IMPROVED SAMPLE COMPLEXITY ANALYSIS FOR FIRST ORDER BILEVEL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel reinforcement learning optimizes an upper-level objective through the policy response to a lower-level reinforcement learning problem. Such problems arise when learning a reward model while adapting a policy to that reward. However, a gap remains between sample-complexity guarantees and practical actor–critic training. Tabular analyses restrict applicability to finite spaces, while continuous-state analyses often impose a uniform Polyak–Łojasiewicz (PL) condition. We show that this condition need not hold even for smooth positive policies with exactly representable values. We develop a single-loop, fully first-order method for continuous states and finite actions with specified multilayer neural actors and a persistently trained nonlinear critic built from fixed features. Under explicit initialization, regularity, feature-excitation, and value-realizability conditions, fixed reference-KL regularization yields local curvature, and we prove containment of the coupled iterates. Shared trajectories control sampling noise, while paired critic analysis controls normalized critic-error differences as the response perturbation shrinks. A joint descent and tracking argument establishes an improved $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity for $O(\varepsilon)$ expected squared gradient norm of the upper objective induced by the selected regularized policy response, including critic training.

1 INTRODUCTION

Bilevel reinforcement learning (BRL) couples reward learning with policy adaptation: changing a reward changes both the responding policy and the trajectories used to evaluate the upper objective. This dependence underlies reward design and reinforcement learning from human feedback. Establishing convergence requires justifying the policy-optimization geometry and controlling value estimation while rewards, policies, and sampling distributions evolve together.

Tabular analyses exploit finite policy and value representations (Shen et al., 2025; Yang et al., 2025). In particular, Yang et al. (2025) use soft Bellman structure to obtain first-order upper-gradient calculations for a global regularized response. Zeng et al. (2026) establish a single-loop first-order tabular actor–critic guarantee under regularization-dependent PL and additional Hessian conditions. Extending these guarantees beyond finite state–action representations requires controlling actor and critic approximation and learning errors.

Beyond tabular formulations, PARL (Chakraborty et al., 2024) studies policy alignment with parameterized policies and accounts for the dependence of the upper sampling distribution on the lower response. Its formulation accommodates continuous states, but its A-PARL algorithm uses implicit hypergradient calculations involving mixed derivatives and an inverse lower-level Hessian. It is therefore not fully first order, and its convergence analysis assumes access to the required derivative oracles. Gaur et al. (2025) establish a first-order $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ sample-complexity bound for continuous state–action spaces under a uniform lower-level Polyak–Łojasiewicz (PL) condition.

Uniform PL does not follow from smooth policy parameterization alone. Our counterexample in Proposition C.52 has smooth positive action probabilities and exactly representable values, yet admits a suboptimal stationary point, violating PL. This motivates specifying an actor construction, deriving its required local geometry, and proving that the learning dynamics remain in the region where that geometry holds. Additionally, PL conditions in general do not hold for arbitrary RL setups. Thus works which rely on PL assumptions are not applicable for arbitrary actor and critic classes.

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.

Additionally, critic learning introduces a distinct difficulty. The value targets change with both the reward and the policy, and the response correction uses differences between nearby policies divided by their separation. Separate bounds on the two critic errors do not directly control their difference after this normalization as the separation shrinks. A coupled guarantee must therefore establish persistent critic stability and track both the individual value errors and their normalized endpoint differences. Thus we have the following question.

Can we obtain improved sample-complexity guarantees for single-loop, first-order bilevel reinforcement learning beyond tabular settings, using explicit neural actors and nonlinear critics, while deriving the required local geometry and controlling their coupled learning errors?

We answer this question in the affirmative and address these challenges with explicit tied multilayer actors and a nonlinear critic built from fixed features. An initialization covariance condition, architectural derivative bounds, and fixed reference-policy KL regularization yield local curvature; a joint containment proof makes this geometry valid throughout training with high probability. Shared trajectories control the noise in the response correction. Under feature excitation and realizability of the regularized value family, our paired critic analysis establishes stability and tracking of the required normalized errors. Residual-controlled actor solves complete the finite, first-order implementation. Together, these ingredients yield $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity for $O(\varepsilon)$ expected squared upper-gradient norm, including persistent critic training, preference sampling, and metric initialization. The guarantee covers continuous states and finite actions. under the stated architecture, initialization, regularity, and value-representation conditions.

- **Variance control for coupled response estimation.** We estimate the policy-response correction by evaluating nearby policies on shared, appropriately weighted trajectories. This coupling keeps sampling variance controlled as the policies being compared become closer, allowing the response approximation to become more accurate without amplifying sampling noise. Together with paired critic-error control and persistent actor–critic tracking, this mechanism supports our improved sample complexity relative to Gaur et al. (2025).
- **Improved interaction complexity with single-loop first-order learning.** We establish the $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ interaction bound while reward, actor, and critic updates proceed together. Shared trajectory sampling controls response-estimation noise as the perturbation shrinks, and a combined descent and tracking potential accounts for all three learning processes and their sampling costs.
- **Explicit actor–critic architectures with coupled learning guarantees.** We specify multilayer neural actors and a nonlinear critic built from fixed features. We show that smooth policies and exact value realizability do not imply uniform PL, and derive local actor curvature from initialization identifiability and fixed reference-KL regularization. We prove containment of the coupled iterates and stability of persistent critic learning, controlling critic errors as rewards, policies change. A continuous-state construction verifies the actor and critic conditions (Appendix C.11).

Work	Continuous State Space	Lower geometry	Nonlinear A–C analysis	SL	FO	Sample complexity
Chakraborty et al. (2024)	✓	PL; Hessian conditions	×	×	×	×
Shen et al. (2025)	×	Policy-space structure	×	×	✓	×
Yang et al. (2025)†	×	Soft Bellman structure	×	×	✓	$\widetilde{\mathcal{O}}(\varepsilon^{-3.5})$
Gaur et al. (2025)	✓	Uniform PL	×	×	✓	$\widetilde{\mathcal{O}}(\varepsilon^{-3})$
Zeng et al. (2026)‡	×	Reg.-dependent PL; Hessian conditions	×	✓	✓	$\widetilde{\mathcal{O}}(\varepsilon^{-3})$
This work	✓	Derived local curvature	✓	✓	✓	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$

Table 1: Analyzed BRL guarantees. SL: single-loop stochastic tracking; FO: first-order implementation. Nonlinear A–C analysis requires explicit nonlinear actors and critics with coupled training and critic-error control. Dashes mark unestablished guarantees or unreported comparable sample complexitys. Rates use $O(\varepsilon)$ squared stationarity under distinct targets and assumptions. †Stoc-SoBiRL; its deterministic model-based variant uses $O(1)$ inner updates. ‡Fixed regularization, Theorem 2 of version 6; best-iterate output. Our uniform-output bound concerns a selected local regularized response, globally optimal on $\mathcal{D}$ under additional PL (Corollary A.2).

2 PROBLEM FORMULATION

MDP and sampling access. For a reward parameter $x \in \mathcal{O}_x = \mathbb{R}^{d_x}$, let $\mathcal{M}_x = (\mathcal{S}, \mathcal{A}, P, r_x, \rho, \gamma)$ be a discounted Markov decision process: $\mathcal{S}$ is compact and possibly continuous, $\mathcal{A}$ is finite, $P(ds' \mid s, a)$ is the transition law, $r_x(s, a)$ is the reward, ρ is the reset distribution, and $\gamma \in (0, 1)$ is the discount. Only the reward and upper sample loss depend explicitly on x. We can independently reset from ρ, generate trajectories under the current policies, and query feedback on trajectory tuples. The reward and its first derivatives are evaluable; the transition kernel need not be known.

Multilayer neural actor. Let L be the hidden depth, $\mathbf{m} = (m_1, \dots, m_L)$ the hidden widths, and $\alpha \in \mathbb{R}^r$ the trainable coordinates. A scalar score $F_{\mathbf{m},\alpha}$ tilts a fixed strictly positive reference policy:

$$\pi_{\mathbf{m},\alpha}(a \mid s) = \frac{\pi_{\mathrm{ref}}(a \mid s) e^{F_{\mathbf{m},\alpha}(s,a)}}{\sum_{b \in \mathcal{A}} \pi_{\mathrm{ref}}(b \mid s) e^{F_{\mathbf{m},\alpha}(s,b)}}. \tag{1}$$

Two coupled smooth multilayer networks define the score. Their weights vary through a fixed map of the same r coordinates around initialization; we call this map the retained actor chart. Depth L and dimension r remain fixed as widths vary, and $F_{\mathbf{m},0} = 0$ gives $\pi_{\mathbf{m},0} = \pi_{\mathrm{ref}}$. Appendix C.1 specifies the chart, normalization, inputs, and activations.

Lower and upper objectives. Let $\mathbb{P}_{\mathbf{m},\alpha}$ be the trajectory law under reset $s_0 \sim \rho$, actions $a_t \sim \pi_{\mathbf{m},\alpha}(\cdot \mid s_t)$, and transitions P; write $\mathbb{E}_{\mathbf{m},\alpha}$ for its expectation. Fix tuple size I and trajectory length K independently of accuracy. The tuple $\boldsymbol{\xi} = (\xi^1, \dots, \xi^I)$ contains independent length-K state–action trajectories, with product law $(\mathbb{P}^K_{\mathbf{m},\alpha})^{\otimes I}$. Feedback follows a fixed conditional law $\mathcal{D}(\cdot \mid \boldsymbol{\xi})$, independent of (x, α) given the tuple. For sample loss $L(x; \boldsymbol{\xi}, Y)$ and fixed Kullback–Leibler (KL) regularization coefficient $\tau > 0$, define

$$g_{\tau,\mathbf{m}}(x, \alpha) = (1 - \gamma)\mathbb{E}_{\mathbf{m},\alpha} \sum_{t \geq 0} \gamma^t \left[-r_x(s_t, a_t) + \tau \log \frac{\pi_{\mathbf{m},\alpha}(a_t \mid s_t)}{\pi_{\mathrm{ref}}(a_t \mid s_t)} \right], \tag{2}$$

$$f_{\mathbf{m}}(x, \alpha) = \mathbb{E}_{\boldsymbol{\xi} \sim (\mathbb{P}^K_{\mathbf{m},\alpha})^{\otimes I},\, Y \sim \mathcal{D}(\cdot \mid \boldsymbol{\xi})} L(x; \boldsymbol{\xi}, Y). \tag{3}$$

Thus lower minimization maximizes regularized return, while the upper loss depends on the policy through its trajectory distribution. For Bradley–Terry comparisons, $I = 2$ and $Y \in \{0, 1\}$, with $Y = 1$ favoring ξ^1. Writing (s_t^i, a_t^i) for trajectory i, the loss is

$$\Delta_x = \sum_{t=0}^{K-1} [r_x(s_t^1, a_t^1) - r_x(s_t^2, a_t^2)], \qquad L(x; \boldsymbol{\xi}, Y) = \log(1 + e^{\Delta_x}) - Y\Delta_x. \tag{4}$$

Correct specification of the preference model is unnecessary (Appendix C.2).

Bilevel problem and stationarity target. Let $\mathcal{B}_{\mathrm{geom}} = \{\alpha : \|\alpha\|_2 < R_{\mathrm{geom}}\}$. For the fixed τ and radius supplied by Section 4, we study

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}} \quad & \Phi_{\tau,\mathbf{m}}(x) := f_{\mathbf{m}}(x, \alpha^*_{0,\mathbf{m}}(x)), \\ \text{where} \quad & \alpha^*_{0,\mathbf{m}}(x) \in \operatorname*{arg\,min}_{\alpha \in \mathcal{B}_{\mathrm{geom}}} g_{\tau,\mathbf{m}}(x, \alpha). \end{aligned} \tag{5}$$

The later conditions make this minimizer unique, interior, and smooth in x, with $\nabla_\alpha g_{\tau,\mathbf{m}}(x, \alpha^*_{0,\mathbf{m}}) = 0$. The ball selects a local response in the full actor space $\mathbb{R}^r$; subscript 0 denotes the unperturbed response. For accuracy $\varepsilon > 0$, our goal is an output satisfying $\mathbb{E}\|\nabla\Phi_{\tau,\mathbf{m}}(x_{\mathrm{out}})\|^2 = O(\varepsilon)$.

Value function and nonlinear critic. Hereafter suppress actor widths, write $g = g_\tau$, and use h_x, h_α for partial gradients of a function h. Define $q_{x,\alpha} = -r_x + \tau \log(\pi_\alpha / \pi_{\mathrm{ref}})$ and the unnormalized discounted action value

$$Q^{\mathrm{true}}_{x,\alpha}(s, a) = \mathbb{E}^{s,a}_\alpha \left[\sum_{t \geq 0} \gamma^t q_{x,\alpha}(s_t, a_t) \right],$$

where $\mathbb{E}_\alpha^{s,a}$ starts the process at $(s_0, a_0) = (s, a)$. For a fixed bounded feature map $\varphi_Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_Q}$, coordinates $\theta \in \mathbb{R}^{d_Q}$, and scale $m_Q \geq 1$, use

$$Q_{m_Q,\theta}(s, a) = \sqrt{m_Q} \tanh\big(\varphi_Q(s, a)^\top \theta / \sqrt{m_Q}\big) . \tag{6}$$

The scale controls nonlinearity at fixed dimension. Section 3 trains this critic persistently; Appendix C.11 jointly verifies the actor and critic assumptions in a continuous-state instance.

Response-chain correction. With response Jacobian $D_x\alpha_0^*(x) \in \mathbb{R}^{r\times d_x}$,

$$\nabla\Phi_\tau(x) = f_x(x, \alpha_0^*(x)) + D_x\alpha_0^*(x)^\top f_\alpha(x, \alpha_0^*(x)). \tag{7}$$

The second term accounts for policy adaptation. The next section estimates this total gradient.

3 Proposed Approach

Response correction through paired actors. Following first-order penalty constructions (Kwon et al., 2024), let $\alpha_c^*(x)$ solve $g_\alpha + cf_\alpha = 0$ locally in $\mathcal{B}_{\text{geom}}$ for perturbation $0 < c \leq \min\{c_0, 1\}$ and sufficiently small fixed c_0 (Section 4). Define

$$S_c^*(x) := \frac{\alpha_c^*(x) - \alpha_0^*(x)}{c}. \tag{8}$$

Persistent coordinates $U, S \in \mathbb{R}^r$ track $\alpha_0^*(x), S_c^*(x)$; the actors are $U, U + cS$. For $p = (x, U, S)$, set $a(p) = g_\alpha(x, U)$ and

$$R_c(p) = \frac{g_\alpha(x, U + cS) - g_\alpha(x, U)}{c} + f_\alpha(x, U + cS), \tag{9}$$

$$V_c(p) = \frac{g_x(x, U + cS) - g_x(x, U)}{c} + f_x(x, U + cS). \tag{10}$$

At these response targets, $R_c = 0$ and $V_c = \nabla\Phi_\tau(x) + O(c)$ under the conditions of Section 4 (Appendix B.1). Thus gradient differences recover policy adaptation using first-order evaluations.

