and gradient-Lipschitz constant
μ_{PL}	PL modulus (Theorem 9); $d \ge 2\mu_{\mathrm{PL}}\Delta$
$\rho_{\bar{\chi}}$	Conditional Polyak-region contraction in Theorem 9
$[4\mu_{\mathrm{PL}}\delta,\ 4L_F\delta/c^2_{\bar{\chi}}]$	Sufficient PL transition band
M, σ_R	Per-iteration sample budget and sub-Gaussian return parameter
$\delta_{\mathrm{crit}}(\theta)$	Local threshold $d(\eta^{\mathrm{smPol}})^2/2$ for objective-cap binding
C^{emp}	Descriptive empirical return ratio; not a proved approximation factor

A.4 REGULARITY ASSUMPTIONS AND NOTATION

Assumption 2 (Regularity and scope)**.** *For probability calculations the sample spaces are standard Borel. Sufficient kernels admit a reverse Markov kernel recovering every distribution in the specified statistical experiment. For the step certificate, L is β-smooth on every trial segment and $F \succeq \mu_F I$ on the iterate region. A cubic KL remainder additionally requires third-order regularity; with only twice differentiability the remainder is $o(\|\theta' - \theta\|^2)$. Ambient metric uniqueness uses the separate hypotheses of Theorem 3, not just these optimization assumptions.*

The loss is in minimization convention, $L = -J_r$. L_{lb} denotes a certified lower bound and $L^\star$ denotes the global optimal value in the objective-suboptimality analyses. Those analyses explicitly require $L_{\mathrm{lb}} = L^\star$ and do not cover a boundary-constrained optimum, a changing Lagrangian, an active safety cap, or an uncertified reward proxy. Proposition 1 instead uses a shrinking margin and proves stationarity without an optimal target, under its separate exact-oracle and acceptance assumptions. If the gradient vanishes, the algorithm takes zero displacement; formulas dividing by $d = \nabla L^\top F^{-1} \nabla L$ are used only when $d > 0$.

A.5 FIRST-PRINCIPLES BUILDING BLOCKS

The following standard inequalities support the later derivations. Their roles are smoothness control, local KL geometry, the scalar root, and uniform concentration over policies actually evaluated.

A.5.1 LEMMA D1 — DESCENT LEMMA IN THE FISHER METRIC

Lemma 1. *Let $L : \mathbb{R}^n \to \mathbb{R}$ be β-smooth in the Euclidean metric: $\|\nabla L(\theta_1) - \nabla L(\theta_2)\| \le \beta\|\theta_1 - \theta_2\|$ for all θ_1, θ_2. Let $F(\theta) \succ 0$ be a smooth Riemannian metric on $\mathbb{R}^n$ with $F \succeq \mu I$. For any step $\eta \ge 0$ and direction $\nabla L(\theta)$:*

$$L\big(\theta - \eta F^{-1}\nabla L\big) \ \le\ L(\theta) - \eta\, d + \tfrac{\beta\,\eta^2}{2}\,\tilde{d}, \quad \textit{with } d := \nabla L^\top F^{-1}\nabla L,\ \tilde{d} := \nabla L^\top F^{-2}\nabla L.$$

Proof. Let $v = F(\theta)^{-1}\nabla L(\theta)$. The fundamental theorem of calculus gives the exact identity

$$L(\theta - \eta v) = L(\theta) - \eta \nabla L(\theta)^\top v - \int_0^\eta [\nabla L(\theta - sv) - \nabla L(\theta)]^\top v\, ds.$$

The absolute value of the integrand is at most $\beta s\|v\|^2$. Integration gives a remainder of absolute value at most $\beta\eta^2\|v\|^2/2$, proving both the displayed upper bound and the analogous lower Taylor bound.

□

A.5.2 LEMMA D2 — KL SECOND-ORDER TAYLOR FOR EXPONENTIAL FAMILIES

Lemma 2. *For a regular exponential family $\pi_\theta(x) = \exp\big(\theta^\top T(x) - A(\theta)\big)h(x)$ on an open subset $\Theta \subset \mathbb{R}^n$ of its natural parameter domain, with common carrier h and C^3 log-partition A, the KL divergence has the second-order expansion*

$$\mathrm{KL}(\pi_\theta \,\|\, \pi_{\theta'}) = \tfrac{1}{2}(\theta - \theta')^\top F(\theta)(\theta - \theta') + O\big(\|\theta - \theta'\|^3\big),$$

where $F(\theta) = \nabla^2 A(\theta)$ is the Fisher information.

Proof. By definition, $\mathrm{KL}(\pi_\theta\|\pi_{\theta'}) = \mathbb{E}_{\pi_\theta}[\log(\pi_\theta/\pi_{\theta'})] = \mathbb{E}_{\pi_\theta}[(\theta-\theta')^\top T(x)] - A(\theta) + A(\theta')$. The MGF property of exponential families gives $\mathbb{E}_{\pi_\theta}[T(x)] = \nabla A(\theta)$, so $\mathrm{KL} = (\theta - \theta')^\top \nabla A(\theta) - A(\theta) + A(\theta')$. Taylor-expand A at θ to third order: $A(\theta') = A(\theta) + (\theta' - \theta)^\top \nabla A(\theta) + \frac{1}{2}(\theta' - \theta)^\top \nabla^2 A(\theta)(\theta' - \theta) + O(\|\theta' - \theta\|^3)$. Substituting and simplifying: $\mathrm{KL} = \frac{1}{2}(\theta - \theta')^\top \nabla^2 A(\theta)(\theta - \theta') + O(\|\theta - \theta'\|^3)$. Differentiating the log-partition gives $\nabla^2 A(\theta) = \mathrm{Cov}_{\pi_\theta}(T)$. Since the score is $T - \mathbb{E}_{\pi_\theta} T$, this covariance is $F(\theta)$, completing the proof. □

A.5.3 LEMMA D3 — POLYAK STEP FROM FIRST PRINCIPLES

Lemma 3. *Let L be β-smooth on the trial segment, let $\Delta = L(\theta) - L_{\mathrm{lb}} \geq 0$, and suppose $d > 0$. The largest step admitted by the lower Taylor inequality in the natural-gradient direction $-F^{-1}\nabla L$ is*

$$\eta^{\mathrm{smPol}} = \frac{2(L(\theta) - L_{\mathrm{lb}})}{d + \sqrt{d^2 + 2\beta\tilde{d}(L(\theta) - L_{\mathrm{lb}})}},$$

with $\eta^{\mathrm{smPol}} = (L - L_{\mathrm{lb}})/d$ if $\beta\tilde{d} = 0$. At a zero gradient we take zero displacement. This is maximality within the specified Taylor certificate, not a general descent or constraint-feasibility theorem. A valid global lower bound already holds for the true objective, independently of the update.

Proof. The lower Taylor bound is $L(\theta - \eta F^{-1}\nabla L) \geq L(\theta) - \eta d - (\beta\eta^2/2)\tilde{d}$. Requiring the right-hand side to be at least L_{lb} gives $(\beta\tilde{d}/2)\eta^2 + d\eta - (L - L_{\mathrm{lb}}) \leq 0$. The positive root is the displayed expression after rationalization. Since the denominator is at least $2d$, this certified step is no larger than the classical first-order Polyak step $(L - L_{\mathrm{lb}})/d$. □

A.5.4 LEMMA D4 — HOEFFDING INEQUALITY + UNIFORM UNION BOUND

Lemma 4. *Let $\bar{R}_s$ be the empirical mean of $m_s \geq 1$ conditionally independent sub-Gaussian samples with parameter σ_R^2 and conditional mean μ_s, with $\sigma_R \geq 0$ and the integer sample size chosen before those samples are drawn. Then for any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$, simultaneously for every $s \geq 1$,*

$$|\bar{R}_s - \mu_s| \leq \sigma_R \sqrt{\frac{2\log(\pi^2 s^2/(3\alpha))}{m_s}}.$$

Proof. For $\sigma_R > 0$, condition on the pre-sampling history $\mathcal{H}_s$, which determines m_s and μ_s. Hoeffding's inequality gives, almost surely, $\mathbb{P}(|\bar{R}_s - \mu_s| \geq u \mid \mathcal{H}_s) \leq 2\exp(-m_s u^2/(2\sigma_R^2))$ for $u > 0$. Substitute the displayed, history-measurable radius, then take expectations to obtain unconditional failure probability $6\alpha/(\pi^2 s^2)$. Summing over $s \geq 1$ gives total failure probability at most α. If $\sigma_R = 0$, the centered samples vanish almost surely and the zero-radius conclusion holds directly, including simultaneously over the countable sequence of evaluations. □

A.5.5 READING THE APPENDIX

The step certificate uses two-sided smoothness and a quadratic movement bound. The convergence results additionally require an exact optimal value, a Polyak-binding phase, and controlled metric drift or PL. The stochastic results transfer this inequality conditional on oracle-error budgets; they do not derive these budgets for neural policies.

A.6 PROOF OF THEOREM 1

Fix the current parameter chart, objective lower bound, smoothness bound, and nonzero-gradient direction. Assume each trial segment lies where the smoothness bound holds. We solve the two scalar certificate inequalities and then state the separate ambient-metric implication.

Step 1 - lower-Taylor no-undershoot root. Let $g = \nabla L(\theta)$, $d_M = g^\top M g$, $q_M = \|Mg\|_2^2$, and $\Delta = L(\theta) - L_{\mathrm{lb}} \geq 0$. The two-sided Taylor remainder bound implied by β-smoothness gives, for $\theta^+ = \theta - \eta M g$,

$$L(\theta^+) \geq L(\theta) - \eta d_M - \frac{\beta \eta^2}{2} q_M.$$

Thus the local certificate $L(\theta^+) \geq L_{\mathrm{lb}}$ is guaranteed whenever

$$\frac{\beta q_M}{2}\eta^2 + d_M \eta - \Delta \leq 0.$$

For $\beta q_M > 0$ this interval is $0 \leq \eta \leq \eta_M^{\mathrm{smPol}}$, where

$$\eta_M^{\mathrm{smPol}} = \frac{-d_M + \sqrt{d_M^2 + 2\beta q_M \Delta}}{\beta q_M} = \frac{2\Delta}{d_M + \sqrt{d_M^2 + 2\beta q_M \Delta}}. \tag{15}$$

If $\beta q_M = 0$, the same formula is interpreted by continuity as $\eta_M^{\mathrm{smPol}} = \Delta / d_M$. Since the denominator in the second form is at least $2d_M$, $\eta_M^{\mathrm{smPol}} \leq \Delta / d_M$: curvature can only shrink the classical first-order Polyak step.