Paired critic and certified neighborhoods. We learn two critics together through coordinates $z = (u, v)$: the base critic has parameter u, and the perturbed critic has parameter $u + cv$. Thus v tracks their separation relative to the perturbation scale c. To keep critic updates bounded, we project onto $\mathcal{K}_Q$, the set of pairs satisfying $\|u\| \leq R_u$ and $\|v\| \leq R_v$. The projection $\Pi_{\mathcal{K}_Q}^W$ uses the weighted distance induced by $\|z\|_W^2 = M\|u\|^2 + \|v\|^2$. Appendix B.2 constructs the positive radii and weight.

For the actors, let $s > 0$ bound the normalized separation and $r_{\text{prox}} > 0$ be the radius of each local actor solve. The analysis uses a prospective neighborhood, defined by $\|S\| < s$ and $\|U\| < R_{\text{geom}} - c_0 s - r_{\text{prox}}/2$, within which proposed updates can be analyzed. The algorithm checks the stricter condition

$$\mathsf{Safe}(U, S) : \qquad \|S\| < s/2, \qquad \|U\| < R_{\text{geom}} - c_0 s - r_{\text{prox}}.$$

The gap between these neighborhoods provides room for local actor updates before the next safety check. Appendix B.5 specifies the radii and proves that the iterates satisfy this test throughout training with high probability.

Finite actor updates. Each actor update balances following an estimated descent direction with keeping policy changes small. Set the actor step $\eta_y = r_Q\beta$ and upper step $\eta_x = q\eta_y$, where $\beta > 0$ is the critic step and $r_Q, q > 0$ are fixed ratios. Draw n independent reset states from ρ, separately from training data, and denote their empirical distribution by $\widehat{\rho}_n$. Choose n for rejection budget δ_{met} using (188), and check the initialization covariance using (191). This empirical certificate supplies the local KL geometry needed for the updates; rejection returns x_0. Store the accepted states and reuse them throughout training for both actors.

For a frozen actor parameter b and estimated gradient h, define

$$\ell_{b,h}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} D_{\text{KL}}\big(\pi_b(\cdot \mid s_i)\|\pi_\alpha(\cdot \mid s_i)\big) + \eta_y h^\top(\alpha - b). \tag{11}$$

Algorithm 1 Single-loop bilevel RL with a persistent paired critic

Require: Initial $x_0 \in \mathbb{R}^{d_x}$; iteration count T; $c, \tau, m_Q, H, \delta_{\mathrm{met}}$; batch sizes; step sizes $\beta, \eta_y = r_Q\beta, \eta_x = q\eta_y$; the specified geometry and radii.

1: Fix the chart; draw n independent reset states to form $\widehat{\rho}_n$ and certify it using (188) and (191).
2: Set $U_0 = S_0 = 0$ and deterministic $z_0 \in \mathcal{K}_Q$; **return** x_0 if the metric certificate or $\mathsf{Safe}(U_0, S_0)$ fails.
3: **for** $t = 0, \dots, T-1$ **do**
4: Freeze $p_t = (x_t, U_t, S_t)$ and z_t.
5: $(\widehat{a}_t, \widehat{R}_t) \leftarrow \mathsf{LowerPair}(p_t, z_t; H)$; $\widehat{C}_t \leftarrow \widehat{a}_t + c\widehat{R}_t$ exactly.
6: $U_{t+1} \leftarrow \widetilde{\mathsf{M}}(U_t, \widehat{a}_t)$; $\alpha^c_{t+1} \leftarrow \widetilde{\mathsf{M}}(U_t + cS_t, \widehat{C}_t)$.
7: $S_{t+1} \leftarrow (\alpha^c_{t+1} - U_{t+1})/c$.
8: $\widehat{V}_t \leftarrow \mathsf{OuterPair}(p_t; H)$; $x_{t+1} \leftarrow x_t - \eta_x \widehat{V}_t$.
9: $\widehat{F}_t \leftarrow \mathsf{TDPair}(p_t, z_t; H)$; $z_{t+1} \leftarrow \Pi^W_{\mathcal{K}_Q}(z_t - \beta\widehat{F}_t)$.
10: **return** x_0 if $\mathsf{Safe}(U_{t+1}, S_{t+1})$ fails.
11: **end for**
12: Draw an independent uniform index $J \in \{0, \dots, T-1\}$; **return** x_J.

The KL term discourages large policy changes on the stored states, while the linear term encourages descent along $-h$. The map $\widetilde{\mathsf{M}}(b, h)$ minimizes this objective approximately by projected gradient descent, starting at b and remaining within distance r_{prox}. Under the certified conditions, the stopping rule in Proposition C.33 guarantees distance at most $c\eta_y^2$ from the exact minimizer within $J_{\max}$ updates. These numerical steps use stored states and finite action sums, requiring no fresh environment interactions. Appendix C.4 gives the implementation details and constants.

Sampling routines and timing. At each iteration, all sampling routines use frozen pre-update parameters; the rollout cap H is distinct from the preference length K. LowerPair uses the paired critics and preference tuples to estimate (a, R_c), with $\widehat{C} = \widehat{a} + c\widehat{R}$ exactly. OuterPair estimates V_c using reward derivatives and separate preference tuples. TDPair estimates the paired temporal-difference (TD) field in (222), reusing each base-policy transition for both endpoint residuals. The sampling blocks are conditionally independent given the pre-draw history, which includes the metric sample. Preference tuples and labels come from the old perturbed actor, and the updated critic enters the next iteration. Appendix C.6 specifies the estimators and their interaction costs. Algorithm 2 gives a detailed expansion of one iteration of Algorithm 1, spelling out the sampling batches, estimator construction, and parameter updates.

Why shared samples matter. The response correction divides differences between nearby policy evaluations by c, which can amplify noise from independent endpoint estimates. Shared, appropriately weighted trajectories make the centered endpoint difference shrink with c, keeping the normalized sampling variance controlled as the deterministic approximation bias decreases. Steps 3 and 6 of the proof outline in Section 4 establish this variance control and explain its role in the sample complexity.

4 Theoretical Analysis

Fix $A > 0$, $\overline{B}_A = \{\alpha \in \mathbb{R}^r : \|\alpha\|_2 \le A\}$, and $0 < R_{\mathrm{geom}} \le A/2$. Derivatives use multilinear operator norms; order zero denotes the function. Bounds are uniform in admissible actor widths. After fixing the construction, global constants are independent of x, c, m_Q and accuracy unless stated otherwise. First we give the needed assumption as follows.

Assumption 4.1 (A1: primitive local regularity)**.** *For every compact $Q \Subset \mathcal{O}_x$ and $\alpha \in \overline{B}_A$, require:*

(i) ***Reward:*** *$r_x(s,a)$ is C^4 in x, with $\sup_{x\in Q,s,a} \|D^k_x r_x(s,a)\| \le C_{r,k}(Q)$ for $0 \le k \le 4$.*

(ii) ***Upper sample loss:*** *$L(x; \boldsymbol{\xi}, Y)$ is C^4 in x, with $\sup_{x\in Q,\boldsymbol{\xi},Y} \|D^k_x L\| \le C_{L,k}(Q)$ for $0 \le k \le 4$.*

(iii) ***Actor score:*** *the normalized tied construction satisfies $\sup_{\mathbf{m},\alpha\in\overline{B}_A,s,a} \|D^k_\alpha F_{\mathbf{m},\alpha}\| \le C_{F,k}(A)$ for $1 \le k \le 4$.*

All envelopes are finite and uniform in sample arguments and admissible widths; state or action differentiability is unnecessary.

Context. Related regularity appears in Marbach & Tsitsiklis (2001, Assumption 2) and Kwon et al. (2024, Assumptions 2 and 7(2)). Fourth-order bounds control paired KL-update remainders; the specified architecture verifies item (iii).

Assumption 4.2 (A2: initialization identifiability)**.** *In the retained r-dimensional chart, let $Z_{\mathbf{m},0} = D_\alpha F_{\mathbf{m},0}$. For a width-independent $\lambda_{\mathrm{eff,L}} > 0$, require*

$$\mathcal{G}_{\mathbf{m},0} := \mathbb{E}_{s\sim\rho}\,\mathrm{Cov}_{a\sim\pi_{\mathrm{ref}}(\cdot|s)}(Z_{\mathbf{m},0}(s,a)) \succeq \lambda_{\mathrm{eff,L}} I_r. \tag{12}$$

Context, verification, and derived geometry. Unlike the domain-wide Fisher conditions of Liu et al. (2020b, Assumption 2.1) and Fatkhullin et al. (2023, Assumption 3), A2 concerns only initialization. It admits a finite-sample certificate for an independently fixed chart (Proposition C.13) and analytical verification in Appendix C.11. Architectural derivative bounds and global regularity in G below propagate this gap locally. Sufficiently strong fixed reference-KL regularization then gives

$$\nabla^2_\alpha[g_{\tau,\mathbf{m}}(x,\alpha) + c f_{\mathbf{m}}(x,\alpha)] \succeq \mu_\tau I_r, \tag{13}$$

uniformly over x, $\alpha \in \mathcal{B}_{\mathrm{geom}}$, and $0 \le c \le c_0$, for $\mu_\tau > 0$ (Proposition B.1). This yields unique smooth interior responses; under C1–C2 and the stated schedule, containment of the coupled iterates is proved (Theorem B.13).

Initialization identifiability and derived geometry. Gaur et al. (2025) imposes uniform lower-level PL, which need not hold even for smooth positive policies (Proposition C.52). Our analysis instead starts from the initialization covariance condition in Assumption 4.2, verified analytically for our explicit construction in Appendix C.11. Together with regularity and fixed reference-KL regularization, this condition yields local response curvature (Proposition B.1). The coupled-update analysis then establishes containment within this region (Theorem B.13).

Assumption 4.3 (G: global upper regularity for preference losses)**.** *Use Section 2's fixed-dynamics, finite-action reset setting and objective (3) on $\mathcal{O}_x = \mathbb{R}^{d_x}$, with fixed tuple size I and trajectory length K. Tuples and labels are freshly sampleable. The conditional feedback law is independent of (x,α) given the trajectories, and L has no additional explicit actor dependence. For finite constants, uniformly in x and sample arguments,*

$$\begin{aligned} &|r_x| \le B_r, \quad \|\nabla_x r_x\| \le G_r, \quad \|\nabla^2_x r_x\|_{\mathrm{op}} \le L_r, \\ &|L| \le B_L, \quad \|\nabla_x L\| \le G_L, \quad \|\nabla^2_x L\|_{\mathrm{op}} \le L_L. \end{aligned} \tag{14}$$

These bounds hold for each realized label; A1's local C^4 regularity is retained.

Context and role. Related global reward and loss envelopes appear in Yang et al. (2025, Assumptions 7.2–7.3). For Bradley–Terry, the reward bounds imply the loss envelopes (Proposition C.14). These bounds remain uniform as x changes. Together with A1–A2, they allow sufficiently strong fixed reference-KL regularization to yield actor curvature uniformly in x, without restricting the upper iterates to a prescribed compact set.

Critic assumptions. The next two assumptions concern feature excitation and value representation. Let ν_U denote the normalized discounted state–action occupancy under π_U with reset distribution ρ. For a state–action function h, define $K_\alpha h(s,a) = \mathbb{E}[h(s',a') \mid s,a]$, where $s' \sim P(\cdot \mid s,a)$ and $a' \sim \pi_\alpha(\cdot \mid s')$. In C1–C2, (U,S) ranges over the prospective neighborhood in Section 3; C2 covers every interpolation $\alpha = U + tS$, $0 \le t \le c_0$, including the prospective update segments of Appendix B.5.

Assumption 4.4 (C1: fixed bounded critic features and excitation)**.** *The map $\varphi_Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_Q}$ and dimension d_Q are fixed independently of x, actors, m_Q, and accuracy, with $\sup_{s,a} \|\varphi_Q(s,a)\| \le F_Q < \infty$. Every base coordinate U in the stated neighborhood satisfies $\mathbb{E}_{\nu_U}[\varphi_Q \varphi_Q^\top] \succeq \lambda_Q I_{d_Q}$ for a uniform $\lambda_Q > 0$.*

Context and role. C1 rules out unidentifiable critic coefficient directions under discounted reset occupancy, supporting stable TD learning. Related feature conditions appear in linear TD (Bhandari et al., 2021, Section 2) and neural TD (Ke et al., 2024, Assumption 3.1).

Assumption 4.5 (C2: realizability of the regularized lower value)**.** *Choose sufficiently large finite τ for actor geometry and initialization as in Proposition B.1 and Theorem B.13, then keep it fixed before selecting critic constants. For every $x \in \mathbb{R}^{d_x}$ and actor α specified above, require*

$$Q^{\mathrm{true}}_{x,\alpha} = (\mathrm{Id} - \gamma K_\alpha)^{-1} q_{x,\alpha} \in \mathrm{span}\,\varphi_Q, \tag{15}$$

where $q_{x,\alpha}, Q^{\text{true}}_{x,\alpha}$ *are in Section 2,* Id *is the identity operator, and* $\operatorname{span}\varphi_Q = \{\varphi_Q^\top b : b \in \mathbb{R}^{d_Q}\}$.

Context and role. C2 requires the fixed feature span to represent the regularized value family, including its actor-dependent KL cost. Related representation conditions appear in Jin et al. (2020, Assumption A and Proposition 2.3); neural TD retians projection error (Ke et al., 2024, Theorem 3.7).

Paired accuracy of the nonlinear critic. At the paired population TD target $(u^*, v^*) = Z(p)$ from (50), with $p = (x, U, S)$, define the value errors $e_0^*(p) = Q_{m_Q,u^*} - Q^{\text{true}}_{x,U}$ and $e_c^*(p) = Q_{m_Q,u^*+cv^*} - Q^{\text{true}}_{x,U+cS}$. Appendix B.2 proves stability and

$$\mathcal{A}_Q := \sup_p \left\{ \|e_0^*(p)\|_\infty^2 + \left\| \frac{e_c^*(p) - e_0^*(p)}{c} \right\|_\infty^2 \right\} \le C m_Q^{-2}, \tag{16}$$

uniformly over the prescribed domain, with C independent of admissible c, m_Q. The key point is that the critic-error difference is $O(c/m_Q)$: division by c preserves the approximation scale. With $c^2 \asymp \varepsilon$ and $m_Q \asymp \varepsilon^{-1/2}$, the paired floor is $O(\varepsilon)$, whereas separate endpoint bounds yield only $O(1)$. This is a population approximation bound; sampling and moving-target errors from persistent training are controlled jointly with actor tracking, starting from any deterministic $z_0 \in \mathcal{K}_Q$ (Appendix B.3).(Appendix B.3).