Step 2 - trust-region root. The local trust model associated with M is

$$\frac{1}{2}(\theta^+ - \theta)^\top M^{-1}(\theta^+ - \theta) = \frac{1}{2}\eta^2 d_M \leq \delta,$$

so the trust-region interval is $0 \leq \eta \leq \eta_M^{\mathrm{TR}} := \sqrt{2\delta / d_M}$. For the Fisher choice $M = F^{-1}$ this is exactly the second-order KL model used in TRPO, because $\mathrm{KL}(\pi_\theta \| \pi_{\theta - \eta F^{-1} g}) = \frac{1}{2}\eta^2 g^\top F^{-1} g + o(\eta^2)$ under twice differentiability (or $O(\eta^3)$ with bounded third derivatives). The quadratic constraint alone does not certify an exact KL radius at a finite step.

Step 3 - intersection. Both certificates define intervals that start at zero. Their intersection is therefore

$$\left[0, \min\{\eta_M^{\mathrm{smPol}}, \eta_M^{\mathrm{TR}}\}\right],$$

and its endpoint is the supremum in Theorem 1. This proves Thm. 1.

Step 4 - coordinate covariance is not Fisher uniqueness. If $\theta = \phi(\tilde{\theta})$ has Jacobian J, then $\tilde{g} = J^\top g$ and a legitimate inverse metric transforms as $\tilde{M} = J^{-1} M J^{-\top}$. Consequently $\tilde{g}^\top \tilde{M} \tilde{g} = g^\top M g$. This verifies coordinate covariance of any Riemannian preconditioner; Fisher uniqueness requires the stronger sufficient-statistic invariance below.

Step 5 — Conditional ambient-metric implication.

Theorem 4 (Conditional ambient Chentsov pullback)**.** *Suppose one ambient metric family on action distributions satisfies Theorem 3 or an applicable continuous-space extension. Pull it back along each $\theta \mapsto \pi_\theta(\cdot \mid s)$ and average with the specified weights $d^{\pi_\theta}(s)$. If the resulting metric is h, then $h(\theta) = cF(\theta)$ for a common $c > 0$.*

Proof. The ambient theorem gives $g^\bullet = c g^{\mathrm{FR}}$ with a common scalar. Pullback gives cF_s at each state; averaging gives cF. This establishes the claim within that ambient class. State weighting itself is an additional modeling choice. □

Remark 1 (Scope of the characterization)**.** *Parameter covariance alone does not select Fisher. Even after Fisher is selected, the certificate uses a Euclidean smoothness constant and $\|F^{-1}g\|_2^2$, which depend on the chart. A finite additive step also need not commute with nonlinear reparametrization. The result characterizes the maximal certified scalar along the fixed direction with that certificate and chart fixed; it does not establish a fully invariant discrete algorithm or chance feasibility from the objective bound. For $M = c^{-1}F^{-1}$, its displacement agrees with the F^{-1} formula after replacing the movement radius by δ/c.*

Practical use of empirical returns - reward-side confidence bound.

Theorem 5 (Uniform confidence bound for the best reward seen)**.** *At iteration s, suppose CP-SPO evaluates the current policy with a positive integer m_s of conditionally independent returns, with m_s chosen before drawing those returns, $R_{s,1}, \dots, R_{s,m_s}$ whose conditional mean is $J_s = J_r(\pi_{\theta_s})$ and whose centered noise is σ_R^2-sub-Gaussian, where $\sigma_R \geq 0$. Let $\bar{J}_s = m_s^{-1}\sum_i R_{s,i}$ and, for a confidence level $\alpha \in (0,1)$, define*

$$r_s(\alpha) = \sigma_R\sqrt{\frac{2\log(\pi^2 s^2/(3\alpha))}{m_s}}, \qquad \widehat{J}_t^{+} = \max_{s\leq t}\{\bar{J}_s + r_s(\alpha)\}, \qquad J_t^{\mathrm{seen}} = \max_{s\leq t} J_s.$$

Then with probability at least $1-\alpha$, simultaneously for all $t \geq 1$,

$$J_t^{\mathrm{seen}} \leq \widehat{J}_t^{+} \leq J_t^{\mathrm{seen}} + 2\max_{s\leq t} r_s(\alpha). \tag{16}$$

Equivalently for the minimization loss $L = -J$, the quantity $\widehat{L}^{\mathrm{lb}}_{\mathrm{seen},t} := -\widehat{J}_t^{+}$ is a valid lower confidence bound on the best loss among policies evaluated so far: $\widehat{L}^{\mathrm{lb}}_{\mathrm{seen},t} \leq \min_{s\leq t} L(\theta_s)$. If all evaluated policies are feasible and $J^\star$ is the globally optimal feasible reward, then the additional gap to the global optimum is explicit:

$$-2\max_{s\leq t} r_s(\alpha) \leq J^\star - \widehat{J}_t^{+} \leq \left(J^\star - J_t^{\mathrm{seen}}\right) + 2\max_{s\leq t} r_s(\alpha)$$

(from (16)*, the left inequality additionally using $J_t^{\mathrm{seen}} \leq J^\star$). The term $J^\star - J_t^{\mathrm{seen}}$ is an optimization gap, not a concentration error; it must be controlled by a separate convergence theorem or by an analytic certificate.*

Proof. For $\sigma_R > 0$, condition on the history $\mathcal{H}_s$ before drawing the evaluation returns. Sub-Gaussian concentration gives, almost surely,

$$\mathbb{P}\{|\bar{J}_s - J_s| > r_s(\alpha) \mid \mathcal{H}_s\} \leq 2\exp\left(-\frac{m_s r_s(\alpha)^2}{2\sigma_R^2}\right) = \frac{6\alpha}{\pi^2 s^2}.$$

Taking expectations gives the same unconditional failure budget. When $\sigma_R = 0$, each centered return is zero almost surely, so the same simultaneous conclusion follows directly without dividing by σ_R. Summing over $s \geq 1$ gives failure probability at most $\sum_s 6\alpha/(\pi^2 s^2) = \alpha$. On the complementary event, $J_s \in [\bar{J}_s - r_s, \bar{J}_s + r_s]$ for every s. Taking maxima yields the lower inequality $J_t^{\mathrm{seen}} \leq \widehat{J}_t^{+}$. For the upper inequality, let s_t maximize $\bar{J}_s + r_s$ over $s \leq t$. Then

$$\widehat{J}_t^{+} = \bar{J}_{s_t} + r_{s_t} \leq J_{s_t} + 2r_{s_t} \leq J_t^{\mathrm{seen}} + 2\max_{s\leq t} r_s.$$

The loss-side statement follows by multiplying rewards by -1. The final display is the decomposition $J^\star - \widehat{J}_t^{+} = (J^\star - J_t^{\mathrm{seen}}) + (J_t^{\mathrm{seen}} - \widehat{J}_t^{+})$ and the bound just proved. □

Remark 2 (What Thm. 5 does not certify)**.** *The running maximum of empirical rewards estimates the best policy that has been* seen*; it is not by itself a certificate of the global optimum. Therefore Thm. 1 should be read with an analytic lower bound L_{lb}. If a separate certified bound $J^\star - J_t^{\mathrm{seen}} \leq B_t$ is available, then $-\widehat{J}_t^{+} - B_t$ is a valid loss lower bound on the confidence event. The unadjusted best-seen bound need not be globally valid. This distinction is essential: confidence intervals handle sampling noise, not unvisited policies.*

This gives a certified best-seen reward-target construction under the theorem's sampling assumptions. It does not certify the EWMA running-best-plus-current-margin proxy used in Algorithm 2; the maximum of individual upper confidence bounds is a different construction. Global lower-bound claims still require additional optimization assumptions.

A.7 AN EXACT RUNNING-BEST SPECIALIZATION WITHOUT AN OPTIMAL TARGET

Proposition 1 (Tail-window stationarity with a shrinking reward margin)**.** *Let L be bounded below by $L_{\inf} > -\infty$ and β-smooth on all trial segments. Use exact population rewards $J_t = -L(\theta_t)$, exact gradients $g_t = \nabla L(\theta_t)$, and matrices $\mu_F I \preceq F_t \preceq M_F I$ with $0 < \mu_F \le M_F < \infty$. Let $\delta > 0$ and $\alpha_{\mathrm{margin}} > 0$ be fixed, and set*

$$m_t = \alpha_{\mathrm{margin}} \sqrt{\frac{\log(t+e)}{t+1}}, \qquad J_t^+ = \max_{0 \le s \le t} J_s + m_t, \qquad \widehat{\Delta}_t = J_t^+ - J_t.$$

For $d_t = g_t^\top F_t^{-1} g_t > 0$, initialize the candidate scalar at

$$\eta_t^0 = \min\left\{ \frac{\widehat{\Delta}_t}{d_t}, \sqrt{\frac{2\delta}{d_t}} \right\}.$$

Fix $b \in (0,1)$ and $N_{\mathrm{bt}} \ge 1$. Test the finite grid $\eta_t^0 b^j$, $j = 0, \dots, N_{\mathrm{bt}} - 1$, and accept the first trial satisfying

$$L(\theta_t - \eta F_t^{-1} g_t) \le L(\theta_t) - c_A \eta d_t, \qquad c_A \in (0,1).$$

Retain the iterate if every trial fails. At $g_t = 0$, stop and regard all subsequent iterates as the same stationary point, with $d_t = 0$. There is no additional acceptance test or subsequent controller modification. The margin decreases to zero and has divergent sum. For each integer $T \ge 1$, define

$$D_T = \max\left\{ \frac{\beta m_T}{2(1-c_A)\mu_F}, \frac{m_T^2}{2\delta} \right\}.$$

Whenever $c_A T m_{2T} > L(\theta_0) - L_{\inf}$,

$$\min_{T \le t < 2T} d_t < D_T. \tag{17}$$

This condition holds for every sufficiently large T, and $D_T = O(\sqrt{\log T / T})$. Thus $\liminf_t d_t = 0$ and $\min_{T \le t < 2T} \|g_t\|^2 < M_F D_T$, so the Euclidean gradient norm also has limit inferior zero.

Proof. Every accepted step decreases the true loss, and rejection retains it. Consequently $J_t = \max_{s \le t} J_s$ and $\widehat{\Delta}_t = m_t$ exactly. The sequence m_t decreases to zero. Suppose, for contradiction, that $d_t \ge D_T$ at every $t = T, \dots, 2T-1$. Since $m_t \le m_T$, $m_t^2 \le 2\delta d_t$, so the objective branch binds: $\eta_t^0 = m_t / d_t$. Moreover, $q_t := \|F_t^{-1} g_t\|^2 \le d_t / \mu_F$. Smoothness therefore gives

$$L(\theta_t - \eta_t^0 F_t^{-1} g_t) \le L(\theta_t) - \eta_t^0 d_t \left(1 - \frac{\beta \eta_t^0}{2\mu_F}\right) \le L(\theta_t) - c_A m_t,$$

because $\eta_t^0 \le 2(1-c_A)\mu_F / \beta$ when $\beta > 0$; the affine case $\beta = 0$ is immediate. Hence the first trial is accepted at each such iterate, regardless of the remaining trial budget. Summing gives

$$L(\theta_T) - L(\theta_{2T}) \ge c_A \sum_{t=T}^{2T-1} m_t \ge c_A T m_{2T} > L(\theta_0) - L_{\inf},$$

a contradiction. The schedule satisfies $T m_{2T} \to \infty$, yielding the tail-window statement and rate. Finally $\|g_t\|^2 \le M_F d_t$ gives the Euclidean-gradient conclusion. □

Remark 3 (Scope of this specialization)**.** *The target need not estimate the globally optimal reward: monotonicity makes its gap equal the prescribed margin. The conclusion is a tail-window stationarity bound, not global optimality, convergence of all iterates, or a bound for constrained training. The argument requires exact unsmoothed returns and a true-loss Armijo test. It neither specifies an undocumented EWMA update nor verifies these assumptions for noisy gradients, empirical surrogate acceptance, or post-cap cost control in the reported RL implementation.*

Remark 4 (A fixed target error can stop or cycle)**.** *Instability from underestimating the target is known for deterministic Polyak steps (Orabona & D'Orazio, 2025, Sec. 6). For this first-order trust-capped candidate, a simple exact example quantifies the obstruction. Take $L(x) = x^2/2$, $F = 1$,*

and $\widehat{\Delta} = (L(x) - \widetilde{L})_+$. *If* $\widetilde{L} = B > 0$, *every point with* $0 < L(x) \le B$ *has zero candidate step despite a nonzero gradient. If* $\widetilde{L} = -B$, $B > 0$, *then for any* $\delta > 0$ *the points*

$$x = \pm\sqrt{\min\{2B/3, \delta/2\}}$$

form a two-cycle of the unmodified candidate: at either point, $\min\{1/2 + B/x^2, \sqrt{2\delta}/|x|\} = 2$, *so* $x^+ = -x$ *and* $L(x) = \min\{B/3, \delta/4\} > 0$ *remains constant. A strict loss-decrease acceptance test rejects this cycle; it does not thereby calibrate the target. These fixed-target examples do not assert persistent bias in the reported time-varying empirical target.*

A.8 Theorem 6 — Variable-Fisher convergence calculation

The rate is conditional on an exact optimal value, a Polyak-binding phase, bounded curvature ratio, and controlled metric drift. It analyzes the objective step without a safety cap or fallback. The stochastic statements transfer the inequality when independently justified error budgets are available.

Theorem 6 (Variable-Fisher Polyak-phase rate)**.** *Let* $L : \mathbb{R}^n \to \mathbb{R}$ *be convex and* β*-smooth with attained global minimum* $L^\star$. *Run the exact-gradient CP-SPO update*

$$\theta_{t+1} = \theta_t - \eta_t F_t^{-1} \nabla L(\theta_t), \qquad F_t := F(\theta_t),$$

with the exact value $L_{\mathrm{lb}} = L^\star$ *and no extra clipping or fallback. Stop at a minimizer; formulas apply while* $\Delta_t > 0$. *Assume* $F_t \succeq \mu_F I$ *on the iterate region. Let* $\theta^\star$ *be any minimizer, define*

$$\Phi_t = \frac{1}{2}\|\theta_t - \theta^\star\|_{F_t}^2,$$

and write $\Delta_t = L(\theta_t) - L^\star$, $d_t = \nabla L_t^\top F_t^{-1} \nabla L_t$, $q_t = \|F_t^{-1}\nabla L_t\|^2$, *and* $\chi_t = \beta q_t \Delta_t / d_t^2$. *For a nonempty, nonstationary phase take* $\beta > 0$ *and* $T > t_0$; *the globally affine case with an attained minimum is constant and already terminated. Suppose that on this phase* $t = t_0, \dots, T-1$:

1. *the smooth Polyak certificate binds and* $\chi_t \le \bar{\chi}$;
2. *the Fisher metric has one-step drift* $F_{t+1} \preceq (1 + \nu_t) F_t$ *for some* $\nu_t \ge 0$.

Let

$$\Gamma_T := \prod_{s=t_0}^{T-1} (1 + \nu_s), \qquad c_{\bar{\chi}} = \frac{2}{1 + \sqrt{1 + 2\bar{\chi}}}, \qquad \kappa_{\bar{\chi}} = c_{\bar{\chi}} - \frac{1}{2} c_{\bar{\chi}}^2.$$

Then

$$\min_{t_0 \le t < T} \Delta_t \le \frac{2\beta\, \Gamma_T\, \Phi_{t_0}}{\mu_F\, \kappa_{\bar{\chi}}\, (T - t_0)}. \tag{18}$$

In particular, if the trust-region branch stops clipping after a finite burn-in and $\Gamma_T = O(1)$, *the certified Polyak phase gives the usual* $O(1/T)$ *convex rate even with a changing Fisher metric.*

Proof. For $t = t_0, \dots, T-1$, Theorem 1 gives $\eta_t = a_t \Delta_t / d_t$ with $a_t = 2/(1 + \sqrt{1 + 2\chi_t}) \in [c_{\bar{\chi}}, 1]$. First expand the next iterate in the current Fisher metric:

$$\Psi_{t+1} := \frac{1}{2}\|\theta_{t+1} - \theta^\star\|_{F_t}^2 = \Phi_t - \eta_t (\theta_t - \theta^\star)^\top \nabla L_t + \frac{\eta_t^2}{2} d_t.$$

Convexity gives $(\theta_t - \theta^\star)^\top \nabla L_t \ge \Delta_t$, hence

$$\Psi_{t+1} \le \Phi_t - \left(a_t - \frac{1}{2} a_t^2\right) \frac{\Delta_t^2}{d_t} \le \Phi_t - \kappa_{\bar{\chi}} \frac{\Delta_t^2}{d_t}.$$

For convex β-smooth L, $\|\nabla L_t\|^2 \le 2\beta\Delta_t$, and $F_t \succeq \mu_F I$ implies $d_t \le \|\nabla L_t\|^2 / \mu_F \le 2\beta\Delta_t / \mu_F$. Therefore $\Delta_t^2 / d_t \ge \mu_F \Delta_t / (2\beta)$ and

$$\Psi_{t+1} \le \Phi_t - \frac{\mu_F \kappa_{\bar{\chi}}}{2\beta} \Delta_t \qquad (t = t_0, \dots, T-1).$$

The metric-drift assumption gives

$$\Phi_{t+1} = \frac{1}{2}\|\theta_{t+1} - \theta^\star\|_{F_{t+1}}^2 \le (1 + \nu_t) \Psi_{t+1}.$$

Let $P_t = \prod_{s=t_0}^{t-1}(1+\nu_s)$ with $P_{t_0} = 1$. Since $P_{t+1} = P_t(1+\nu_t)$,

$$\frac{\Phi_{t+1}}{P_{t+1}} \le \frac{\Phi_t}{P_t} - \frac{\mu_F \kappa_{\bar{\chi}}}{2\beta}\frac{\Delta_t}{P_t}.$$

Summing and using $\Phi_T \ge 0$ gives

$$\frac{\mu_F \kappa_{\bar{\chi}}}{2\beta}\sum_{t=t_0}^{T-1}\frac{\Delta_t}{P_t} \le \Phi_{t_0}.$$

Because $P_t \le \Gamma_T$, $\sum_{t=t_0}^{T-1}\Delta_t \le 2\beta\Gamma_T\Phi_{t_0}/(\mu_F\kappa_{\bar{\chi}})$, which implies the displayed minimum bound. □

Lemma 5 (Trust-region control of Fisher drift)**.** *Assume that on the iterate region $F(\theta) \succeq \mu_F I$ and F is operator-Lipschitz:*

$$\|F(\theta') - F(\theta)\|_{\mathrm{op}} \le \Lambda_F \|\theta' - \theta\|_2.$$

If the CP-SPO trust constraint at step t is enforced with radius δ_t, then

$$F_{t+1} \preceq (1+\nu_t)F_t, \qquad \nu_t = \frac{\Lambda_F}{\mu_F}\sqrt{\frac{2\delta_t}{\mu_F}}.$$

Consequently

$$\Gamma_T \le \exp\left(\frac{\Lambda_F}{\mu_F}\sum_{t=t_0}^{T-1}\sqrt{\frac{2\delta_t}{\mu_F}}\right).$$

Thus finite accumulated trust-region length gives a bounded metric-drift product.

Proof. The trust constraint gives $\|\theta_{t+1} - \theta_t\|_{F_t} \le \sqrt{2\delta_t}$, hence $\|\theta_{t+1} - \theta_t\|_2 \le \sqrt{2\delta_t/\mu_F}$. For any vector v,

$$v^\top F_{t+1} v \le v^\top F_t v + \|F_{t+1} - F_t\|_{\mathrm{op}}\|v\|^2 \le \left(1 + \frac{\Lambda_F}{\mu_F}\sqrt{\frac{2\delta_t}{\mu_F}}\right) v^\top F_t v,$$

where the last step uses $v^\top F_t v \ge \mu_F \|v\|^2$. Multiplying the one-step bounds and using $1 + x \le e^x$ proves the product bound. □

Corollary 1 (Fixed-metric specialization)**.** *If $F_t \equiv F_0$ on the Polyak-binding phase, then $\Gamma_T = 1$ and Theorem 6 reduces to the standard fixed-metric $O(1/T)$ Polyak-phase calculation.*

Corollary 2 (Oracle-error reduction for stochastic implementations)**.** *Under the assumptions of Theorem 6, suppose an implemented stochastic update produces iterates for which, on an event $\mathcal{E}$ of probability at least $1-\alpha$, the one-step current metric Lyapunov inequality is perturbed by non-negative errors ε_t:*

$$\Psi_{t+1} \le \Phi_t - \frac{\mu_F \kappa_{\bar{\chi}}}{2\beta}\Delta_t + \varepsilon_t, \qquad t = t_0, \dots, T-1,$$

where $\Psi_{t+1} = \frac{1}{2}\|\theta_{t+1} - \theta^\star\|_{F_t}^2$, and the same metric-drift bound $F_{t+1} \preceq (1+\nu_t)F_t$ holds. Then, on $\mathcal{E}$,

$$\min_{t_0 \le t < T}\Delta_t \le \frac{2\beta\,\Gamma_T}{\mu_F\,\kappa_{\bar{\chi}}\,(T - t_0)}\left(\Phi_{t_0} + \sum_{t=t_0}^{T-1}\varepsilon_t\right).$$

Thus any concentration argument for empirical gradients, Fisher solves, or lower-bound estimates only has to provide a high-probability budget for $\sum_t \varepsilon_t$; the variable-Fisher telescoping is unchanged. Theorem 7 gives the explicit budget used by the implementation.