4.1 Main Result

Now with the assumptions in place the main theorem is stated as follows,

Theorem 4.6 (Single-loop stationarity with a trained paired critic)**.** *Fix* γ, I, K*, actor depth and retained dimension, and critic feature dimension. Assume A1, A2, G, and C1–C2 for the specified actor and critic* (6)*, at the selected fixed* τ *and neighborhoods. Initialize* $U_0 = S_0 = 0$ *and any deterministic* $z_0 \in \mathcal{K}_Q$*. There exist fixed projection radii, actor buffers, positive step ratios* r_Q, q*, and positive constants independent of* ε *such that, for sufficiently small* $\varepsilon > 0$*, use*

$$\begin{gathered} \beta = k_\beta \varepsilon, \qquad \eta_y = r_Q \beta, \qquad \eta_x = q\eta_y, \qquad c^2 = k_c \varepsilon, \qquad T = \lceil k_T \varepsilon^{-2} \rceil, \\ H = \Theta\big((1-\gamma)^{-1} \log(1/\varepsilon)\big), \qquad m_Q \ge \max\{m_{Q,0}, C_m \varepsilon^{-1/2}\}. \end{gathered} \tag{17}$$

Choose constants as in Appendix B.5, with squared truncation biases summing to $o(\varepsilon)$*, positive TD/evaluation batches at most polynomial in* H*, and fixed preference batches as in Appendix C.6. Run Algorithm 1 with* $\delta_{\text{met}} = \varepsilon/2$*, its certified metric, and finite KL solves. Total certificate-failure probability is at most* ε*. The output* x_{out}*, uniform among training iterates on success and equal to* x_0 *otherwise, satisfies*

$$\mathbb{E}\|\nabla \Phi_{\tau,\mathbf{m}}(x_{\text{out}})\|^2 = O(\varepsilon), \qquad N_{\text{seg}}, N_{\text{trans}} = \widetilde{\mathcal{O}}(\varepsilon^{-2}), \tag{18}$$

where $N_{\text{seg}}, N_{\text{trans}}$ *count all sampled trajectory segments and environment transitions, including critic, evaluation, preference, and initialization data. The tilde suppresses logarithmic factors in* $1/\varepsilon$*.*

Interpretation of the guarantee. The theorem establishes stationarity of the upper objective induced by the selected policy response at fixed reference-KL regularization. The sample complexity includes all training and initialization data, including persistent critic learning from potentially inaccurate initial value estimates. Finite KL solves reuse stored metric states, with their numerical work accounted for separately (Appendices C.4 and B.6). Appendix A.1 compares our optimization target and sample complexity with prior guarantees. Under an additional domain PL condition, the selected response is a global lower minimizer as shown in (Corollary A.2).

4.2 Proof Outline

Write $\Phi = \Phi_{\tau,\mathbf{m}}$. The derived bounds $\|\nabla\Phi\| \le G_\Phi$ and $\Phi \ge \underline{\Phi}$ hold globally in x, with $\underline{\Phi} = 0$ for Bradley–Terry and $-B_L$ otherwise (Equation (43)). Conditional expectations $\mathbb{E}_t$ use the pre-draw history, including the accepted metric sample. Appendix B follows the six steps below.

1. Derived actor geometry and first-order response correction. (Appendix B.1.) A2, the reset component of occupancy, and the architectural derivative bounds yield local reference-KL curvature. Sufficiently strong fixed τ dominates the reward and perturbation curvature, giving unique smooth interior base and perturbed responses uniformly in x (Proposition B.1). Their normalized difference yields $V_c(x, \alpha_0^*(x), S_c^*(x)) = \nabla\Phi(x) + O(c)$ (Lemma B.3). This identifies the first-order population direction to be tracked, with squared response-approximation bias $O(c^2)$ and without assuming uniform lower-level PL.

2. Stable paired targets and normalized approximation errors. (Appendix B.2.) The reset identity and Jensen's inequality give, for square-integrable h,

$$\mathbb{E}_{\nu_U}[h(h - \gamma K_U h)] \geq (1 - \sqrt{\gamma})\|h\|^2_{L^2(\nu_U)}. \tag{19}$$

C1 transfers this coercivity to linear TD; the $O(c)$ actor and $O(m_Q^{-1})$ nonlinear field and derivative perturbations preserve stability. This gives an interior paired target $Z(p)$ with two bounded derivatives (Propositions B.5 and B.6). Under C2, comparing the TD roots and their actor derivatives with the realizable value family gives $O(m_Q^{-1})$ control of both the value error and its derivative along the actor path. Integrating this derivative between the endpoints yields $\|e_c^*(p) - e_0^*(p)\|_\infty = O(c/m_Q)$. Together with $\|e_0^*(p)\|_\infty = O(m_Q^{-1})$, this establishes $\mathcal{A}_Q = O(m_Q^{-2})$ in (16), uniformly in the response perturbation (Proposition B.7). Thus shrinking c does not amplify the normalized approximation error of the population critic targets. The next step accounts for stochastic tracking of these targets.

3. Shared samples, finite solves, and moving-target tracking. (Appendix B.3.) Shared, appropriately weighted trajectories control endpoint fluctuations before division by c. On the certified neighborhoods, the centered endpoint difference has per-sample second moment $O(c^2)$, so the normalized correction has variance $O(B^{-1})$ for a batch of B shared trajectories, uniformly in c. Uniform geometric-length moments permit fixed positive batches (Lemma B.8). Evaluation uses the old critic and $\widehat{C} = \widehat{a} + c\widehat{R}$, while a conditionally independent TD block tracks $Z(p)$ (Lemma B.10). For moving targets, conditional independence and centering cancel linear noise cross terms, leaving the $1/\beta$ penalty only on the mean target drift (Lemma B.9). The empirical metric and finite KL solves give endpoint error at most $c\eta_y^2$, hence normalized error $O(\eta_y^2)$. These solves reuse stored metric states and require no additional environment samples (Propositions C.33 and C.37).

4. Joint tracking contraction. (Appendix B.4.) Define the critic tracking error $\mathcal{C}_t = \|z_t - Z(p_t)\|^2_W$. Let $\mathcal{T}_t$ be the nonnegative actor tracking potential in (41), controlling errors in U_t, S_t. For a fixed positive weight λ, set

$$\mathcal{D}_t = \mathcal{T}_t + \lambda r_Q \mathcal{C}_t. \tag{20}$$

Sufficiently small fixed step ratios absorb the mutual actor–critic error terms and give, for fixed $\rho_D, C_D > 0$ (Theorem B.11),

$$\begin{aligned} \mathbb{E}_t\mathcal{D}_{t+1} &\leq (1 - \rho_D\eta_y)\mathcal{D}_t + C_D\eta_y(qG_\Phi^2 + \omega + \beta), \\ \omega &= c^2 + \mathcal{A}_Q + b_{\text{ev}}^2 + b_{\text{TD}}^2, \end{aligned} \tag{21}$$

where $b_{\text{ev}}, b_{\text{TD}}$ bound truncation biases. This inequality holds before exit and through the prospective next step in the larger analytic neighborhood. The term qG_Φ^2 bounds target movement for containment; Step 6 retains the gradient term explicitly to avoid introducing this constant into the stationarity floor.

5. Proving containment of the coupled iterates. (Appendix B.5.) Choose actor geometry and fixed τ, then critic radii and potential weights, followed by r_Q to accommodate the permitted initial critic error and q to control target drift. Buffered prospective-step estimates justify the local recursions before containment is known. A stopped martingale bound then controls $\mathcal{D}_t$: under (17), fluctuations of order $\sqrt{\eta_y \log(T/\varepsilon)}\,\text{poly}(H)$ fit inside the actor buffers. This excludes a first actor exit with conditional probability at least $1 - \varepsilon/2$ for every accepted metric realization (Theorems B.12 and B.13). Projection contains the critic, and G supplies global bounds in the upper variable.

6. Upper stationarity and sample complexity. (Appendix B.6.) With fixed positive weights and constants, and $\Delta\Psi_t = \Psi_{t+1} - \Psi_t$, upper descent and tracking give (Proposition B.14)

$$\begin{aligned} \Psi_t &= \Phi(x_t) - \underline{\Phi} + a_\Psi q\mathcal{T}_t + b_\Psi q r_Q \mathcal{C}_t, \\ \mathbb{E}_t\Delta\Psi_t &\leq -a_*\eta_x\big(\|\nabla\Phi(x_t)\|^2 + \mathcal{T}_t + \mathcal{C}_t\big) + C_*\eta_x(\omega + \beta + \eta_y + \eta_x). \end{aligned} \tag{22}$$

Telescoping with predictable pre-exit indicators and terminal nonnegativity avoids conditioning on future containment. Uniform output on success and fallback on failure yield

$$\mathbb{E}\|\nabla\Phi(x_{\mathrm{out}})\|^2 \le C\left[\frac{\Psi_0}{\eta_x T} + \omega + \beta + \eta_y + \eta_x\right] + \varepsilon G_\Phi^2. \tag{23}$$

The initial potential includes critic error. Under (17), $c^2 = \Theta(\varepsilon)$ makes the squared response-secant bias $O(\varepsilon)$, while $m_Q^{-2} = O(\varepsilon)$ gives the same order for the paired critic approximation floor. Truncation ensures $b_{\mathrm{ev}}^2 + b_{\mathrm{TD}}^2 = o(\varepsilon)$, hence $\omega = O(\varepsilon)$. The choices $\eta_x, \eta_y, \beta = \Theta(\varepsilon)$ and $T = \Theta(\varepsilon^{-2})$ also give $\Psi_0/(\eta_x T) = O(\varepsilon)$, so every term in the bound is $O(\varepsilon)$. The uniform sampling bounds from Step 3 allow these choices without a polynomial increase in batch size as c shrinks. Counting TD, response-evaluation, preference, and initialization samples therefore gives $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity. Appendix A.1.1 explains the $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity established by Gaur et al. (2025) and the source of our improved bound.

5 Empirical Results

We focus on interaction scaling in the exact continuous-state construction and end-to-end performance on Walker Walk.

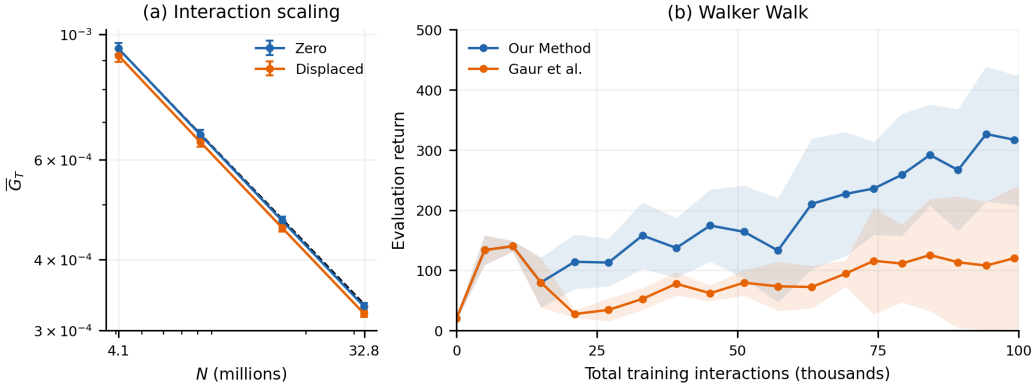

Figure 1: **Interaction scaling and continuous control.** (a) Average squared upper-gradient norm, $\overline{G}_T = T^{-1}\sum_{t<T}|\Phi'(x_t)|^2$, versus total training cost N, for zero and displaced critic initialization. Error bars show 95% t intervals over ten seeds; the dashed line is an $N^{-1/2}$ reference. (b) Walker Walk evaluation returns: means and one-standard-deviation bands across five matched training seeds.

We use the exact continuous-state, two-action construction in Appendix C.11, with a tied multilayer actor and a fixed-feature nonlinear critic. Training uses sampled trajectories; exact reference calculations provide the upper-gradient evaluation metric. Figure 1(a) shows similar decay of the average squared upper-gradient norm for an initially exact zero critic and a displaced critic with inaccurate value estimates, consistent with the predicted $N^{-1/2}$ trend and $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ interaction complexity. The prior works in Table 1 do not report this direct check of the predicted stationarity scaling against total sampled interactions. On Walker Walk, we learn rewards from trajectory preferences and incorporate the response correction into reward updates while training a continuous-control policy. Figure 1(b) shows higher mean returns than the adapted method of Gaur et al. (2025). Appendices D.1.1 and D.1.2 explain the scaling and initializations; Appendices D and E give protocols and implementation qualifications for the exact and end-to-end cases resepctively.

6 Conclusion

We developed a single-loop, fully first-order bilevel RL method with explicit neural actors and a persistently trained nonlinear critic. We proved $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity for $O(\varepsilon)$ expected squared upper-gradient norm under the stated assumptions.

AI USE STATEMENT

Generative AI tools were used for writing assistance. The authors verified all mathematical statements, proofs, citations, and experimental results, and take full responsibility for the content of the submission.

REPRODUCIBILITY STATEMENT

The supplementary material includes the experiment code, dependencies, configurations, and seeds, with setup instructions in the README in order to rerun the experiments from scratch, generate all given figures.

REFERENCES

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

Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. *Advances in Neural Information Processing Systems*, 35:8013–8026, 2022.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021.

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 24410–24449, 2024.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pp. 9827–9869. PMLR, 2023.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018.

Mudit Gaur, Utsav Singh, Amrit Singh Bedi, Raghu Pasupathy, and Vaneet Aggarwal. On the sample complexity bounds of bilevel reinforcement learning. In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020.

Kaiyi Ji. Lower complexity bounds for nonconvex-strongly-convex bilevel optimization with first-order oracles. In *International Conference on Machine Learning*, 2026.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

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

Zhifa Ke, Zaiwen Wen, and Junyu Zhang. An improved finite-time analysis of temporal difference learning with deep neural networks. In *Forty-first International Conference on Machine Learning*, 2024.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. On penalty methods for non-convex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, volume 2024, pp. 47473–47518, 2024.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International conference on machine learning*, pp. 6305–6315. PMLR, 2020a.

Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *International conference on machine learning*, pp. 6882–6892. PMLR, 2021.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020b.

P. Marbach and J.N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. Pmlr, 2015.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Yan Yang, Bin Gao, and Ya-xiang Yuan. Bilevel reinforcement learning via the development of hyper-gradient without lower-level convexity. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4780–4788. PMLR, 03–05 May 2025.

Sihan Zeng, Sujay Bhatt, Sumitra Ganesh, and Alec Koppel. A hessian-free actor-critic algorithm for bi-level reinforcement learning with applications to llm fine-tuning. *arXiv preprint arXiv:2601.16399*, 2026.

CONTENTS OF THE APPENDIX

A Related Work **12**
A.1 Optimization target and comparison with prior guarantees . . . 13

B Main Results **15**
B.1 Actor geometry and response correction . . . 15
B.2 Stable paired TD targets and derived accuracy . . . 18
B.3 Shared samples and persistent tracking . . . 20
B.4 A contracting potential for containment . . . 22
B.5 Constructively excluding a first exit . . . 22
B.6 Stationarity and complete data accounting . . . 24

C Supporting Results **25**
C.1 Actor construction and initialization certificates . . . 25
C.2 Preference losses and global mixed-derivative bounds . . . 32
C.3 Response and tracking-potential calculus . . . 34
C.4 Common-KL geometry and finite local updates . . . 36
C.5 Critic calculus and auxiliary TD estimates . . . 45
C.6 Sampling routines and paired oracle estimates . . . 46
C.7 Conditional tracking estimates . . . 49
C.8 Analytic and probabilistic tools . . . 50
C.9 Assumption verification and the PL counterexample . . . 52
C.10 KL-regularized counterexamples and degenerate neural policies . . . 52
C.11 A concrete continuous-state instance with a nontrivial response correction . . . 55
C.12 Conditional global lower responses on a prescribed domain . . . 67

D Continuous-state experimental details **69**
D.1 Instance, parameterization, and reference calculations . . . 69
D.2 Critic initialization and recovery . . . 70

E End-to-end continuous-state benchmark details **71**
E.1 Cohort, uncertainty, and interaction accounting . . . 71
E.2 End-to-end implementation . . . 71

A RELATED WORK

Bilevel optimization and RL. Implicit bilevel methods control Hessian inverses or iterative hypergradients (Franceschi et al., 2018; Grazzi et al., 2020; Ji et al., 2021); penalty methods permit first-order stochastic updates (Kwon et al., 2024). In BRL, PARL uses implicit response calculations (Chakraborty et al., 2024), penalty approaches prescribe lower-solution accuracy (Shen et al., 2025), and soft Bellman structure yields tabular hypergradients (Yang et al., 2025). Gaur et al. (2025) analyze continuous state–action spaces under uniform lower PL, with value-estimation error controlled by an approximation assumption and an imported actor–critic bound. Zeng et al. (2026) give a single-loop tabular actor–critic guarantee under regularization-dependent PL and Hessian conditions. Table 1 compares assumptions; Section A.1 distinguishes targets.

Critic construction and coupled learning. Prior BRL implementations already use neural critics (Shen et al., 2025; Yang et al., 2025). We analyze an explicit nonlinear critic whose persistent paired updates track moving values and control the normalized difference of endpoint errors as the response perturbation shrinks. Separate endpoint accuracy need not provide this control after division by the perturbation. The proof requires fixed bounded features, excitation, and realizability of the regularized value family (Appendices B.2 and C.11); it does not establish unrestricted feature learning or nonlinear expressive superiority.

Policy geometry and TD. Natural policy gradients and trust regions motivate Fisher geometry (Kakade, 2001; Schulman et al., 2015; Agarwal et al., 2021). Our initialization certificate yields local response curvature, in contrast to domain-wide Fisher nondegeneracy (Liu et al., 2020b; Fatkhullin et al., 2023). Feature-conditioned TD stability (Bhandari et al., 2021; Ke et al., 2024) supports our nonlinear paired tracking proof. The context after A1–C2 details the restrictions.

Bilevel optimization. Bilevel optimization couples an upper objective to the solution of a parameterized lower problem. With smooth, strongly convex lower objectives, the unique lower response supports convergence analyses based on implicit differentiation or differentiation through lower updates (Grazzi et al., 2020; Ji et al., 2021). For convex lower objectives with multiple minimizers, bilevel descent aggregation incorporates the upper objective when selecting among lower solutions (Liu et al., 2020a). Nonconvex lower problems introduce further difficulties: stationary points can be suboptimal, and the response reached by optimization can depend on initialization. Approaches to this setting include value-function reformulations (Liu et al., 2021) and selection maps that explicitly specify which lower critical point defines the upper objective (Arbel & Mairal, 2022). Our work develops this local-response perspective for BRL, deriving response curvature for explicit neural actors and controlling the coupled learning of policies and persistent nonlinear critics.

Comparison with lower bounds. Ji (2026) establish an $\Omega(\delta^{-4})$ stochastic oracle lower bound for zero-respecting first-order methods on nonconvex-strongly-convex bilevel problems, where δ denotes expected gradient-mapping norm accuracy. With fixed problem constants and $\delta = \sqrt{\varepsilon}$, this yields an $\Omega(\varepsilon^{-2})$ benchmark for squared stationarity accuracy, matching the accuracy exponent of our $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ interaction bound. This comparison does not establish minimax optimality for our structured RL setting, since the problem classes and oracle models differ; matching condition-number dependence also remains open.

A.1 Optimization target and comparison with prior guarantees

The target $\Phi_{\tau,\mathbf{m}}$ uses fixed reference-KL regularization: c is an estimation perturbation with vanishing bias, whereas τ remains part of the lower problem. Yang et al. (2025) use a global entropy-regularized tabular response; Gaur et al. (2025) global lower responses in their parameter class under uniform PL; Shen et al. (2025) a penalty reformulation linked to lower optimality; and Zeng et al. (2026) fixed and vanishing regularization under their geometric conditions. Our explicit structure yields local curvature and joint containment. We can show that under similar assumptions to Gaur et al. (2025) we can attain the same stationarity guarantee for a selected globally optimal lower response.

Assumption A.1 (Optional PL on a prescribed actor domain)**.** *Fix the finite actor architecture* $\mathbf{m}$ *and* τ *of Theorem 4.6, and a compact convex* $\mathcal{D} \subset \mathbb{R}^r$ *with* $\overline{B}_{R_{\mathrm{geom}}} \subset \operatorname{int} \mathcal{D}$*, independent of* x, c, m_Q, ε*. With* $g^*_{\tau,\mathbf{m},\mathcal{D}}(x) = \min_{\alpha\in\mathcal{D}} g_{\tau,\mathbf{m}}(x,\alpha)$*, assume a constant* $\mu_{\mathrm{PL}} > 0$ *independent of* x, ε *such that, for all* $x \in \mathbb{R}^{d_x}$ *and* $\alpha \in \mathcal{D}$*,*

$$\tfrac{1}{2}\|\nabla_\alpha g_{\tau,\mathbf{m}}(x,\alpha)\|^2 \geq \mu_{\mathrm{PL}}\big[g_{\tau,\mathbf{m}}(x,\alpha) - g^*_{\tau,\mathbf{m},\mathcal{D}}(x)\big]. \tag{24}$$

The gradient is the ambient Euclidean gradient, including at the boundary.

Corollary A.2 (Selected global regularized response)**.** *Under Theorem 4.6's conditions and Assumption A.1,* $\alpha^*_{0,\mathbf{m}}(x) \in \arg\min_{\alpha\in\mathcal{D}} g_{\tau,\mathbf{m}}(x,\alpha)$ *for every* x*. Thus* $\Phi^{\mathrm{sel}}_{\tau,\mathbf{m},\mathcal{D}}(x) := f_{\mathbf{m}}(x, \alpha^*_{0,\mathbf{m}}(x)) \equiv \Phi_{\tau,\mathbf{m}}(x)$ *is a selected global-response objective, and the same output satisfies* (18) *with unchanged label and numerical costs.*

Appendix C.12 proves the corollary, gives selection-independence conditions, and rules out full-space PL for this bounded architecture. Initialization, fixed τ, and upper stationarity are unchanged.

A.1.1 Source of the sample-complexity improvement

We compare the sample-complexity exponents for squared-gradient stationarity, retaining the distinctions between optimization targets and assumptions discussed above.

The published analysis. In the notation of Gaur et al. (2025, Theorem 1), let σ denote the perturbation, $T_{\mathrm{G}}, K_{\mathrm{G}}$ the outer and inner iteration counts, $n_{\mathrm{G}}, B_{\mathrm{G}}$ the sampling budgets, and H_{G} the rollout horizon. Writing $\mathcal{S}_{\mathrm{G}}$ for their averaged squared-stationarity criterion and $\varepsilon_{\mathrm{approx,G}}$ for their approximation term, their bound has the form

$$\mathcal{S}_{\mathrm{G}} \leq \widetilde{\mathcal{O}}\left(\frac{1}{T_{\mathrm{G}}} + \frac{e^{-K_{\mathrm{G}}}}{\sigma^2} + \frac{1}{\sigma^2 n_{\mathrm{G}}} + \frac{\gamma^{2H_{\mathrm{G}}}}{\sigma^2 B_{\mathrm{G}}} + \sigma^2 + \varepsilon_{\mathrm{approx,G}}\right). \tag{25}$$

Their stated schedule uses, up to logarithmic factors,

$$\sigma^2 \asymp \varepsilon, \qquad T_{\mathrm{G}} \asymp \varepsilon^{-1}, \qquad n_{\mathrm{G}}, B_{\mathrm{G}} \asymp \varepsilon^{-2}, \\ K_{\mathrm{G}}, H_{\mathrm{G}} = O(\log(1/\varepsilon)). \tag{26}$$

This gives $\mathcal{S}_{\mathrm{G}} = \widetilde{\mathcal{O}}(\varepsilon + \varepsilon_{\mathrm{approx,G}})$ and the sample count

$$N_{\mathrm{G}} = n_{\mathrm{G}} K_{\mathrm{G}} T_{\mathrm{G}} + B_{\mathrm{G}} K_{\mathrm{G}} H_{\mathrm{G}} T_{\mathrm{G}} + B_{\mathrm{G}} H_{\mathrm{G}} T_{\mathrm{G}} = \widetilde{\mathcal{O}}(\varepsilon^{-3}). \tag{27}$$

The term $1/(\sigma^2 n_{\mathrm{G}})$ exposes the estimation cost: with $\sigma^2 \asymp \varepsilon$, taking $n_{\mathrm{G}} \asymp \varepsilon^{-2}$ makes this contribution $O(\varepsilon)$.

Joint control of normalized response errors. Our analysis controls endpoint differences jointly before dividing by the response perturbation. Let $\bar{\zeta}_0, \bar{\zeta}_c$ denote the centered endpoint errors averaged over B shared trajectories. On the prescribed actor and critic neighborhoods,

$$\mathbb{E}_t \|\bar{\zeta}_c - \bar{\zeta}_0\|^2 \le \frac{C c^2}{B}, \qquad \mathbb{E}_t \left\| \frac{\bar{\zeta}_c - \bar{\zeta}_0}{c} \right\|^2 \le \frac{C}{B}, \tag{28}$$

with C independent of c and B (Lemma B.8). The shared-trajectory construction and its weighting therefore control the sampling variance uniformly as the perturbation shrinks.

The critic approximation requires a corresponding joint estimate. The paired population TD targets satisfy

$$\|e_0^*(p)\|_\infty^2 + \left\| \frac{e_c^*(p) - e_0^*(p)}{c} \right\|_\infty^2 \le C m_Q^{-2}, \tag{29}$$

uniformly over the prescribed neighborhood (Proposition B.7). Thus neither centered sampling noise nor the normalized critic approximation floor introduces an inverse power of c. The moving-target analysis additionally accounts for the error incurred while learning these critics persistently (Appendix B.3).

Propagation to the final sample complexity. The joint descent and tracking bound (23) combines these estimates with the initial-error contribution $\Psi_0/(\eta_x T)$. Under (17),

$$c^2 = \Theta(\varepsilon), \qquad m_Q^{-2} = O(\varepsilon), \\ \eta_x, \eta_y, \beta = \Theta(\varepsilon), \qquad T = \Theta(\varepsilon^{-2}). \tag{30}$$

The squared secant bias, paired critic approximation floor, step-size-weighted stochastic terms, and initial-error contribution are then all $O(\varepsilon)$. The prescribed truncation and metric-certificate accuracies control the remaining terms.

Uniform sampling moments allow fixed or logarithmically growing batches. Consequently, counting all TD, response-evaluation, preference, and initialization samples gives

$$N_{\mathrm{ours}} = \widetilde{\mathcal{O}}(T) = \widetilde{\mathcal{O}}(\varepsilon^{-2}), \tag{31}$$

including critic training. The improved exponent follows from joint control of errors in the divided response correction together with persistent actor–critic tracking. This comparison concerns the guarantees established by the two analyses under their respective assumptions.

Organization of the Appendix

The appendix separates the proof of the main theorem, its supporting constructions and estimates, and the experimental evidence. Appendix B gives the main proof sequence; Appendix C supplies its technical ingredients. Appendices D–E describe the experiments.

Appendix B: Main results. This section proves Theorem 4.6 in the order of the six-step proof outline in the main text. Section B.1 establishes the actor geometry and response correction. Section B.2 derives stable paired-TD targets and their approximation accuracy, while Section B.3 controls shared-sample estimation and persistent tracking. Sections B.4 and B.5 construct a contracting potential and exclude a first exit from the admissible region. Section B.6 then proves upper-objective stationarity and accounts for the data used in training and metric estimation.

Appendix C: Supporting results. Sections C.1–C.3 develop the multilayer actor construction, initialization certificates, preference-loss derivatives, and response calculus. Section C.4 establishes the common-KL geometry and implements the actor updates through an empirical metric and a finite local solver, including the effect of numerical errors. Sections C.5–C.8 provide the critic, sampling, conditional-tracking, and concentration estimates used in the main proof. Sections C.9 and C.10 discuss assumption verification and counterexamples to uniform lower-level PL guarantees. Section C.11 constructs a continuous-state instance with an explicit response, verified critic-feature conditions, and a nonzero chain correction; it also clarifies the interpretation of the on-policy preference objective. Section C.12 proves the optional global-response corollary, states selection-independence conditions, and explains the obstruction to uniform PL on the full actor space.