Proof. Repeat the normalized telescoping proof of Theorem 6. The additional term contributes $\sum_t \varepsilon_t/P_t \le \sum_t \varepsilon_t$ before multiplying by the final factor Γ_T. □

Theorem 7 (Implementation-error budget)**.** *Assume the pointwise hypotheses and Polyak-binding condition of Theorem 6 at every implemented iterate. Let $s_t^\star = -\eta_t F_t^{-1}\nabla L_t$ be its ideal exact-gradient CP-SPO displacement, and let the implemented displacement be $\widehat{s}_t = -\widehat{\eta}_t\widehat{n}_t$, where $n_t = F_t^{-1}\nabla L_t$ is the exact natural-gradient direction. Suppose that on an event $\mathcal{E}$,*

$$\|\widehat{n}_t - n_t\|_{F_t} \le \zeta_t, \qquad |\widehat{\eta}_t - \eta_t| \le b_t, \qquad \|\theta_t - \theta^\star\|_{F_t} \le R_t.$$

Then the implemented displacement satisfies

$$\|\widehat{s}_t - s_t^\star\|_{F_t} \le r_t := \eta_t\zeta_t + b_t(\sqrt{d_t} + \zeta_t),$$

and the one-step current-metric Lyapunov inequality of Cor. 2 holds with

$$\varepsilon_t = (R_t + \sqrt{2\delta_t})\, r_t + \frac{1}{2}r_t^2. \tag{19}$$

Thus direction-estimation, Fisher-solve, and step-length errors enter the global rate only through the explicit scalar budget $\sum_t \varepsilon_t$.

Proof. The displacement error decomposes as

$$\widehat{s}_t - s_t^\star = -(\widehat{\eta}_t - \eta_t)n_t - \widehat{\eta}_t(\widehat{n}_t - n_t).$$

Equivalently, $\widehat{\eta}_t \le \eta_t + b_t$ and $\|n_t\|_{F_t} = \sqrt{d_t}$ give

$$\|\widehat{s}_t - s_t^\star\|_{F_t} \le b_t\sqrt{d_t} + (\eta_t + b_t)\zeta_t = r_t.$$

Let $e_t = \widehat{s}_t - s_t^\star$ and $\Psi_{t+1}^\star = \frac{1}{2}\|\theta_t + s_t^\star - \theta^\star\|_{F_t}^2$. The exact proof of Theorem 6 gives

$$\Psi_{t+1}^\star \le \Phi_t - \frac{\mu_F\kappa_{\bar{\chi}}}{2\beta}\Delta_t.$$

For the implemented iterate,

$$\widehat{\Psi}_{t+1} = \frac{1}{2}\|\theta_t + s_t^\star + e_t - \theta^\star\|_{F_t}^2 \le \Psi_{t+1}^\star + \|\theta_t + s_t^\star - \theta^\star\|_{F_t}\|e_t\|_{F_t} + \frac{1}{2}\|e_t\|_{F_t}^2.$$

The exact CP-SPO step obeys the trust bound $\|s_t^\star\|_{F_t} \le \sqrt{2\delta_t}$, so $\|\theta_t + s_t^\star - \theta^\star\|_{F_t} \le R_t + \sqrt{2\delta_t}$. Substituting $\|e_t\|_{F_t} \le r_t$ gives (19). □

Corollary 3 (End-to-end high-probability variable-Fisher rate)**.** *Assume the hypotheses of Theorem 6 and the metric drift bound $F_{t+1} \preceq (1+\nu_t)F_t$ along the implemented trajectory on a phase $t = t_0, \dots, T-1$ where the ideal exact-gradient update is Polyak-binding. The drift lemma applies only if its movement bound holds for the actual implemented step. Suppose that with probability at least $1-\alpha$, the implementation budgets in Theorem 7 hold and*

$$\frac{1}{T-t_0}\sum_{t=t_0}^{T-1}\left((R_t + \sqrt{2\delta_t})\, r_t + \frac{1}{2}r_t^2\right) \le \mathcal{E}_T.$$

Then, on the same event,

$$\min_{t_0 \le t < T}\Delta_t \le \frac{2\beta\,\Gamma_T\,\Phi_{t_0}}{\mu_F\kappa_{\bar{\chi}}(T-t_0)} + \frac{2\beta\,\Gamma_T}{\mu_F\kappa_{\bar{\chi}}}\mathcal{E}_T. \tag{20}$$

In particular, if $R_t \le R$, $\delta_t \le \delta_{\max}$, $\Gamma_T \le \Gamma$, and the natural-direction and step-length budgets satisfy $\zeta_t, b_t = O(\sqrt{n\log(T/\alpha)/M})$ uniformly with bounded d_t and η_t, and $n\log(T/\alpha)/M \le 1$, then

Writing $e_M = \sqrt{n\log(T/\alpha)/M}$, the specialization is

$$\min_{t_0 \le t < T}\Delta_t = O\left(\frac{\beta\Gamma}{\mu_F\kappa_{\bar{\chi}}}\left[\frac{\Phi_{t_0}}{T-t_0} + (R + \sqrt{\delta_{\max}})e_M + e_M^2\right]\right),$$

with constants depending only on bounded-gradient and step-length envelope constants. The conditioning factor is explicit in both the optimization and implementation-error terms.

Proof. Combine Theorem 7 with Cor. 2 and divide the summed error budget by the phase length. The displayed specialization follows by substituting the uniform $O(\sqrt{n\log(T/\alpha)/M})$ budgets into r_t and using the boundedness assumptions; the quadratic error term is retained. □

Lemma 6 (Dimension dependence of MC gradient noise)**.** *Conditionally on the iterate, let $G_{t,1},\dots,G_{t,M}$ be independent unbiased gradient estimates whose centered coordinates have variance at most σ_g^2. For their average $\widehat{g}_t$, $\mathbb{E}[\|\widehat{g}_t - \nabla L_t\|^2 \mid \theta_t] \le n\sigma_g^2/M$.*

Proof. Independence gives variance at most σ_g^2/M for each sample-mean coordinate; summing gives the result. □

This assumption is on the full gradient estimator, not the score alone. Return concentration does not imply it. High-probability direction and step-length budgets additionally require tail bounds, empirical-Fisher stability, solve accuracy, and control near a zero gradient.

Remark 5 (Strong-convexity and radius scope)**.** *Geometric rates need additional curvature and metric control, as in the PL result below. A decaying radius can increase the number of clipped iterations; no schedule-dependent rate or minimax optimality is claimed here.*

A.9 Insensitivity once the objective cap binds

Corollary 4 (Pointwise insensitivity to oversized radii)**.** *At a fixed nonstationary iterate with fixed objective information, let $\delta_{\mathrm{crit}}(\theta) = \frac{1}{2}d(\theta)(\eta^{\mathrm{smPol}}(\theta))^2$. Every $\delta \ge \delta_{\mathrm{crit}}(\theta)$ produces the same CP-SPO objective step. Runs from the same initialization with identical objective and gradient information have identical iterates if the Polyak cap binds in all runs at every step.*

Proof. The condition is equivalent to $\eta^{\mathrm{TR}} \ge \eta^{\mathrm{smPol}}$, so the min-form equals the radius-independent η^{smPol}. The trajectory claim follows by induction. □

This does not bound the clipped phase or compare pre-cap trajectories to radius-selected TRPO. A universal multiplicative comparison is false: for $L(\theta) = \theta^2/2$, $F = 1$, $\theta_0 = 1$, $L_{\mathrm{lb}} = 0$, $\beta = 1$, and one iteration, TRPO at $\delta = 1/2$ reaches zero loss. CP-SPO takes $\eta = 1/(1+\sqrt{2})$ and retains positive loss. Empirical return ratios are descriptive, not theoretical approximation constants.

A.9.1 Theorem 8 — Strict-safety guarantee under interior trust radius

The objective cap is separate from feasibility. The following sufficient condition controls cost change through movement geometry and a maintained interior margin.

Assumption 3 (Locally Fisher-Lipschitz cost surrogate)**.** *There exists $L_c \ge 0$ such that $|Dc(\theta)[v]| \le L_c\,\|v\|_{F(\theta)}$ for every tangent vector v and every θ in the iterate region, where $\|v\|_{F(\theta)}^2 = v^\top F(\theta)v$ is the local Fisher metric norm controlled by the KL trust-region model.*

Theorem 8 (One-step and inductive strict-safety guarantee)**.** *Suppose Assumption 3 holds with $L_c > 0$ on the balls below, which lie inside the domain, and $K_t < \infty$. If $L_c = 0$, the surrogate is constant on the segments and no movement cap is needed for feasibility. At step t, define*

the current safety margin $m_t := -c(\theta_t) > 0$ and

$$K_t := \sup_{\theta' \in B_t}\ \sup_{v \neq 0} \sqrt{\frac{v^\top F(\theta')v}{v^\top F(\theta_t)v}}, \qquad B_t = \{\theta' : \|\theta' - \theta_t\|_{F(\theta_t)} \le \sqrt{2\delta_t}\}. \tag{21}$$

If the trust radius used at step t satisfies, for some $\gamma \in (0,1]$,

$$\delta_t \ \le\ \frac{\gamma^2 m_t^2}{2L_c^2 K_t^2}, \tag{22}$$

then the CP-SPO update satisfies

$$c(\theta_{t+1}) \le -(1-\gamma)m_t.$$

In particular, if (22) is enforced at every iteration before a horizon T with any fixed $\gamma < 1$, then $\theta_0,\dots,\theta_T$ are strictly feasible. For a constant trust radius, a sufficient finite-horizon condition is $m_t \ge \tilde{c}$ and $K_t \le K$ on the horizon together with $\delta \le \gamma^2\tilde{c}^2/(2L_c^2K^2)$.

Proof. The CP-SPO step rule produces $\theta_{t+1} = \theta_t - \eta_t F^{-1}\nabla L$ with $\eta_t \le \sqrt{2\delta_t/d(\theta_t)}$ from the trust-region branch. The Fisher norm of the displacement is $\|\theta_{t+1}-\theta_t\|^2_{F(\theta_t)} = \eta_t^2 d(\theta_t) \le 2\delta_t$, so the per-step displacement is at most $\sqrt{2\delta_t}$ in $\|\cdot\|_{F(\theta_t)}$. Let $\gamma_s = \theta_t + s(\theta_{t+1}-\theta_t)$, $s \in [0,1]$. By Assumption 3 and the definition of K_t,

$$|c(\theta_{t+1}) - c(\theta_t)| \le \int_0^1 L_c \,\|\theta_{t+1}-\theta_t\|_{F(\gamma_s)}\, ds \le L_c K_t \sqrt{2\delta_t}.$$

Condition (22) makes this change at most γm_t, hence

$$c(\theta_{t+1}) \le c(\theta_t) + \gamma m_t = -(1-\gamma)m_t.$$

If $\gamma < 1$, this is strict feasibility, and induction gives the finite-horizon statement whenever the displayed condition is enforced at each step.

□

Remark 6 (Practical use)**.** *Theorem 8 gives a sufficient δ regime for strict-safety throughout training; it requires either adaptive per-step radii tied to the current margin, or a separate argument that a uniform margin $\tilde{c}$ is maintained on the horizon. Two practical implications: (i) for cost-sensitive deployment one can prove strict-safety by initializing strictly feasible and shrinking δ_t as the margin shrinks; (ii) the smooth Polyak branch η^{smPol} can reduce the actual displacement. The sharper test $L_c K_t \|\theta_{t+1}-\theta_t\|_{F_t} < m_t$ can then hold even when the radius-based test does not; a fixed radius itself does not shrink near the objective optimum. The condition is sufficient, not necessary: in practice, CP-SPO often stays feasible at much larger δ than* (22) *predicts, especially once the Polyak branch binds.*

A.9.2 Theorem 9 — PL extension of the CP-SPO rate

The Fisher-norm PL inequality $d(\theta) \ge 2\mu_{\mathrm{PL}}(L(\theta) - L^\star)$ is an additional optimization assumption. Positive definiteness of Fisher does not establish it. Euclidean strong convexity with modulus m and an upper bound $F \preceq M_F I$ imply Fisher-norm PL with $\mu_{\mathrm{PL}} = m/M_F$. No PL property is established here for the Safety-Gymnasium policies.

Theorem 9 (CP-SPO phase bounds and global entry under PL)**.** *Suppose $L : \mathbb{R}^n \to \mathbb{R}$ is β-smooth with attained global minimum $L^\star$, and use $L_{\mathrm{lb}} = L^\star$. Analyze the exact-gradient objective update without a safety cap or fallback, stopping at $\Delta_t = 0$. Assume $F(\theta) \succeq \mu_F I$ on the iterate region and the Fisher-norm PL condition*

$d_t \ge 2\mu_{\mathrm{PL}}\Delta_t$, *where $d_t = \nabla L_t^\top F_t^{-1}\nabla L_t$ and $\Delta_t = L(\theta_t) - L^\star$. Let $q_t = \|F_t^{-1}\nabla L_t\|^2$ and assume $\chi_t = \beta q_t \Delta_t / d_t^2 \le \bar{\chi} < 4$ at every nonstationary iterate, with $\bar{\chi} > 0$. Define*

$$c_{\bar{\chi}} = \frac{2}{1+\sqrt{1+2\bar{\chi}}}, \qquad \rho_{\bar{\chi}} = 2(1-c_{\bar{\chi}}) < 1, \qquad L_F = \beta/\mu_F.$$

Then CP-SPO with constant trust radius $\delta > 0$ has the following certified phase bounds and convergence guarantee.

1. *If $\Delta_t \le 4\mu_{\mathrm{PL}}\delta$, the smooth Polyak branch is guaranteed to bind.*
2. *If $\Delta_t \ge 4L_F\delta/c_{\bar{\chi}}^2$, the trust-region branch is guaranteed to bind.*
3. *In the Polyak-binding region, $\Delta_{t+1} \le \rho_{\bar{\chi}}\Delta_t$.*
4. *In the trust-region-binding region, $\Delta_{t+1} \le \Delta_t - 2\sqrt{\delta\mu_{\mathrm{PL}}\Delta_t} + \beta\delta/\mu_F$.*
5. *If $\Delta_0 > 0$, define*

$$h_{\bar{\chi}} = \frac{3-\sqrt{1+2\bar{\chi}}}{2} > 0, \qquad \underline{a}_0 = \min\left\{c_{\bar{\chi}}, 2\sqrt{\frac{\mu_{\mathrm{PL}}\delta}{\Delta_0}}\right\},$$
$$\rho_{\mathrm{all}} = 1 - h_{\bar{\chi}}\underline{a}_0 \in (0,1).$$

Across both branches and their transition band, $\Delta_t \le \rho_{\mathrm{all}}^t \Delta_0$. If $\Delta_0 = 0$, the method stops at the optimum.

Thus every nonterminated trajectory satisfying the assumptions reaches the sufficient Polyak-binding region in finite time. The result does not specify the actual branch-switching time and does not apply to an uncertified target, a changed controller direction, or extra clipping. The curvature-ratio condition follows, for example, from $\chi_t \le \beta/(2\mu_F\mu_{\mathrm{PL}}) < 4$, using $q_t \le d_t/\mu_F$ and PL; it is an analysis assumption, not an additional input to the step rule.

Remark 7 (What is and is not new in Theorem 9)**.** *Linear objective convergence of fixed-step Euclidean gradient descent under smoothness and PL is classical (Karimi et al., 2016). The contraction for our Fisher-preconditioned step is derived below. The useful addition is the explicit branch* band*: PL alone gives a lower bound on d_t, so it certifies that Polyak binds below $4\mu_{\mathrm{PL}}\delta$, while smoothness and $F \succeq \mu_F I$ give a sufficient trust-binding region above $4L_F\delta/c_{\bar\chi}^2$. The root identity also supplies a common descent bound through that band when $\bar\chi < 4$, rather than assuming that the trajectory reaches the local region. This is a specialized exact-oracle calculation, not a new general PL principle or an exact switching law.*

Proof. Branch conditions. The first-order Polyak cap is Δ_t/d_t, and Theorem 1 gives $\eta_t^{\mathrm{smPol}} \le \Delta_t/d_t$. Hence $\Delta_t^2 \le 2\delta d_t$ is sufficient for $\eta_t^{\mathrm{smPol}} \le \eta_t^{\mathrm{TR}}$, i.e. for Polyak binding. By PL, $d_t \ge 2\mu_{\mathrm{PL}}\Delta_t$, so $\Delta_t \le 4\mu_{\mathrm{PL}}\delta$ implies $\Delta_t^2 \le 2\delta d_t$.

For the converse sufficient region, the curvature bound gives $\eta_t^{\mathrm{smPol}} \ge c_{\bar\chi}\Delta_t/d_t$. Smoothness and $F_t \succeq \mu_F I$ imply $d_t \le \|\nabla L_t\|^2/\mu_F \le 2L_F\Delta_t$. Therefore $\eta_t^{\mathrm{TR}} \le \eta_t^{\mathrm{smPol}}$ is guaranteed when $\sqrt{2\delta/d_t} \le c_{\bar\chi}\Delta_t/d_t$, i.e. $\Delta_t^2 \ge 2\delta d_t/c_{\bar\chi}^2$, which follows from $\Delta_t \ge 4L_F\delta/c_{\bar\chi}^2$.

Polyak-region contraction. In the Polyak-binding region write $\eta_t = a_t\Delta_t/d_t$ with $a_t = 2/(1+\sqrt{1+2\chi_t}) \ge c_{\bar\chi}$. The smooth descent lemma gives

$$\Delta_{t+1} \le \Delta_t - a_t\Delta_t + \frac{1}{2}a_t^2\chi_t\Delta_t = 2(1-a_t)\Delta_t \le 2(1-c_{\bar\chi})\Delta_t = \rho_{\bar\chi}\Delta_t.$$

The equality uses $a_t + \chi_t a_t^2/2 = 1$. In particular, $\bar\chi < 4$ implies $c_{\bar\chi} > 1/2$ and contraction.

Trust-region descent. If the trust branch binds, then

$$L_{t+1} - L^\star \;\le\; L_t - L^\star - \eta_t d_t + \tfrac{\beta\eta_t^2}{2}\tilde d_t \;=\; (L_t - L^\star) - \sqrt{2\delta\, d_t} + \beta\delta\tilde d_t/d_t.$$

Here $\tilde d_t/d_t \le 1/\mu_F$ and PL gives $\sqrt{2\delta d_t} \ge 2\sqrt{\delta\mu_{\mathrm{PL}}\Delta_t}$, yielding the displayed trust-region inequality.

Uniform progress through both branches. At any nonstationary iterate set

$$b_t := \frac{\eta_t d_t}{\Delta_t} = \min\left\{a_t, \frac{\sqrt{2\delta d_t}}{\Delta_t}\right\}.$$

Since $0 < b_t \le a_t$ and $\chi_t a_t = \sqrt{1+2\chi_t} - 1$, smoothness yields

$$\Delta_{t+1} \le \big[1 - b_t(1 - \chi_t b_t/2)\big]\Delta_t \le (1 - h_{\bar\chi} b_t)\Delta_t.$$

This first establishes $\Delta_{t+1} \le \Delta_t$, so induction gives $\Delta_t \le \Delta_0$. PL now implies

$$\frac{\sqrt{2\delta d_t}}{\Delta_t} \ge 2\sqrt{\frac{\mu_{\mathrm{PL}}\delta}{\Delta_t}} \ge 2\sqrt{\frac{\mu_{\mathrm{PL}}\delta}{\Delta_0}}, \qquad b_t \ge \underline{a}_0.$$

Substitution and induction prove the all-branch rate. At a reached optimum the stopped sequence keeps gap zero. □

Corollary 5 (Finite entry and subsequent ε-accuracy under PL)**.** *Under Theorem 9, suppose $\Delta_0 > 0$. The first entry time τ into $\Delta_t \le 4\mu_{\mathrm{PL}}\delta$ obeys*

$$\tau \le \left\lceil \frac{\max\{0, \log(\Delta_0/(4\mu_{\mathrm{PL}}\delta))\}}{\log(1/\rho_{\mathrm{all}})} \right\rceil.$$

The objective branch binds thereafter unless the method has already terminated. From entry, the gap is at most $\rho_{\bar\chi}^k\Delta_\tau$ after k further iterations. Consequently $\lceil \log(\Delta_\tau/\varepsilon)/\log(1/\rho_{\bar\chi})\rceil$ further iterations suffice when $0 < \varepsilon < \Delta_\tau$. These upper bounds need not equal the observed crossing or convergence times.