Appendix D: Controlled continuous-state experiments. This section specifies the actor and nonlinear critic, sampled training protocol, initialization comparisons, and reference calculations. It reports response-secant bias, shared versus independent sampling variance, critic approximation and tracking, and stationarity versus charged interactions. It also records the controls, uncertainty estimates, numerical safeguards, pilot history, and qualifications of the finite-budget scaling results.

Appendix E: End-to-end benchmarks. This section documents Walker Walk and Highway-fast, including the SAC implementation, preference collection, baselines, architectures, hyperparameters, learning curves, and checkpoint and per-seed returns. It also gives the uncertainty estimates and accounting of main and auxiliary training interactions.

B Main Results

This section proves Theorem 4.6 under A1, A2, G and C1–C2, following the six steps of the main-text proof outline. The target is the selected fixed-reference-KL response objective; the dynamics are fixed, actions are finite, and preferences use fixed finite tuples. The proofs of the actor construction and auxiliary calculations are in Appendix C.

Notation and sampling history. Suppress actor-width subscripts and write $g = g_\tau$, $\Phi = \Phi_{\tau,\mathbf{m}}$, $p = (x, U, S)$, $\alpha_0 = U$, and $\alpha_c = U + cS$. All norms are Euclidean or induced operator norms unless a subscript is given. Let $\mathcal{F}_t$ be the history before the current data blocks, including the old actors, upper parameter, critic, and independent initial metric sample. On acceptance, the common-KL law is fixed at $\widehat{\rho}_n$ throughout training, with the certificate of Proposition C.30. Write $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. The lower, outer and TD blocks are independent conditional on this history; the two preference blocks are separate and use the old perturbed actor. If needed, $\mathcal{G}_t$ additionally reveals the lower block and actor updates; outer samples remain fresh under that conditioning. The new critic is first used at the next iteration.

Constants denoted C may change between inequalities. They may depend on the fixed discount, tuple sizes, actor depth and retained dimensions, feature conditioning, and the sufficiently large but fixed τ. They are independent of x, accuracy, $0 < c \le c_0$, $m_Q \ge m_{Q,0}$, and admissible actor widths. Uniformity under later reductions of step ratios is stated where needed. L denotes actor depth, K preference trajectory length, H the truncation cap and N a geometric prefix length.

Policy transition and reset occupancy. For bounded measurable h, the policy transition operator and normalized discounted state–action occupancy used in C1–C2 are

$$\begin{aligned} (K_\alpha h)(s,a) &= \mathbb{E}_{s'\sim P(\cdot|s,a)} \sum_{a'\in\mathcal{A}} \pi_\alpha(a' \mid s') h(s',a'), \\ \rho_U(ds,da) &= \rho(ds)\pi_U(da \mid s), \qquad \nu_U = (1-\gamma)\sum_{k\ge 0} \gamma^k \rho_U K_U^k. \end{aligned} \tag{32}$$

Thus $\nu_U = (1-\gamma)\rho_U + \gamma \nu_U K_U$; its state marginal is d_U^γ.

B.1 Actor geometry and response correction

Write $j(x,\alpha) = (1-\gamma)\mathbb{E}_{\pi_\alpha} \sum_{t\ge 0} \gamma^t r_x(s_t,a_t)$ and $\mathcal{K}(\alpha) = (1-\gamma)\mathbb{E}_{\pi_\alpha} \sum_{t\ge 0} \gamma^t \log(\pi_\alpha/\pi_{\mathrm{ref}})(a_t \mid s_t)$, so $g = -j + \tau\mathcal{K}$. Put $h_c = g + cf$. The construction in Appendix C.1 supplies actor derivative

bounds $C_k = C_{k,L}$ and the KL Hessian Lipschitz constant L_K, independent of x, width, and τ. Let $a_0 = (1-\gamma)\lambda_{\mathrm{eff,L}}$ and $\kappa = a_0/4$, and choose

$$0 < R = R_A = R_{\mathrm{geom}} \le \min\{A/2, a_0/(2L_K), \sqrt{\lambda_{\mathrm{eff,L}}}/(2C_2)\}. \tag{33}$$

Omit a restriction with zero denominator. Write B_R for the open ball and $\Theta_R = \overline{B}_R$ for its closure. Define $J_q = B_r A_q(\gamma)$, $G_f = B_L b_1(IK)$ and $L_f = B_L b_2(IK)$ using the likelihood polynomials in Appendix C.2. Fix a separation radius $s = S_{\mathrm{tube}} > 0$ independently of τ; reduce the preliminary cap c_0 so $c_0 s \le R/8$. The common-KL radius $r_{\mathrm{prox}} \le R/8$ is chosen in Appendix C.4. All actor bounds below are used where $\|S\| \le s$ and $U + c\overline{B}_s \subset B_R$, with the additional local-update buffer specified in Step B.5.

Proposition B.1 (One certified actor neighborhood for every x)**.** *For a fixed perturbation cap c_0, choose a finite $\tau \ge 1$ satisfying*

$$\tau \ge \frac{4(J_2 + c_0 L_f)}{a_0}, \qquad \frac{J_1 + c_0 G_f}{\kappa\tau} < \frac{R}{4}. \tag{34}$$

Then $h_c = g_\tau + cf$ has actor Hessian at least $\mu_\tau I$, $\mu_\tau = \kappa\tau$, on B_R for every x and $0 \le c \le c_0$. Its unique root on this ball obeys

$$\|\alpha_c^\star(x)\| \le \frac{J_1 + cG_f}{\mu_\tau} < R/4, \qquad \|\alpha_c^\star(x) - \alpha_0^\star(x)\| \le \frac{cG_f}{\mu_\tau}. \tag{35}$$

The locally defined roots agree on overlaps and give a globally defined selected response on $\mathbb{R}^{d_x}$, smooth to the orders used below.

Proof. Lemma C.10 and the KL Hessian Lipschitz bound give $\mathcal{K}_{\alpha\alpha} \succeq (a_0/2)I$ on B_R. Lemma C.15 yields $h_{c,\alpha\alpha} \succeq (\tau a_0/2 - J_2 - c_0 L_f)I \succeq \mu_\tau I$. Since the KL gradient vanishes at initialization, $\|h_{c,\alpha}(x,0)\| \le J_1 + cG_f$. On any sphere of radius strictly between $(J_1 + cG_f)/\mu_\tau$ and R, strong monotonicity makes the gradient point outward. A minimizer on the closed smaller ball is therefore interior. Strong convexity gives uniqueness and the first distance bound. Subtract the root equations for c and 0 and use strong monotonicity to obtain the second bound. The implicit-function theorem applies because the actor Hessian is invertible; uniqueness identifies the local branches wherever they overlap. This asserts a selected local neural response, not an unrestricted global optimum. □

Let $G_{\alpha x}$, $G_{\alpha xx}$, $G_{\alpha\alpha x}$, and $G_{\alpha\alpha\alpha}$ denote the corresponding uniform derivative bounds for g_τ, and write $F_x, F_\alpha, F_{xx}, F_{x\alpha}, F_{\alpha\alpha}$ for those of f. Define

$$\begin{aligned} a_1 &= \mu_\tau^{-1} G_{\alpha x}, \\ a_2 &= \mu_\tau^{-1}\big(G_{\alpha xx} + 2G_{\alpha\alpha x} a_1 + G_{\alpha\alpha\alpha} a_1^2\big). \end{aligned} \tag{36}$$

Corollary B.2 (Global gradient, Hessian, and value bounds)**.** *The selected response satisfies $\|D_x\alpha_0^\star\| \le a_1$ and $\|D_x^2\alpha_0^\star\| \le a_2$. Consequently,*

$$\|\nabla\Phi_\tau(x)\| \le F_x + F_\alpha a_1 =: G_\Phi, \tag{37}$$

$$\|\nabla^2\Phi_\tau(x)\|_{\mathrm{op}} \le F_{xx} + 2F_{x\alpha}a_1 + F_{\alpha\alpha}a_1^2 + F_\alpha a_2 =: L_\Phi. \tag{38}$$

Also $-B_L \le \Phi_\tau(x) \le B_L$, with lower bound 0 for (138). The gradient is globally Lipschitz, and the usual quadratic upper Taylor bound holds along every upper-update segment.

Proof. Set $a(x) = \alpha_0^\star(x)$, $G_0 = g_{\alpha\alpha}(x, a(x))$, and $B = g_{\alpha x}(x, a(x))$. Differentiating $g_\alpha(x, a(x)) = 0$ gives $D_x a = -G_0^{-1}B$. In unit directions u, v, differentiating once more gives

$$\begin{aligned} G_0 D_x^2 a[u,v] = -\big(&g_{\alpha xx}[u,v] + g_{\alpha\alpha x}[D_x a[u], v] + g_{\alpha\alpha x}[D_x a[v], u] \\ &+ g_{\alpha\alpha\alpha}[D_x a[u], D_x a[v]]\big). \end{aligned}$$

The inverse bound $\|G_0^{-1}\| \le \mu_\tau^{-1}$ proves (36). Applying the chain rule twice to $f(x, a(x))$ proves the two bounds. Integrating the bounded Hessian along a segment in $\mathbb{R}^{d_x}$ proves the final statement. □

Lemma B.3 (Uniform consistency of the response correction)**.** *With the population field V_c in (10),*

$$\|V_c(x, \alpha_0^\star(x), S_c^\star(x)) - \nabla\Phi_\tau(x)\| \le C_{\mathrm{sec}} c, \qquad S_c^\star = (\alpha_c^\star - \alpha_0^\star)/c, \tag{39}$$

uniformly over x, width, and $0 < c \le c_0$. On the actor tube, the field is uniformly Lipschitz in (U, S), with no inverse power of c.

Proof. For fixed x, implicit differentiation in c gives $\partial_c \alpha_c^\star = -h_{c,\alpha\alpha}^{-1} f_\alpha(x, \alpha_c^\star)$, hence a uniform bound G_f/μ_τ. A second derivative satisfies

$$\partial_c^2 \alpha_c^\star = -h_{c,\alpha\alpha}^{-1} \left[h_{c,\alpha\alpha\alpha}[\partial_c \alpha_c^\star, \partial_c \alpha_c^\star] + 2 f_{\alpha\alpha} \partial_c \alpha_c^\star \right],$$

which is uniformly bounded by Lemma C.15. Taylor expansion in c of $g_x(x, \alpha_c^\star)$ and the Lipschitz bound for $f_x(x, \cdot)$ show that V_c differs by $O(c)$ from $f_x - g_{x\alpha} g_{\alpha\alpha}^{-1} f_\alpha$, evaluated at the base response. This is the exact response-chain gradient. Finally,

$$V_c(x, U, S) = \int_0^1 g_{x\alpha}(x, U + \theta c S) S \, d\theta + f_x(x, U + cS).$$

Differentiating this integral in U, S uses only the bounded blocks $g_{x\alpha}, g_{x\alpha\alpha}, f_{x\alpha}$ and bounded S. □

Actor tracking potential. For $c > 0$, define the scaled potential

$$\mathcal{H}_c(x, U, S) = \frac{g(x, U + cS) - g(x, U) - c\langle g_\alpha(x, U), S\rangle}{c^2} + \frac{f(x, U + cS) - f(x, U)}{c}. \tag{40}$$

Its S gradient is R_c from (9). Let $\bar{S}_c(x, U)$ be its unique interior minimizer over $\overline{B}_s$; Appendix C.3 proves existence for sufficiently large fixed τ and the integral extension at $c = 0$. Set $\Delta_0(x, U) = g(x, U) - g(x, \alpha_0^*(x))$ and $\Gamma_c(x, U, S) = \mathcal{H}_c(x, U, S) - \mathcal{H}_c(x, U, \bar{S}_c(x, U))$. For the fixed positive weight λ_S selected in Lemma C.20, the nonnegative actor potential is

$$\mathcal{T}(x, U, S) = \Delta_0(x, U) + \lambda_S \Gamma_c(x, U, S). \tag{41}$$

Lemma B.4 (Certified actor geometry and local update bounds)**.** *For the radii just selected, there are constants $\kappa, E_*, J_1, G_f > 0$. A sufficient finite fixed τ yields $\mu_\tau = \kappa\tau$ and, uniformly in x and sufficiently small c,*

$$\begin{aligned} \|\alpha_0^\star(x)\| &\le J_1/\mu_\tau, & \|S_c^\star(x)\| &\le G_f/\mu_\tau, \\ \|U - \alpha_0^\star(x)\|^2 + \|S - S_c^\star(x)\|^2 &\le E_* \mathcal{T}(x, U, S)/\mu_\tau, \\ \mathcal{T}(x, 0, 0) &\le \frac{J_1^2 + \lambda_S G_f^2}{2\mu_\tau}. \end{aligned} \tag{42}$$

Here E_*, λ_S can be fixed independently of sufficiently large τ. The needed gradients and full (x, U, S) Hessian of $\mathcal{T}$ are uniformly bounded on the prospective actor tube. Moreover,

$$\begin{aligned} \|\nabla\Phi\| &\le G_\Phi, \qquad \left\|\nabla^2\Phi\right\| \le L_\Phi, \qquad \Phi \ge \underline{\Phi}, \\ \|a\|^2 + \|R_c\|^2 &\le C\mathcal{T}, \qquad \|V_c\|^2 \le C(\|\nabla\Phi\|^2 + \mathcal{T} + c^2). \end{aligned} \tag{43}$$

For preference cross-entropy take $\underline{\Phi} = 0$; for a bounded general loss take $-B_L$.

Write $P^0 = \mathcal{F}_{\mathbf{m}, \widehat{\rho}_n}(U)^{-1}$, $P^c = \mathcal{F}_{\mathbf{m}, \widehat{\rho}_n}(U + cS)^{-1}$ and $K_c = (P^0 - P^c)/c$. On metric acceptance these operators obey the same fixed bounds by Proposition C.30. For compatible samples satisfying $\widehat{C} = \widehat{a} + c\widehat{R}$, the finite actor updates admit

$$\begin{aligned} \Delta U &= -\eta_y P^0 \widehat{a} + \mathfrak{r}_U, \\ \Delta S &= -\eta_y P^c \widehat{R} + \eta_y K_c \widehat{a} + \mathfrak{r}_S, \qquad \Delta x = -\eta_x \widehat{V}. \end{aligned} \tag{44}$$

Here $\mathfrak{r}_U = \rho^U + e_U$ and $\mathfrak{r}_S = \rho^S + e_S$, where ρ^U, ρ^S are the exact reference-map remainders and the finite-solver errors satisfy $\|e_U\| \le c\eta_y^2$, $\|e_S\| \le 2\eta_y^2$ samplewise. The predictable operators are uniformly bounded and the noiseless (U, S) direction decreases $\mathcal{T}$ by at least a fixed multiple of $\mathcal{T}$. The combined remainders have conditional first, second and fourth moments $O(\eta_y^2)$, $O(\eta_y^4)$ and $O(\eta_y^8)$, respectively. Deterministic actor increments are $\eta_y \operatorname{poly}(H)$ on sufficiently small prospective steps.

Proof. Proposition B.1 gives curvature and the two response displacements. Lemmas C.19 and C.20 prove the tracking self-bounds, full Hessian bound, distance control and initialization estimate without any compact upper domain. Corollary B.2 gives the objective bounds, and Lemma C.21 gives the corrected-field estimate. For the local updates, Lemmas C.24 and C.25 prove the exact reference expansion, including its inverse-metric correction. Proposition C.33 and Lemma C.35 give the stated implementation errors; no derivative of the numerical stopping rule is required. Lemma C.27 proves descent of its population direction. The moment assertions follow from Lemma C.45 and the sampled-field bounds in Lemma B.8. These are moment conditions derived from the sampling protocol, and do not assume accurate critic values. □

B.2 Stable paired TD targets and derived accuracy

For $p = (x, U, S)$, write $\alpha_0 = U$, $\alpha_c = U + cS$ and $z = (u, v)$, so the critics have coordinates u and $u + cv$. Let $q_{x,\alpha}$ and $Q^{\mathrm{true}}_{x,\alpha}$ have the convention in (15): the true value is an unnormalized discounted cost. The normalized state–action occupancy obeys

$$\nu_U = (1-\gamma)\rho_U + \gamma \nu_U K_U. \tag{45}$$

Define the untruncated TD mean field (distinct from the cap H) by

$$\begin{aligned} \mathsf{H}(x, U, \alpha, \theta) &= \mathbb{E}_{\nu_U}\left[(Q_{m_Q,\theta} - q_{x,\alpha} - \gamma K_\alpha Q_{m_Q,\theta})\nabla_\theta Q_{m_Q,\theta}\right], \\ F(p, z) &= \begin{pmatrix} \mathsf{H}(x, U, U, u) \\ [\mathsf{H}(x, U, U + cS, u + cv) - \mathsf{H}(x, U, U, u)]/c \end{pmatrix}. \end{aligned} \tag{46}$$

The operator K_α includes the next-state and next-action expectations. This is a TD semigradient field. The sample field $\widehat{\mathsf{H}}$, with exact finite next-action sums, is defined in (222); U denotes its base sampling law and has no additional pathwise role in that integrand.

The projection set and weighted norm are $\mathcal{K}_Q = \overline{B}_{R_u} \times \overline{B}_{R_v}$ and $\|z\|_W^2 = M\|u\|^2 + \|v\|^2$, with $W = \mathrm{diag}(MI_{d_Q}, I_{d_Q})$. The radii and $M > 0$ are constructed below; their definition precedes the projected update in (223). Let $\ell_{\max}$ bound $|\log(\pi_\alpha/\pi_{\mathrm{ref}})|$ on the prospective actor neighborhood and put $q_{\max} = B_r + \tau\ell_{\max}$.

Proposition B.5 (Global TD coercivity and interior roots)**.** *Set $\mu_0 = (1 - \sqrt{\gamma})\lambda_Q$. Choose*

$$R_Q > 4q_{\max}F_Q/\mu_0, \qquad \gamma F_Q^2 L_\pi c_0 s \le \mu_0/4, \tag{47}$$

where L_π is a uniform policy ℓ_1-Lipschitz constant. There is a finite $m_{Q,0}$ such that, for every $m_Q \ge m_{Q,0}$, $\mathsf{H}(x, U, \alpha, \cdot)$ is strongly monotone on $\overline{B}_{R_Q}$ with modulus $\mu_Q = \mu_0/2$, and has a unique root $\theta^*(x, U, \alpha)$ with norm strictly less than $R_Q/2$. The bounds are global in x and uniform in m_Q, c.

Proof. Lemma C.39 gives the nonlinear critic derivative and linearization bounds in (220).

For every scalar v, (45) and Jensen imply $\|K_U v\|_{L^2(\nu_U)} \le \gamma^{-1/2}\|v\|_{L^2(\nu_U)}$. Consequently

$$\mathbb{E}_{\nu_U} v(v - \gamma K_U v) \ge (1 - \sqrt{\gamma})\|v\|^2_{L^2(\nu_U)}. \tag{48}$$

For a unit $b \in \mathbb{R}^{d_Q}$ set $v = b^\top \nabla Q_{m_Q,\theta}$. Differentiating the TD field shows that its quadratic Jacobian form equals $\mathbb{E}v(v - \gamma K_\alpha v)$ plus the TD residual times $b^\top \nabla^2 Q b$. The residual has absolute value at most $D_Q = (1+\gamma)F_Q R_Q + q_{\max}$. Replacing K_α by K_U costs at most $\gamma F_Q^2 L_\pi c_0 s$. Also

$$\mathbb{E}_{\nu_U} \nabla Q \nabla Q^\top \succeq \mathrm{sech}^4(F_Q R_Q/\sqrt{m_Q})\lambda_Q I.$$

It follows that

$$\mathrm{sym}\, D_\theta \mathsf{H} \succeq \left[\mu_0 \,\mathrm{sech}^4(F_Q R_Q/\sqrt{m_Q}) - \frac{2D_Q F_Q^3 R_Q}{m_Q} - \gamma F_Q^2 L_\pi c_0 s\right] I. \tag{49}$$

Choose $m_{Q,0}$ so the first two deviations from μ_0 total at most $\mu_0/4$. Integration along line segments proves strong monotonicity with modulus μ_Q.

At zero, $\|\mathsf{H}(x, U, \alpha, 0)\| \le q_{\max}F_Q$. A variational-inequality solution exists on the closed ball by applying Brouwer to a projected field step. At the boundary its outward radial field satisfies $\langle \theta, \mathsf{H}(\theta)\rangle \ge \mu_Q R_Q^2 - R_Q q_{\max} F_Q > 0$, which excludes a boundary variational-inequality solution by testing against zero. Thus the solution is an interior root, unique by strong monotonicity. Its norm is at most $q_{\max}F_Q/\mu_Q < R_Q/2$. □

With θ^* the unique TD root just obtained, define its paired target

$$Z(p) = \left(\theta^*(x,U,U), \frac{\theta^*(x,U,U+cS) - \theta^*(x,U,U)}{c}\right) \tag{50}$$

Proposition B.6 (Uniform paired roots, projection set, and target derivatives)**.** *There are fixed radii R_u, R_v and a weight M for which $R_u + c_0 R_v < R_Q$, both endpoint critic parameters lie in the critic ball for every $z \in \mathcal{K}_Q$, and the target $Z(p)$ lies strictly inside $\mathcal{K}_Q$ with uniform slack. After reducing c_0 if needed, $F(p,\cdot)$ is strongly monotone and Lipschitz in the W norm, with constants uniform in x, m_Q, c. Its sample field is bounded uniformly in c. The target is twice continuously differentiable in p, with*

$$\|DZ\|_W \le L_Z, \qquad \left\|D^2 Z\right\|_W \le H_Z \tag{51}$$

uniformly over x and the prospective actor tube.

Proof. The inverse of $D_\theta \mathsf{H}$ has norm at most μ_Q^{-1}. Implicit differentiation gives

$$D_\alpha \theta^* = -(D_\theta \mathsf{H})^{-1} D_\alpha \mathsf{H},$$

uniformly bounded by Lemma C.40. Integrating this derivative along $U + tS$ bounds the second component of (50) by a constant times s. Choose R_u strictly between the uniform base-root bound and R_Q, and R_v strictly above that response bound. Reduce c_0 to get $R_u + c_0 R_v < R_Q$. The root and derivative constants can first be bounded for the preliminary cap; reduction does not invalidate them.

Keeping the sample and base law fixed gives the integral identity

$$\begin{aligned}&\frac{\widehat{\mathsf{H}}(x,U,U+cS,u+cv) - \widehat{\mathsf{H}}(x,U,U,u)}{c} \\ &= \int_0^1 \big\{ D_\alpha \widehat{\mathsf{H}}(x,U,U+tcS,u+tcv)[S] \\ &\qquad\qquad + D_\theta \widehat{\mathsf{H}}(x,U,U+tcS,u+tcv)[v] \big\}\, dt.\end{aligned} \tag{52}$$

It bounds the sample field without dividing two unrelated endpoint estimates by c. Its mean has block Jacobian

$$F_z = \begin{pmatrix} D_\theta \mathsf{H}_0 & 0 \\ B_c & D_\theta \mathsf{H}_c \end{pmatrix}, \qquad B_c = [D_\theta \mathsf{H}_c - D_\theta \mathsf{H}_0]/c.$$

The same integral formula bounds $\|B_c\| \le C_B$. For increments (a,b) its weighted quadratic form is at least $M\mu_Q \|a\|^2 + \mu_Q \|b\|^2 - C_B \|a\| \|b\|$. Choosing $M \ge 1 + C_B^2/\mu_Q^2$ gives strong monotonicity with modulus $\mu_F = \mu_Q/2$ in the W norm. Bounded blocks give Lipschitz continuity, and the equations $F = 0$ identify (50).

Twice differentiating the mean integral (52) uses derivatives of H of total order at most three, with at most two x derivatives and at least one actor or critic derivative in each third-order block. Lemma C.40 is exactly sufficient, including the derivatives of the base law. Thus $F_p, F_z, F_{pp}, F_{pz}, F_{zz}$ have uniform bounds. Applying the implicit-function theorem in the open neighborhood of each interior target gives

$$\begin{aligned} DZ &= -F_z^{-1} F_p, \\ D^2 Z[a,b] &= -F_z^{-1}\{F_{pp}[a,b] + F_{pz}[a, DZb] + F_{zp}[DZa, b] + F_{zz}[DZa, DZb]\}. \end{aligned} \tag{53}$$

Strong monotonicity bounds the inverse by μ_F^{-1} in the weighted norm. For bounds L_1 on F_p and L_2 on the second blocks, one may take $L_Z = L_1/\mu_F$ and $H_Z = L_2(1+L_Z)^2/\mu_F$. Uniqueness identifies the local branches on overlaps for every $x \in \mathbb{R}^{d_x}$. □

Proposition B.7 (The global paired approximation floor is derived)**.** *Under C2, write $(u^*, v^*) = Z(p)$ and define the endpoint errors*

$$e_0^*(p) = Q_{m_Q, u^*} - Q^{\text{true}}_{x,U}, \qquad e_c^*(p) = Q_{m_Q, u^* + cv^*} - Q^{\text{true}}_{x, U+cS}.$$

Then the squared paired approximation floor satisfies

$$\mathcal{A}_Q := \sup_p \left\{ \|e_0^*(p)\|_\infty^2 + \left\| \frac{e_c^*(p) - e_0^*(p)}{c} \right\|_\infty^2 \right\} \le C m_Q^{-2}. \tag{54}$$

This bound and the required projected-root coefficient derivative bounds are uniform in x.

Proof. Fix x, U, S and vary the endpoint $\alpha_t = U + tS$, $0 \le t \le c_0$, keeping ν_U fixed. The linear-feature TD field is $A_t\theta - b_t$, with

$$A_t = \mathbb{E}_{\nu_U}[\varphi_Q(\varphi_Q - \gamma K_{\alpha_t}\varphi_Q)^\top], \qquad b_t = \mathbb{E}_{\nu_U}[\varphi_Q q_{x,\alpha_t}].$$

Equation (48) and the perturbation bound give $\operatorname{sym} A_t \succeq (\mu_0/2)I$. Consequently its root $w_t = A_t^{-1}b_t$ is uniformly bounded and lies inside $B_{R_Q/2}$. All required coefficient derivatives follow by differentiating $A_t w_t = b_t$: A_t has only actor dependence, and b_t uses only the available reward derivatives and actor likelihood derivatives. In particular no smooth-coefficient hypothesis is needed in addition to C2. Exact realizability makes w_t the true-value coefficient, since the Bellman equation implies the TD equation and the latter has a unique root.

Let $\mathsf{H}_t^{m_Q}$ be the nonlinear field, with root $\theta_t^{m_Q}$. From (220), uniformly over the critic ball,

$$\begin{aligned}\left\|\mathsf{H}_t^{m_Q} - (A_t\theta - b_t)\right\| + \left\|D_\theta \mathsf{H}_t^{m_Q} - A_t\right\| \\ + \left\|\partial_t \mathsf{H}_t^{m_Q} - (\dot{A}_t\theta - \dot{b}_t)\right\| \le C/m_Q.\end{aligned}$$

Strong monotonicity implies $\left\|\theta_t^{m_Q} - w_t\right\| \le C/m_Q$. Differentiating both root equations, with $B_t = D_\theta \mathsf{H}_t^{m_Q}(\theta_t^{m_Q})$, gives

$$B_t(\dot{\theta}_t^{m_Q} - \dot{w}_t) = -[\partial_t \mathsf{H}_t^{m_Q}(\theta_t^{m_Q}) - (\dot{A}_t w_t - \dot{b}_t)] - (B_t - A_t)\dot{w}_t.$$

The inverse, coefficient, and derivative bounds imply $\left\|\dot{\theta}_t^{m_Q} - \dot{w}_t\right\| \le C/m_Q$. Therefore $e_{m_Q}(t) = Q_{m_Q,\theta_t^{m_Q}} - \varphi_Q^\top w_t$ has $\left\|e_{m_Q}(t)\right\|_\infty + \left\|\dot{e}_{m_Q}(t)\right\|_\infty \le C/m_Q$. Integrating $\dot{e}_{m_Q}$ from 0 to c bounds the normalized error difference by C/m_Q. Squaring proves (54). All constants were global in x before this calculation. □

B.3 Shared samples and persistent tracking

Condition on a realized accepted metric sample, included in $\mathcal{F}_t$. The three sampling routines in Appendix C.6 use the old p_t, z_t and conditionally independent data blocks. The numerical actor solvers use only the stored states and frozen lower block; Lemma C.36 preserves the conditional independence from the current TD block. All constants below are uniform over accepted metric realizations. Write $b_{\mathrm{TD}} = C_{\mathrm{TD}}\gamma^H$ and $b_{\mathrm{ev}} = C_{\mathrm{ev}}(H+1)^\ell\gamma^H$ for uniform truncation-bias bounds, with fixed constants and integer ℓ supplied there. Define

$$\mathcal{T}_t = \mathcal{T}(p_t), \quad \mathcal{C}_t = \|z_t - Z(p_t)\|_W^2, \quad X_t = \|\nabla\Phi(x_t)\|^2, \quad \omega = c^2 + \mathcal{A}_Q + b_{\mathrm{ev}}^2 + b_{\mathrm{TD}}^2. \tag{55}$$

All conditional inequalities below hold at a pre-exit point and through its prospective next step in the larger analytic actor neighborhood.