Remark 8 (Scope of PL). *PL permits some nonconvex objectives (Karimi et al., 2016), but requires a justification for the particular objective, parameterization, and metric. A Fisher lower eigenvalue bound or overparameterization alone does not imply it. The synthetic experiment illustrates the mechanism; it does not establish these assumptions for neural policies.*

Remark 9 (Radius dependence in the stated PL class). *Under the assumptions of Theorem 9, far from the optimum the trust-region branch controls step size, while near the optimum the Polyak certificate can take over and contract geometrically. The local contraction factor is radius independent, but the entry bound and ρ_{all} depend on δ. This is not a comparison against the best tuned trust-region trajectory.*

Proposition 2 (Eventual Polyak-binding under PL, sufficient form). *Let $\phi_T := \frac{1}{T}\sum_{t<T} \mathbf{1}\{\eta_t = \eta_t^{\mathrm{smPol}}\}$ denote the Polyak-binding fraction over T iterations. Under the hypotheses of Theorem 9, let t_0 be the finite entry time from Corollary 5. Then the smooth Polyak branch binds for every $t \geq t_0$ and $\Delta_t \leq \rho_{\bar{\chi}}^{t-t_0}\Delta_{t_0}$. For a trajectory that does not terminate at the optimum, $\phi_T \to 1$ as $T \to \infty$.*

Proof. Theorem 9 shows that $\Delta_t \leq 4\mu_{\mathrm{PL}}\delta$ is sufficient for smooth-Polyak binding. In that region the same theorem gives $\Delta_{t+1} \leq \rho_{\bar{\chi}}\Delta_t$. Since $\rho_{\bar{\chi}} < 1$, induction gives $\Delta_t \leq \rho_{\bar{\chi}}^{t-t_0}\Delta_{t_0} \leq \Delta_{t_0} \leq 4\mu_{\mathrm{PL}}\delta$, so the sufficient binding condition remains true for all later iterates. If the hitting time t_0 is finite, then $\phi_T \geq (T - t_0)/T \to 1$. □

B Empirical support

This appendix specifies the experimental protocols and separates the practical first-order variant from the certified finite-curvature rule. E1–E4 evaluate first-order objective-gap clipping; E5 tests the separate feasibility cap on a synthetic problem; E6 evaluates a practical policy-learning implementation with a running-best reward proxy and Lagrangian cost control. The E6 safety cap is disabled. The finite-curvature theorem does not directly certify the first-order experiments, and the E6 proxy is not a certified objective lower bound.

B.1 Experimental record and reproducibility

Hardware and software. Synthetic runs take less than five seconds on a laptop CPU. The 45-run, 500k-step Safety-Gymnasium sweep ran on an Apple M3 Max with 128 GB unified memory and MPS, taking approximately 4 h 45 min with four concurrent processes, at approximately 300–350 environment steps/s per process. The software configuration is Python 3.11, PyTorch 2.11, and `safety-gymnasium` 1.0 with compatibility patches for mutable dataclass defaults and Gymnasium's removed compatibility wrapper. These resource measurements describe the stated 45-run sweep; the additional studies below have separate training budgets.

Shared policy-learning settings. The shared settings are $\delta = 0.01$ for CP-SPO, cost limit $C = 25$, controller margin $\tau = 2.5$, dual stepsize $\rho = 0.05$, return EWMA weight $\beta_{\mathrm{ewma}} = 0.9$, and return scale 100. Policies and value networks have two 64-unit tanh hidden layers. Rollouts contain 4096 environment steps; GAE uses $\gamma = 0.99$ and $\lambda = 0.97$; value optimization uses Adam with learning rate 3×10^{-4} and 20 updates per batch. Conjugate gradient uses at most 10 iterations with damping 0.1; line search uses at most 10 trials, shrinkage 0.8, and empirical KL threshold 1.5δ. The environment horizon was not overridden. The submitted trainer uses a return-margin coefficient 0.05 and the unified trainer uses 0.1. These are operational target margins, not confidence-failure probabilities; these distinct settings must be retained when comparing protocols. The theoretical smoothness bound and the EWMA weight have distinct roles. No smoothness certificate was used in the first-order E6 cap. The stated running-best-plus-margin target is

$$\widehat{J}_t^{+} = \max_{s \leq t} \widehat{J}_s + \alpha_{\mathrm{margin}} \sqrt{\frac{\log(t+e)}{t+1}}, \qquad \widetilde{L}_t = -\widehat{J}_t^{+}.$$

The return estimates use the reported EWMA coefficient and return-scale settings above. This is an empirical target specification, not the confidence bound of Theorem 5.

B.2 CERTIFIED RULE AND EMPIRICAL PROPOSAL

The certified update uses the smooth root from Theorem 1 with a certified lower bound and a valid smoothness upper bound. The empirical proposal below instead uses a running-best reward target. It describes only the reward-step component; the full update also applies the Lagrangian cost controller and line search. The Lagrangian cost-control path and the feasibility cap must not be identified with this reward proposal. In loss-minimization convention, a Lagrangian loss $L + \lambda(J_c - C)$ has gradient $\nabla L + \lambda \nabla J_c$.

Algorithm 2: First-order reward proposal used in the empirical variant

Input: estimated loss gradient $\widehat{g}$, positive-definite metric $\widehat{F}$, current reward $\widehat{J}$, reward target $\widehat{J}^+$, quadratic movement radius $\delta > 0$

$v \leftarrow \widehat{F}^{-1}\widehat{g}, \quad d \leftarrow \widehat{g}^\top v$;
if $d = 0$ **then**
  Return a zero reward proposal;
else
  $\widehat{\Delta} \leftarrow [\widehat{J}^+ - \widehat{J}]_+$;
  $\eta^{\mathrm{P}} \leftarrow \widehat{\Delta}/d, \quad \eta^{\mathrm{TR}} \leftarrow \sqrt{2\delta/d}$;
  $\eta \leftarrow \min\{\eta^{\mathrm{P}}, \eta^{\mathrm{TR}}\}$;
  Return proposal $-\eta v$ to the cost-control and line-search procedure;

This first-order cap is the algebraic $\beta = 0$ specialization of the smooth formula. On a nonlinear objective, setting $\beta = 0$ does not establish the finite-curvature lower-Taylor certificate. Likewise, a running-best target bounds the best evaluated policy only under appropriate concentration assumptions; it does not bound the global optimum without an additional optimization-gap certificate.

B.3 SYNTHETIC MECHANISM EXPERIMENTS

E1 uses a two-dimensional Gaussian policy with identity covariance and quadratic loss $L(\theta) = \frac{1}{2}\theta^\top Q\theta - q^\top\theta$, with surrogate $c(\theta) = \max(0, \theta^\top H\theta - h)^2$, $\delta = 0.05$, 15 initializations, and 60 updates. E2 uses a four-dimensional quadratic with condition number 10, noise scale 0.05, $\theta_0 = 5\mathbf{1}$, 20 seeds, and 200 updates. Both reported objective-gap variants use the analytic optimum; the oracle is not removed in E2. The following table summarizes the original mechanism experiments. Their first-order objective-gap rule is distinct from the finite-curvature sensitivity experiment reported below.

	E1 (oracle $L^\star$, δ=0.05)		E2 (oracle $L^\star$=0, δ-sweep)		
Method	Final $L-L^\star$ (60 iter, med.)	Cum. viol. (60 iter, sum)	Final $L-L^\star$ (δ=0.3, med.)	Polyak share (last 50)	Robustness ratio ($\delta\geq$0.1)
TRPO	2.4×10^{-2}	54	8.5×10^{-2}	0%	47.31
CPO	2.6×10^{-2}	**28**	9.1×10^{-2}	0%	34.82
Polyak only	1.2×10^{-7}	210	3.8×10^{-2}	100%	2.74
CP-SPO	$\mathbf{\leq 10^{-10}}$	54	$\mathbf{3.3\times10^{-2}}$	**95%**	**1.04**

Table 5: Reported E1 and E2 summary. E1 losses are terminal medians and the violation column is the reported cumulative count. E2 losses are reported at $\delta = 0.3$, with branch share over the last 50 iterations. The E2 ratios describe this experiment; the separate E3 sweep below reports its own within-instance comparisons.

The 10^{-10} terminal-gap threshold is a numerical tolerance. The E1 cumulative counts imply approximately $210/54 = 3.89$ times as many surrogate-infeasible iterates for Polyak-only as for CP-SPO on this instance; they are not trajectory-level chance-violation counts.

B.4 Reported synthetic radius sweep

δ	TRPO final L	CP-SPO final L	CP-SPO / best-TRPO
10^{-3}	33	33	4.58×10^{3}
$3\cdot10^{-3}$	1.8	1.8	$250\times$
10^{-2}	$7.2\cdot10^{-3}$	$8.4\cdot10^{-3}$	$1.17\times$
$3\cdot10^{-2}$	$1.5\cdot10^{-2}$	$1.2\cdot10^{-2}$	$1.67\times$
10^{-1}	$3.5\cdot10^{-2}$	$2.0\cdot10^{-2}$	$2.78\times$
$3\cdot10^{-1}$	$8.9\cdot10^{-2}$	$3.4\cdot10^{-2}$	$4.72\times$
1	0.30	$3.4\cdot10^{-2}$	$4.72\times$

Table 6: Reported E3 last-20-iteration loss summary, 200 updates and 20 seeds. The final column is recomputed from the displayed losses using 0.0072 as the best reported TRPO loss. These are within-instance comparisons, not a universal bound on return or radius sensitivity.

For this E3 sweep, across $\delta \in \{0.1, 0.3, 1\}$ the CP-SPO max/min loss ratio is $0.034/0.020 = 1.70$, compared with $0.30/0.035 \simeq 8.57$ for TRPO. Across $\delta \in [0.01, 1]$, the largest displayed CP-SPO/best-TRPO loss ratio is approximately 4.72. These within-instance quantities differ from E2's aggregate summary and should not be pooled. Small radii slow both methods at the fixed training horizon.

PL-phase diagnostic. The five-seed E4 measurements use the first-order rule and describe a transition in branch activity. The reported phase estimate 12 ± 1.5 differs from the heuristic prediction 10 by 1.33 reported standard deviations. The measured contraction 0.78 ± 0.03 is descriptive; applying the finite-curvature rate theorem additionally requires its specified update and assumptions.