Lemma B.8 (Uniform oracle biases and moments)**.** *The squared conditional biases of the base, normalized-response, and corrected outer fields are at most*

$$C(\mathcal{C}_t + \mathcal{A}_Q + b_{\mathrm{ev}}^2), \qquad b_{\mathrm{ev}} \le C(H+1)^\ell\gamma^H \tag{56}$$

for a fixed integer ℓ. For the fixed-dynamics outer estimator its critic-dependent term can be omitted. Every fixed-order conditional sample moment is uniformly bounded in H; deterministic envelopes are polynomial in H. All statements are uniform in x, m_Q, c. Fixed positive batch sizes already satisfy the normalized moment requirements; TD/evaluation batches may grow at most polynomially in H, with fixed preference batch sizes.

Proof. Lemma C.43 controls the base and normalized critic errors before likelihood weighting, then transfers them through the shared-prefix estimators. Lemmas C.41 and C.42 control those weights; Proposition C.44 supplies unbiased bounded preference terms. Their geometric tail and moment estimates prove the displayed bounds. Every data block uses the old critic, so centering is conditional on the stated pre-draw history. For fixed dynamics the outer field contains only reward derivatives and direct preference terms, and hence no critic error. □

Lemma B.9 (Moving-target critic recursion with centered target movement)**.** *For sufficiently small steps, $\eta_x \le \eta_y$, and η_y/β sufficiently small, there are fixed positive k, C such that*

$$\mathbb{E}_t\Delta\mathcal{C}_t \le -k\beta\mathcal{C}_t + C\frac{\eta_y^2}{\beta}\mathcal{T}_t + C\frac{\eta_x^2}{\beta}X_t + C\beta\omega + C(\beta^2 + \eta_y^2 + \eta_x^2). \tag{57}$$

The constants can be fixed uniformly over all sufficiently small ratios η_y/β and η_x/η_y.

Proof. First hold the target at the old p_t. Projection is nonexpansive and $Z(p_t) \in \mathcal{K}_Q$. Strong monotonicity, Lipschitz continuity, bounded noise second moment, and (228) give

$$\mathbb{E}_t \|z_{t+1} - Z(p_t)\|_W^2 \le (1 - a\beta)\mathcal{C}_t + C\beta b_{\mathrm{TD}}^2 + C\beta^2 \tag{58}$$

for a fixed $a > 0$. Indeed expansion of the projected squared distance gives the cross term $-2\beta \langle z_t - Z, F\rangle$; absorb the bias cross term by Young's inequality, and the β^2 terms by a sufficiently small absolute step.

Write $\Delta p_t = d_t + w_{t+1}$, where $d_t = \mathbb{E}_t \Delta p_t$ and $\mathbb{E}_t w_{t+1} = 0$. The current TD block is conditionally independent of this entire increment. Reveal that block first, and set $\bar{e} = z_{t+1} - Z(p_t)$, $J = DZ(p_t)$. Then $\bar{e}, J, d_t$ are fixed under the enlarged conditioning and w_{t+1} stays centered. Taylor expansion gives

$$Z(p_t + \Delta p_t) - Z(p_t) = J(d_t + w_{t+1}) + R_t, \qquad \|R_t\|_W \le \tfrac{1}{2} H_Z \|\Delta p_t\|^2 .$$

Both critic points defining $\bar{e}$ lie in $\mathcal{K}_Q$, so $\|\bar{e}\|_W$ is bounded by its diameter. Prospective increments have a fixed deterministic bound. The remainder cross terms and its square are consequently bounded by $C\|d_t + w_{t+1}\|^2$. Centering cancels the linear cross term involving w_{t+1}. Use Young's inequality only for the deterministic displacement Jd_t, and combine with (58), to obtain

$$\mathbb{E}_t \mathcal{C}_{t+1} \le (1 - a'\beta)\mathcal{C}_t + C\beta b_{\mathrm{TD}}^2 + C\beta^2 + \frac{C}{\beta}\|d_t\|^2 + C\mathbb{E}_t\|w_{t+1}\|^2 . \tag{59}$$

In particular the centered target movement has no $1/\beta$ penalty.

Lemma C.45 gives the conditional mean, second moment and fourth moment estimates in (235). It uses the paired reference-map remainder, the samplewise finite-solver error from Lemma C.35, and the trained-critic oracle bias, including the fact that only the deterministic target drift is penalized by $1/\beta$ in (59).

Substitute (235) into (59). Choose a fixed upper bound on $(\eta_y/\beta)^2$ small enough to absorb the self-coupling $C(\eta_y^2 + \eta_x^2)\mathcal{C}_t/\beta$ into the negative drift. Since $\eta_x \le \eta_y \le \beta \le 1$, the remaining terms have the sizes in (57). The constants before absorption are uniform over ratios at most one; fixing a sufficiently small upper bound and then reducing the ratios preserves them. □

Lemma B.10 (Actor tracking and outer descent with the trained critic)**.** *For sufficiently small steps and η_x/η_y below a fixed threshold, there are fixed positive a, b, d, u, v, w, C with*

$$\mathbb{E}_t \Delta\Phi_t \le -a\eta_x X_t + b\eta_x \mathcal{T}_t + d\eta_x \mathcal{C}_t + C\eta_x\omega + C\eta_x^2, \tag{60}$$

$$\mathbb{E}_t \Delta\mathcal{T}_t \le -u\eta_y \mathcal{T}_t + v\eta_x X_t + w\eta_y \mathcal{C}_t + C\eta_y\omega + C(\eta_y^2 + \eta_x^2). \tag{61}$$

The constants can be fixed uniformly over subsequently smaller step-size ratios.

Proof. Use the simultaneous Taylor estimate of Lemma C.46. The population actor direction decreases $\mathcal{T}$ by $\rho_0\mathcal{T}$, where $\rho_0 > 0$ is fixed, and $\|\nabla\mathcal{T}\| \le C\sqrt{\mathcal{T}}$ by Lemma C.19. Thus lower conditional bias adds at most $C\eta_y\sqrt{\mathcal{T}_t}\sqrt{\mathcal{C}_t + \omega}$. Young's inequality leaves $\rho_0\eta_y\mathcal{T}_t/4$ and $C\eta_y(\mathcal{C}_t + \omega)$. The mean local-map remainders and all quadratic Taylor terms cost $C(\eta_y^2 + \eta_x^2)$. The finite numerical error contributes at most $C\eta_y^2$ because the potential gradient is uniformly bounded on the prospective neighborhood. Equivalently, Proposition C.37 transfers each one-step inequality from the virtual exact step at the same old iterate.

The upper contribution is bounded by $C\eta_x(\mathcal{T}_t + X_t + \mathcal{C}_t + \omega)$ because $\|V_c\| \le C(\sqrt{X_t} + \sqrt{\mathcal{T}_t} + c)$. Fix an upper bound on $q = \eta_x/\eta_y$ small enough to absorb its $\mathcal{T}_t$ coefficient. Since $\eta_x \le \eta_y$, the remaining terms give (61), with constants unchanged when q is subsequently reduced. This calculation uses the full Hessian of the tracking potential and requires no smoothness of the chosen metric as a function of the iterates.

For the outer step, global L_Φ-smoothness gives

$$\mathbb{E}_t \Delta\Phi_t \le -\eta_x \langle \nabla\Phi(x_t), V_c(p_t) + e_{V,t}\rangle + \tfrac{1}{2} L_\Phi \eta_x^2 \mathbb{E}_t \|\widehat{V}_t\|^2 .$$

Here $e_{V,t} = \mathbb{E}_t \widehat{V}_t - V_c(p_t)$. Lemma C.21 bounds the squared secant error by $C(\mathcal{T}_t + c^2)$, and Lemma B.8 bounds the additional squared bias by $C(\mathcal{C}_t + \mathcal{A}_Q + b_{\mathrm{ev}}^2)$. Young's inequality and bounded second moments yield (60). The actual fixed-dynamics outer estimator has no critic dependence; the displayed $d\eta_x\mathcal{C}_t$ is a harmless conservative bound. □

B.4 A CONTRACTING POTENTIAL FOR CONTAINMENT

The step ratios $r_Q = \eta_y/\beta$ and $q = \eta_x/\eta_y$ are positive constants at most one. We first construct a contracting actor–critic potential; the objective-containing potential used for stationarity is defined separately in Step B.6.

Theorem B.11 (Combined tracking contraction)**.** *Choose $\lambda > 0$ with $\lambda k > w$, and reduce the admissible upper bound on r_Q so that $\lambda C r_Q^2 \le u/2$ in (57). Define*

$$\mathcal{D}_t = \mathcal{T}_t + \lambda r_Q \mathcal{C}_t. \tag{62}$$

There are $\rho_D, C_D > 0$, uniform over subsequently smaller r_Q, q, such that

$$\mathbb{E}_t \mathcal{D}_{t+1} \le (1 - \rho_D \eta_y)\mathcal{D}_t + \eta_y D_\varepsilon, \qquad D_\varepsilon = C_D(q G_\Phi^2 + \omega + \beta). \tag{63}$$

Proof. Add (61) to λr_Q times (57). After dividing by η_y, the coefficients of $\mathcal{T}_t, \mathcal{C}_t, X_t$ are

$$-u + \lambda C r_Q^2, \qquad w - \lambda k, \qquad vq + \lambda C q^2 r_Q^2.$$

Let $\delta_C = \lambda k - w > 0$ and choose $\rho_D = \min\{u/2, \delta_C/\lambda\}$. Since $r_Q \le 1$, the negative terms dominate $-\rho_D(\mathcal{T}_t + \lambda r_Q \mathcal{C}_t)$. The X_t coefficient is at most $(v + \lambda C)q$, and $X_t \le G_\Phi^2$. The bias terms are at most $C(1+\lambda)\omega$. The divided step remainders equal, up to a fixed factor,

$$\eta_y + \eta_x^2/\eta_y + \lambda(\beta + \eta_y^2/\beta + \eta_x^2/\beta) \le C(1+\lambda)\beta.$$

Absorb these constants into C_D. No inverse power of a subsequently reduced ratio enters either ρ_D or C_D. The finite solver adds at most $C(1+\lambda r_Q)\eta_y^2$ to the unweighted one-step comparison in Proposition C.37. After division by η_y, this is $O(\eta_y) \le O(\beta)$, already included in D_ε. Thus the recurrence concerns the implemented process, with enlarged fixed remainder constants. □

B.5 CONSTRUCTIVELY EXCLUDING A FIRST EXIT

Use the actor radius $R_A = R$ and separation radius s from Step B.1. The local-solve radius r_{prox} obeys (157). Put

$$\begin{aligned} &S_{\max} = s/2, \quad c_0 s \le R_A/8, \quad r_{\mathrm{prox}} \le R_A/8, \\ &b_{\mathrm{cap}} = \min\{R_A/4, s/8\}, \quad T_{\mathrm{cap}} = b_{\mathrm{cap}}^2 \mu_\tau / E_*. \end{aligned} \tag{64}$$

The canonical observable certificate is

$$\mathsf{Safe}(U,S): \quad \|S\| < s/2 \quad \text{and} \quad \|U\| < R_A - c_0 s - r_{\mathrm{prox}}. \tag{65}$$

Let $\tau_a = \inf\{k \ge 0 : \mathsf{Safe}(U_k, S_k) \text{ fails}\}$, with infimum $+\infty$ for the empty set. The check includes $k = T$. The larger analytic neighborhood is $\|S\| < s$, $\|U\| < R_A - c_0 s - r_{\mathrm{prox}}/2$; it contains the prospective step from a safe point when the increments are sufficiently small. There is no upper-domain exit test. Lemma C.51 controls innovations through that prospective step, including projected critic updates.

Theorem B.12 (Maximum bound through a prospective first exit)**.** *Let τ_a be the first completed iteration at which an actor certificate fails. Suppose (63)–(240) hold before and through that prospective step, and $0 < \rho_D \eta_y \le 1$. With probability at least $1 - \delta$,*

$$\max_{0 \le k \le T \wedge \tau_a} \mathcal{D}_k \le \max\left\{\mathcal{D}_0, \frac{D_\varepsilon}{\rho_D}\right\} + B_D(H)\sqrt{\frac{2\eta_y}{\rho_D}\log\frac{2T}{\delta}}. \tag{66}$$

Proof. Put $a = 1 - \rho_D \eta_y$ and multiply each pre-exit innovation by the predictable indicator $I_j = \mathbf{1}_{\{j < \tau_a\}}$, setting it to zero afterward. For a deterministic k, on the event $k \le \tau_a$, iteration of the conditional inequality gives

$$\mathcal{D}_k \le a^k \mathcal{D}_0 + \frac{D_\varepsilon}{\rho_D}(1 - a^k) + \sum_{j=0}^{k-1} a^{k-1-j} I_j N_{j+1}.$$

The last sum is a weighted martingale sum for this fixed k. Its squared increment bounds sum to at most $\eta_y^2 B_D(H)^2 \sum_{i \ge 0} a^{2i} \le \eta_y B_D(H)^2/\rho_D$. The bounded-increment exponential martingale

inequality bounds its upper tail by $\delta/(2T)$ at the deviation in (66). A union bound over $k \le T$ gives the result (with slack in the failure probability). The first two deterministic terms are a convex combination. The comparison is asserted only through the exit; contraction of a frozen post-exit value is neither used nor needed. □

Theorem B.13 (Constructive containment for the trained critic)**.** *Under Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5, condition on any realized accepted metric sample from Definition C.28. Use the finite actor solvers and specified critic and sampling algorithm, initialize $U_0 = S_0 = 0$ and any deterministic $z_0 \in \mathcal{K}_Q$. There are finite fixed choices of regularization, radii, weights, positive step ratios, and positive sampling/tuning constants, uniform over accepted metric realizations, such that, for all sufficiently small ε,*

$$\beta = k_\beta \varepsilon, \qquad \eta_y = r_Q \beta, \qquad \eta_x = q\eta_y, \quad c^2 = k_c \varepsilon, \qquad T = \lceil k_T \varepsilon^{-2} \rceil, \tag{67}$$

with $m_Q \ge \max\{m_{Q,0}, C_m \varepsilon^{-1/2}\}$ and logarithmic H, satisfies

$$\mathbb{P}(\tau_a \le T \mid s_1, \dots, s_n) \le \delta_a, \qquad \delta_a = \varepsilon/2. \tag{68}$$

The critic remains in $\mathcal{K}_Q$ by construction. No upper compactness, accurate critic initialization, or explicit augmented containment margin is assumed.