B.5 Direct finite-curvature certificate test

A separate deterministic E1 extension evaluates the smooth-root rule with $F = I$, $L_{\mathrm{lb}} = L^\star$, and $\beta_{\mathrm{true}} = \lambda_{\max}(Q)$ known exactly. Seven smoothness multipliers and five radii are evaluated from the same 15 initial points: 525 runs and 31,500 updates. Every valid smoothness overestimate preserves the lower-Taylor certificate. The smallest computed residual is -1.55×10^{-15} against the stated numerical tolerance -10^{-10}. At $\delta \in \{0.1, 0.3, 1\}$, every start reaches $L - L^\star \le 10^{-10}$ within 60 updates for multipliers 1, 2, and 4; the respective median iteration ranges are 30–32, 37–39, and 48–50. The conservative $10\times$ setting yields terminal median gaps 3.16×10^{-9}–4.53×10^{-9}. This tests the certified rule directly, whereas the original E1–E4 experiments test its first-order specialization.

Smoothness multiplier	$1\times$	$2\times$	$4\times$	$10\times$
Terminal-gap max/min ratio	1.00	1.00	1.00	1.43
Smooth-cap binding at $\delta = 0.3$	96.4%	97.3%	98.1%	99.3%

Table 7: Finite-curvature E1 extension. Terminal-gap ratios are over $\delta \in \{0.1, 0.3, 1\}$. A deliberately invalid $0.5\times$ smoothness estimate at $\delta = 0.3$ produces a 37.0% certificate-residual violation fraction, illustrating why the bound must be a valid upper bound.

B.6 Synthetic feasibility cap and held-out tail check

E5 isolates the feasibility cap with 10 seeds and 2000 iterates per seed. The original study reports zero chance-infeasible iterates out of 20,000 with the cap, compared with 6412 out of 20,000 (32.1%) without it. The case-complete rerun again reports zero out of 20,000 and a 5.9% change in final reward. One discriminant-boundary roundoff diagnostic produces no chance-infeasible iterate. The reported counts are tests of policy-level chance feasibility; they are not counts of individual unsafe trajectories or 20,000 independent Bernoulli trials. The guarantee is conditional on a surrogate-feasible start, the sub-Gaussian assumption, and valid cost and curvature bounds.

A separate frozen-policy PointGoal1 check uses 200 held-out evaluation trajectories after calibration. Three cross the tested threshold (1.5%). Under the stated independence assumption, the exact one-

sided 95% Clopper–Pearson upper bound is 3.83%. The empirical survival curve lies below the calibrated sub-Gaussian envelope at the tested thresholds. This is a check of the tail premise at those thresholds; it neither establishes global sub-Gaussianity nor certifies the E6 training policies, which run with the feasibility cap disabled.

B.7 Safety-Gymnasium at 500k steps: three-seed sweep

The primary 500k sweep below uses three seeds for every configuration. All means summarize the final 50k environment steps. The uncertainty entries are the reported across-seed standard deviations. CP-SPO uses $\delta = 0.01$ once per task. This tests transfer of that fixed configuration, but does not measure CP-SPO's own sensitivity to radius.

	PointGoal1		CarGoal1		PointPush1	
Method	$\bar{R}$	$\bar{C}$	$\bar{R}$	$\bar{C}$	$\bar{R}$	$\bar{C}$
TRPO-Lag $\delta{=}10^{-3}$	$10.76{\pm}0.21$	71.6	$4.40{\pm}1.12$	86.2	$0.31{\pm}0.02$	64.8
TRPO-Lag $\delta{=}10^{-2}$	$20.59{\pm}0.40$	51.2	$18.76{\pm}2.13$	78.6	$0.56{\pm}0.10$	59.0
TRPO-Lag $\delta{=}5{\cdot}10^{-2}$	$\mathbf{22.53{\pm}0.39}$	50.8	$\mathbf{25.52{\pm}2.63}$	60.3	$\mathbf{0.66{\pm}0.87}$	46.6
TRPO-Lag $\delta{=}10^{-1}$	$21.00{\pm}2.35$	66.2	$25.19{\pm}1.11$	66.3	$0.63{\pm}0.24$	54.6
CP-SPO ($\delta{=}10^{-2}$)	$19.45{\pm}0.84$	53.7	$16.83{\pm}1.71$	71.2	$0.39{\pm}0.27$	54.6

Table 8: Three-seed 500k Safety-Gymnasium sweep. The selected baseline is the maximum reported TRPO-Lagrangian mean return among the four radii in this sweep. Baseline/CP-SPO ratios from the displayed means are 1.16, 1.52, and 1.69 on PG1, CG1, and PP1. Every displayed mean cost exceeds 25.

B.8 Five-seed PointPush1 summary

A separate 500k summary uses five PointPush1 seeds (the original three plus two additional seeds), while retaining the same three-seed PG1 and CG1 summaries. It is retained below as a separate reporting scope. Adding two PP1 seeds changes both the sample mean and the return-selected baseline radius, so this summary is kept separate from the three-seed sweep.

	PointGoal1		CarGoal1		PointPush1	
Method	$\bar{R}$	$\bar{C}$	$\bar{R}$	$\bar{C}$	$\bar{R}$	$\bar{C}$
TRPO-Lag $\delta{=}10^{-2}$	20.6 ± 0.4	51.2	18.8 ± 2.1	78.6	0.48 ± 0.18	57.2
TRPO-Lag $\delta{=}5{\cdot}10^{-2}$	$\mathbf{22.5 \pm 0.4}$	50.8	$\mathbf{25.5 \pm 2.6}$	60.3	0.52 ± 0.65	43.3
TRPO-Lag $\delta{=}10^{-1}$	21.0 ± 2.4	66.2	25.2 ± 1.1	66.3	$\mathbf{0.69 \pm 0.23}$	58.2
CP-SPO (fixed $\delta{=}10^{-2}$)	19.5 ± 0.8	53.7	16.8 ± 1.7	71.2	0.53 ± 0.26	64.4
$C^{\text{emp}} = R^{\star}_{\text{tuned}} / R_{\text{CP-SPO}}$	**1.16**		**1.52**		**1.30**	

Table 9: Reported 500k summary with three seeds on PG1/CG1 and five on PP1. The reported baseline/CP-SPO return ratios refer to selection within this table (displayed means are rounded). The PP1 ratio 1.30 belongs to this five-seed summary; it is not the three-seed value 1.69. Neither ratio is a theoretical guarantee. All mean costs exceed 25.

B.9 Runtime and observed controller activity

The reported median seconds per 1000 environment steps are 2.75/2.75 (PG1), 2.96/2.90 (CG1), and 2.81/2.80 (PP1), with CP-SPO listed first and TRPO-Lagrangian second. The largest difference computed from these rounded values is approximately 2.1%. These compare CP-SPO at $\delta = 0.01$ with TRPO-Lagrangian at $\delta = 0.05$ and do not include runtime uncertainty; they suggest similar overall runtime rather than establishing zero overhead from the cap alone.

Task	Phase	Polyak	TRPO	Lagr. fallback
PointGoal1	early	0.7%	7.7%	91.7%
	mid	2%	4%	93%
	late	0.7%	5.7%	93.7%
CarGoal1	early	1%	20%	78%
	mid	2.7%	10.7%	86.7%
	late	7%	4%	88%
PointPush1	early	0%	42%	57%
	mid	0%	43%	56%
	late	1%	47%	51%

Table 10: Reported branch classifications in early, middle, and late thirds of the 500k training horizon; three seeds pooled. There are approximately 122 updates per seed, or 366 pooled updates. Each entry in the PointGoal1 early/late and CarGoal1 middle rows includes an author-specified correction of 0.7 percentage points; these displayed rows total 100.1%. The displayed categories are controller-state summaries; the matched cap-only study below isolates the effect of changing the cap.

Standalone Polyak classifications range from 0% to 7%. Lagrangian fallback accounts for 51–93.7% across the displayed rows. This is evidence about the recorded controller states, not a safety certificate or proof that objective clipping causes the observed returns. The stored-candidate analysis further distinguishes which candidate was smaller before cost control from the eventual controller state. Among safeguard-active proposals, the Polyak candidate is smaller on 5.8%, 17.1%, and 7.2% of PG1, CG1, and PP1 proposals, respectively. The conditional median Polyak/TRPO candidate ratios are 0.91, 0.87, and 0.93. These reductions can be hidden by a later controller-active classification.

B.10 Matched cap-only and return-target sensitivity studies

The cap-only ablation uses three paired PointGoal1 seeds and 500k steps, holds the augmented reward–cost controller fixed, and changes only the cap. Full CP-SPO improves final-50k mean return by 8.6% and return area under the curve by 8.1%, while reducing the area of cost excess above the budget by 5.7%. The across-seed return-SD ratio (full/cap-disabled) is 1.10, so this particular ablation does not show a variance reduction. Material applied-step differences occur on 7.0% of accepted controller-active updates. The median of the three seed-level median candidate ratios is 0.83, corresponding to a 17% reduction. This matched comparison identifies a contribution of the cap under the tested controller; it does not establish that every observed cross-task difference is caused by the cap.

A separate three-seed, 500k PointGoal1 sensitivity study varies only the return EWMA weight. The baseline value 0.9 is shared across tasks. Table 11 reports changes relative to that value. Both alternatives reduce return and increase cost-excess area in this study. The largest aggregate change is 13.0%; the result quantifies sensitivity of this operational parameter rather than making the implementation parameter-free.

EWMA weight	Return change	Cost-excess area change	Return-SD change
0.5	-10.4%	$+11.8\%$	$+12.3\%$
0.99	-4.1%	$+7.2\%$	$+13.0\%$

Table 11: EWMA sensitivity relative to the fixed value 0.9 on PointGoal1, three seeds and 500k steps. These weights affect the empirical reward target; they are distinct from the smoothness bound in the certificate.

B.11 COMPACT ADDITIONAL BASELINES

The additional CPO-style and PPO-Lagrangian comparators were reported as compact implementations in the same training pipeline. The CPO-style comparator uses projected/recovery updates and is not the full CPO reference dual solver. These results therefore assess the reported implementations, not the best attainable performance of the named algorithms. The official-implementation comparison below reports a separate reference-CPO protocol. Published results with different horizons are contextual rather than matched comparisons.

	PointGoal1		CarGoal1		PointPush1	
Method	$\bar{R}$	$\bar{C}$	$\bar{R}$	$\bar{C}$	$\bar{R}$	$\bar{C}$
CPO-style	9.37 ± 0.20	55.8	5.66 ± 1.19	58.4	0.27 ± 0.24	39.8
PPO-Lagrangian	1.15 ± 1.99	27.2	-0.02 ± 0.26	7.7	0.43 ± 0.12	27.0

Table 12: Reported compact baselines at 500k, final-50k means over three seeds. Lower mean cost can accompany lower return. Neither mean return nor mean cost alone establishes chance feasibility.

B.12 OFFICIAL CPO REFERENCE ON POINTGOAL1

The reference comparison uses official OmniSafe 0.5.0 CPO with native estimation and line search. Three PointGoal1 seeds share the exposed settings: 4096-step rollouts, two 64-unit tanh layers, cost limit 25, target KL 0.01, conjugate-gradient and damping settings, and critic updates. Aggregation differs: OmniSafe uses rolling-100-episode summaries spanning approximately 100k steps, whereas CP-SPO uses nominal final-50k summaries. Table 13 preserves this distinction. The CP-SPO return is higher at both reported markers, while reference CPO has lower cost; neither reaches the cost budget. The different averaging windows preclude a fully matched endpoint comparison. The CPO entries use sample standard deviations; CP-SPO retains the original reported summaries.

Interaction marker	Method	Return	Mean episodic cost
Approximately 500k	Official CPO	3.07 ± 1.67	39.8 ± 9.2
500k	CP-SPO	19.45 ± 0.84	53.7
Approximately 1M	Official CPO	9.32 ± 3.01	33.2 ± 7.9
1M	CP-SPO	23.86 ± 0.49	53.2

Table 13: Three-seed PointGoal1 reference-CPO comparison. The different aggregation windows are specified in the text; these observations show distinct return–cost operating points, not cost-constrained dominance.

B.13 ONE-MILLION-STEP COMPARISON

The 1M extension compares CP-SPO at $\delta = 0.01$ with TRPO-Lagrangian at the selected radius 0.05, using three seeds and final-50k-step summaries. A full four-radius sweep at 1M is not reported, so the comparator is radius-selected rather than demonstrably optimal over radii at that horizon. On PP1 the five-seed 500k summary selected 0.1, whereas the three-seed summary selected 0.05; the 1M comparison follows the latter protocol.

Method	PointGoal1 $\bar{R}$	PointGoal1 $\bar{C}$	CarGoal1 $\bar{R}$	CarGoal1 $\bar{C}$	PointPush1 $\bar{R}$	PointPush1 $\bar{C}$
TRPO-Lag δ=5·10^{-2}	**24.65±0.04**	53.9	**28.19±1.19**	70.5	**1.38±0.51**	62.3
CP-SPO (δ=10^{-2}, fixed radius)	23.86±0.49	53.2	26.13±1.61	69.0	0.56±0.05	**41.7**
C^{emp} at 1M	**1.03**		**1.08**		**2.46**	

Table 14: Reported 1M Safety-Gymnasium results, three seeds and final-50k-step means. CP-SPO retains 96.8%, 92.7%, and 40.6% of the selected TRPO-Lagrangian mean return on PG1, CG1, and PP1, respectively. The PP1 cost is lower (41.7 versus 62.3), but all mean costs exceed 25; these runs demonstrate no cost-budget compliance or chance-feasibility guarantee.

B.14 Separate seven-method PointPush1 protocol

The seven-method PP1 comparison uses a different recorded CP-SPO variant (v3), a common nominal trust-radius setting of 0.01 where applicable, and final-50-epoch means. Fifty epochs at 4096 steps/epoch span approximately 204,800 steps, so this aggregation is not the final-50k window above. CP-SPO's reported cost also differs (58.2 versus 41.7). The two CP-SPO rows therefore represent different protocol summaries. PPO-Lagrangian and SPS-PG do not become radius-based methods merely because the common configuration contains a nominal radius field.

Method	$\bar{R}$±std	$\bar{C}$±std	per-seed returns
CP-SPO v3	**0.58±0.02**	58.2±12.0	[0.58, 0.61, 0.56]
trpo_pid (highest mean return)	**0.66±0.29**	35.9±9.5	[0.40, 0.50, 1.06]
c_trpo	0.58±0.25	31.2±2.3	[0.44, 0.92, 0.37]
cpo	0.30±0.11	42.6±2.2	[0.38, 0.38, 0.15]
trpo_lag δ=10^{-2}	0.10±0.42	**21.5±4.8**	[0.29, 0.51, −0.48]
ppo_lag	0.20±0.13	22.6±12.3	[0.11, 0.38, 0.11]
sps_pg	−0.17±0.05	21.7±6.5	[−0.21, −0.10, −0.18]
$C^{\text{emp}}_{\text{default}} = R^{\star}_{\text{best at } \delta=0.01} / R_{\text{CP-SPO}}$		**0.66/0.58 ≈ 1.14×**	

Table 15: Separate PP1 suite: 1M steps, three seeds, final-50-epoch aggregation. The reported return standard deviations are consistent, to rounding, with population standard deviations of the displayed seed means. CP-SPO has the smallest observed return spread in this small sample; this does not establish a general variance reduction or attribute it to the cap. The best-suite/CP-SPO mean-return ratio is approximately 1.14 within this protocol only.

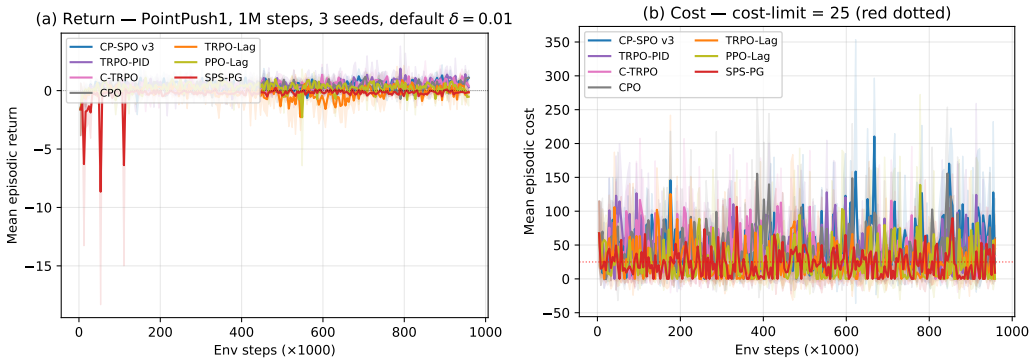

Figure 1: PP1 training curves for the separate seven-method protocol at 1M steps, three seeds. The original panels are enlarged and stacked: return above, cost below. The plotted bands are reported as one across-seed standard deviation. The dotted horizontal line marks mean episodic cost 25. The broad transient variation and distinct final aggregation window should be kept in view when interpreting the final-mean table.

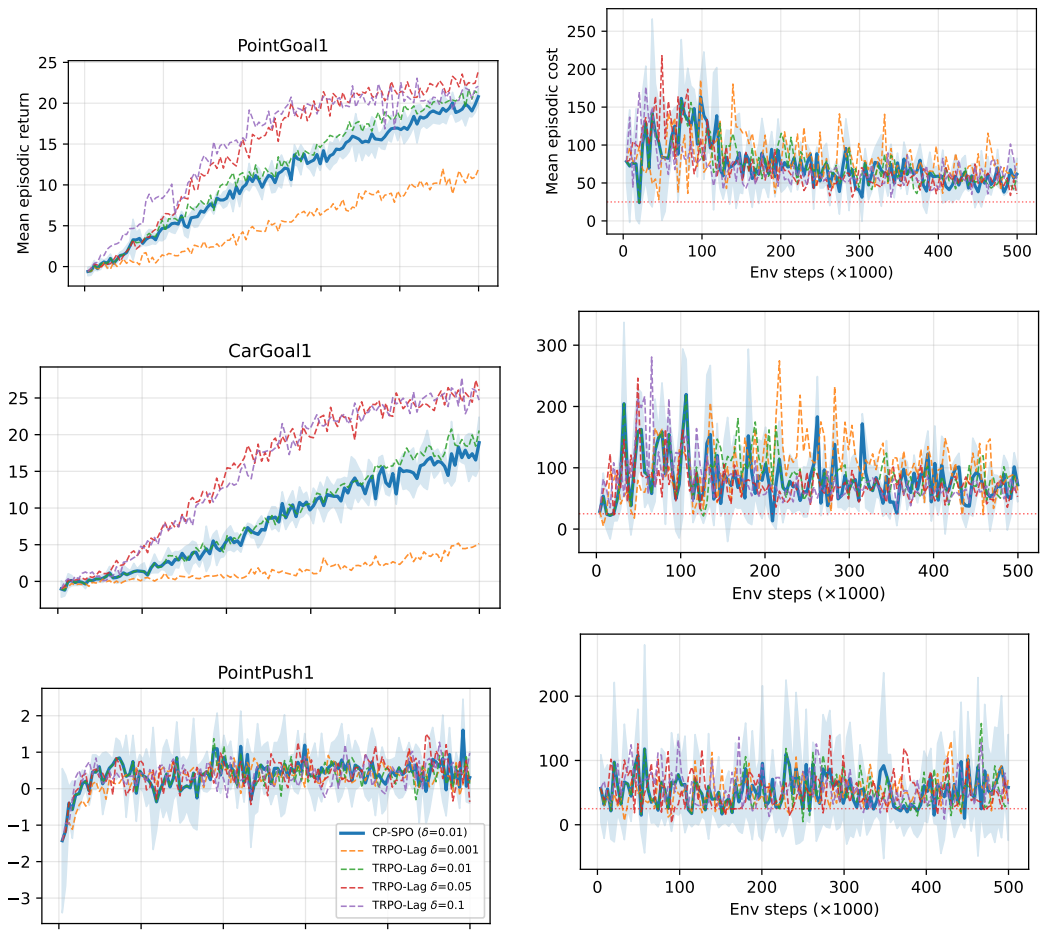

Figure 2: Three-seed, 500k Safety-Gymnasium curves. The solid black curve is CP-SPO at $\delta = 0.01$; the four TRPO-Lagrangian radii use distinct dash patterns. Panels are enlarged and rearranged into task rows (PG1, CG1, PP1). Left: mean episodic return. Right: mean episodic cost, with budget 25 shown as a horizontal reference line. The curves illustrate the reported fixed-configuration comparison; they do not measure a CP-SPO radius sweep or establish safety.

Scope of empirical conclusions. The synthetic tests distinguish objective-gap clipping from the separate requirements of the safety cap. The finite-curvature study tests valid and invalid smoothness inputs, and the matched PointGoal1 ablation identifies a contribution of the cap within a fixed controller. The real-policy results use a running-best reward proxy and a Lagrangian fallback. They show a return–cost tradeoff, not chance-feasible learning: all mean costs in the primary 1M comparison exceed the budget. Three tasks and three to five seeds limit claims about transfer and variability. A single shared radius is a tested configuration, not proof that radius or target-estimation sensitivity has disappeared.