Proof. **1. Fix actor geometry before critic constants.** The actor radius and tube are selected from the actor certificate. Increase the finite regularization τ until the actor curvature conditions hold and

$$J_1/\mu_\tau < R_A/4, \quad G_f/\mu_\tau < s/4, \qquad \mathcal{T}_0 < T_{\mathrm{cap}}/4. \tag{69}$$

This is possible by (42), since $\mathcal{T}_0 = O(\tau^{-1})$ and $T_{\mathrm{cap}} = \Theta(\tau)$. Now keep τ fixed. C2 concerns the lower criterion at this selected value. Construct $R_Q, m_{Q,0}, R_u, R_v, M$ as in Step B.2, reducing c_0 within the preliminary cap as needed. The resulting constants may depend on this τ; the proof never increases τ afterward.

2. Fix weights and then the ratios. Let $C_{\max} = \mathrm{diam}_W(\mathcal{K}_Q)^2$, which bounds $\mathcal{C}_0$ uniformly in x, m_Q, c. Fix the weight λ from Theorem B.11. Also fix the stationarity weights a_Ψ, b_Ψ from Proposition B.14 in Step B.6 before choosing ratios. Those weights depend only on the already fixed recursion constants, so this forward reference introduces no circular choice. Choose $r_Q > 0$ below all critic and combined-drift thresholds, with

$$\lambda r_Q C_{\max} < T_{\mathrm{cap}}/4. \tag{70}$$

It follows that $\mathcal{D}_0 < T_{\mathrm{cap}}/2$. This controls any bounded initial critic error without a preliminary training phase. Choose $q > 0$ below both tracking and stationarity thresholds and so small that

$$C_D q G_\Phi^2/\rho_D < T_{\mathrm{cap}}/4. \tag{71}$$

If $G_\Phi = 0$, this last restriction is vacuous. The constants C_D, ρ_D were fixed uniformly over smaller ratios, so these choices are noncircular.

3. Make the remaining floor and fluctuations small. After the deterministic prospective-step justification in Step 4, the maximum theorem applies with the following choices. Choose any fixed positive k_β, k_c, k_T, and take ε sufficiently small that all absolute step restrictions and $c \le c_0$ hold. Proposition B.7 gives $\mathcal{A}_Q = O(\varepsilon)$ with the stated m_Q. Take $H = O((1-\gamma)^{-1}\log(1/\varepsilon))$ large enough that $b_{\mathrm{ev}}^2 + b_{\mathrm{TD}}^2 = o(\varepsilon)$, and use the normalized batches from Lemma B.8. Hence $\omega + \beta = O(\varepsilon)$ and $D_\varepsilon/\rho_D < T_{\mathrm{cap}}/2$ for sufficiently small ε. With $\delta = \delta_a = \varepsilon/2$ and the schedule (67), the fluctuation term in (66) tends to zero: it is $\sqrt{\varepsilon}$ times a polynomial in $\log(1/\varepsilon)$. It is eventually less than $T_{\mathrm{cap}}/4$.

4. Justify the prospective step before using its inequality. At a pre-exit point, (65) leaves room for $U + c\overline{B}_s$ and the local solves. The samplewise normalized-pair bounds imply $\|\Delta U\| < r_{\mathrm{prox}}/2$ and $\|\Delta S\| < s/4$ when ε is sufficiently small, because each is at most $\varepsilon\,\mathrm{poly}(H)$. Thus the prospective endpoint and every segment to it lie in the larger actor tube with $\|S\| < s$ and with the endpoint policies in B_{R_A}. Choose the preliminary neighborhoods with strict extra slack if a closed ball is used in a derivative bound. All local-solver and target derivative estimates therefore apply to this step before it is known to remain in the smaller certified set. The critic update stays in $\mathcal{K}_Q$ by projection, and its target at the prospective endpoint lies in that same set by Proposition B.6.

Upper segments require no buffer restriction because the estimates are global in x. In detail, each numerical iterate remains in its buffered local ball by projection. The virtual reference update has the stated deterministic increment bound; the implemented update differs by at most $\sqrt{5}\eta_y^2$ in (U, S) by Lemma C.35. This proves the prospective claim for both updates, and the segment between them, without first assuming containment (Lemma C.36).

5. Contradict the first exit. On the event from Theorem B.12, the previous choices give $\mathcal{D}_k < T_{\mathrm{cap}}$ for every $k \le T \wedge \tau_a$. Since $\mathcal{T}_k \le \mathcal{D}_k$, (42) makes each actor error strictly smaller than b_{cap}. At a candidate first exit this implies

$$\|U_k\| < J_1/\mu_\tau + b_{\mathrm{cap}} < R_A/2, \qquad \|S_k\| < G_f/\mu_\tau + b_{\mathrm{cap}} < 3s/8 < s/2.$$

Also $c_0 s + r_{\mathrm{prox}} \le R_A/4$, so the distance of U_k from ∂B_{R_A} is strictly greater than the required buffer. Both certificates hold strictly, a contradiction. Therefore this event excludes $\tau_a \le T$ and has conditional probability at least $1 - \delta_a$. Averaging this uniform conditional bound and adding the independent metric rejection budget $\delta_{\mathrm{met}} = \varepsilon/2$ gives total failure probability at most ε. □

B.6 Stationarity and complete data accounting

The contracting potential $\mathcal{D}$ proves containment. For stationarity we use positive weights on the actor and critic errors and include the lower-bounded selected objective.

Proposition B.14 (Augmented objective descent)**.** *Choose $a_\Psi, b_\Psi > 0$ with $a_\Psi u > b + 1$ and $b_\Psi k > d + a_\Psi w + 1$, using the constants in (60)–(61). For sufficiently small fixed r_Q, q,*

$$\Psi_t = \Phi(x_t) - \underline{\Phi} + a_\Psi q \mathcal{T}_t + b_\Psi q r_Q \mathcal{C}_t \tag{72}$$

satisfies

$$\mathbb{E}_t \Delta\Psi_t \le -a_* \eta_x (X_t + \mathcal{T}_t + \mathcal{C}_t) + C_* \eta_x (\omega + \beta + \eta_y + \eta_x) \tag{73}$$

through the prospective actor exit, with fixed positive constants.

Proof. Add (60), $a_\Psi \eta_x/\eta_y$ times (61), and $b_\Psi \eta_x/\beta$ times (57). After division by η_x, the three coefficients are bounded above by

$$-a + a_\Psi v q + b_\Psi C q^2 r_Q^2, \qquad b - a_\Psi u + b_\Psi C r_Q^2, \qquad d + a_\Psi w - b_\Psi k.$$

Choose the weights first, then r_Q to make $b_\Psi C r_Q^2 < 1/2$, and q to make $a_\Psi v q + b_\Psi C q^2 r_Q^2 < a/2$. These can be imposed simultaneously with the containment choices. Weighted remainder terms have the bound in (73). □

Proof of Theorem 4.6. Fix the actor initialization and retained chart, and take $\delta_{\mathrm{met}} = \delta_a = \varepsilon/2$. Proposition C.29 bounds metric rejection by δ_{met}. Conditional on any accepted metric sample, Proposition C.30 supplies the same Fisher constants uniformly on the actor ball. All following constants are independent of the accepted realization. The finite-solver errors are already included in the tracking recursions by Proposition C.37. In particular their weighted contribution to Ψ is at most $Cq\eta_y^2 = C\eta_x\eta_y$, which is absorbed by (73).

For each such metric realization, stop at τ_a and freeze the process afterward. Multiply (73) by the predictable pre-exit indicator and telescope, including the prospective exit step:

$$a_* \eta_x \sum_{t<T} \mathbb{E}[\mathbf{1}_{\{t<\tau_a\}} X_t \mid s_1, \dots, s_n] \le \Psi_0 + C_* \eta_x T(\omega + \beta + \eta_y + \eta_x).$$

The terminal potential is nonnegative. For $t < T$, $\mathbf{1}_{\{\tau_a > T\}} \le \mathbf{1}_{\{t<\tau_a\}}$; Theorem B.13 gives the uniform conditional bound $\mathbb{P}(\tau_a \le T \mid s_1, \dots, s_n) \le \delta_a$. Draw the independent uniform output index on success and return x_0 on either failure. Averaging over the metric data and using $\|\nabla\Phi(x_0)\| \le G_\Phi$ gives

$$\mathbb{E}\,\|\nabla\Phi(x_{\mathrm{out}})\|^2 \le C\left[\frac{\Psi_0}{\eta_x T} + c^2 + \mathcal{A}_Q + b_{\mathrm{ev}}^2 + b_{\mathrm{TD}}^2 + \beta + \eta_y + \eta_x\right] + (\delta_{\mathrm{met}} + \delta_a) G_\Phi^2. \tag{74}$$

Here the initial potential is bounded uniformly over metric realizations: $\Phi(x_0) - \underline{\Phi} \le 2B_L$ (or B_L for Bradley–Terry), and $\mathcal{T}_0, \mathcal{C}_0$ have the fixed bounds proved above. This argument conditions

only on initial metric data and uses predictable stopping; it never conditions a descent inequality on future successful training. Metric acceptance reproduces the existing geometric constants, so no enlargement of the C2 neighborhood or increase in fixed τ is needed. The union of metric rejection and actor exit has probability at most $\delta_{\mathrm{met}} + \delta_a = \varepsilon$.

Under schedule (17), $m_Q \geq \max\{m_{Q,0}, C_m \varepsilon^{-1/2}\}$ gives $\mathcal{A}_Q = O(\varepsilon)$, and logarithmic H makes $b_{\mathrm{ev}}^2 + b_{\mathrm{TD}}^2 = o(\varepsilon)$. Also $(\eta_x T)^{-1} = O(\varepsilon)$ and every remaining term has order $O(\varepsilon)$. This proves the claimed stationarity bound.

Let B_C, B_H, B_V count TD, lower evaluation and outer discounted sample draws per iteration, and $B_{f,a}, B_{f,x}$ the tuple-batch sizes. A TD or discounted evaluation draw uses at most $H+1$ transitions; a tuple uses at most IK transitions. The complete conservative ledger is

$$\begin{aligned} N_{\mathrm{seg}} &\leq T[B_C + B_H + B_V + I(B_{f,a} + B_{f,x})] + n + N_{\mathrm{pilot,seg}}, \\ N_{\mathrm{trans}} &\leq T[(H+1)(B_C + B_H + B_V) + IK(B_{f,a} + B_{f,x})] + n + N_{\mathrm{pilot,trans}}. \end{aligned} \tag{75}$$

The n charges the one-time metric reset sample from (188). Each reset is conservatively charged as one segment and at most one transition; when no transition follows the reset its actual transition contribution is zero. Optional, separately collected initialization-pilot data are added once; they are unnecessary when A2 is assumed. The fixed-gap pilot in Proposition C.13, if used, needs $O(\log(1/\varepsilon))$ draws. For fixed $\gamma, r, G_Z, \lambda_0, I, K$, $n = O(\log(1/\varepsilon))$. Discounted training batches are fixed or at most polynomial in H, preference batches are fixed, and labels are charged for both preference blocks. Thus $N_{\mathrm{seg}}, N_{\mathrm{trans}} = \widetilde{O}(\varepsilon^{-2})$ and $N_{\mathrm{labels}} = O(\varepsilon^{-2})$; every persistent TD update is counted. Reusing metric states during numerical optimization incurs no additional environment interactions or preference labels.

There are at most two finite local solves per completed training iteration. Proposition C.33 therefore gives

$$N_{\text{KL-grad}} \leq 2T(J_{\max} + 1) = O\big(\kappa_K \varepsilon^{-2} \log(1/\varepsilon)\big), \qquad \kappa_K = M_{K,L}/m_{K,L}. \tag{76}$$

Indeed $G_{\mathrm{prox}}(H) = \mathrm{poly}(H)$, $H = O(\log(1/\varepsilon))$ and $c\eta_y = \Theta(\varepsilon^{3/2})$ in (205), while κ_K is independent of n and accuracy. A finite-sum gradient has at most $n|\mathcal{A}|$ state–action terms. If $C_{\mathrm{act}}(\mathbf{m}, r)$ bounds the arithmetic for one actor term, including its first derivative, the numerical work is

$$O(T(J_{\max} + 1)[n|\mathcal{A}|C_{\mathrm{act}}(\mathbf{m}, r) + r]).$$

Constructing the metric additionally needs $n|\mathcal{A}|$ actor-feature evaluations, covariance accumulation and one $r \times r$ spectral certificate. These finite computations are separate from environment interactions and from the other network arithmetic. Although endpoint parameter accuracy is $c\eta_y^2 = \Theta(\varepsilon^{5/2})$, geometric convergence of the local solver makes its additional iteration cost logarithmic; it does not introduce an $\varepsilon^{-5/2}$ sample exponent. □

C Supporting Results

These construction details and auxiliary results support the six-step proof in Appendix B. Assumptions and the selected target are those in the main text; no additional upper localization or oracle-accuracy assumption is introduced for the main theorem. The optional global-response results in Section C.12 explicitly state their additional conditions.

C.1 Actor construction and initialization certificates

The following normalized architecture supplies A1(iii). A2 is the separate initialization covariance condition; hidden width alone does not imply it. All construction constants are fixed before choosing τ. Fix the hidden depth L, input width $m_0 = d_\psi$, and hidden widths

$$\mathbf{m} = (m_1, \ldots, m_L).$$

For a branch parameter

$$\vartheta = (W^1, \ldots, W^L), \qquad W^\ell \in \mathbb{R}^{m_\ell \times m_{\ell-1}},$$

define

$$h_\vartheta^0(s,a) = \psi(s,a), \tag{77}$$

$$z_\vartheta^\ell(s,a) = \frac{1}{\sqrt{m_{\ell-1}}} W^\ell h_\vartheta^{\ell-1}(s,a), \tag{78}$$

$$h_\vartheta^\ell(s,a) = \sigma_\ell\big(z_\vartheta^\ell(s,a)\big), \qquad \ell = 1,\dots,L, \tag{79}$$

$$G_{\mathbf{m},\vartheta}(s,a) = \frac{1}{\sqrt{m_L}} b^\top h_\vartheta^L(s,a), \qquad b_j \in \{-1,+1\}. \tag{80}$$

The activations are applied coordinatewise and the readout b is fixed.

Equip the branch parameter space with the product Frobenius norm

$$\|\Delta\vartheta\|_{\mathrm{par}}^2 := \sum_{\ell=1}^{L} \big\|\Delta W^\ell\big\|_F^2 . \tag{81}$$

Let $B_{\mathbf{m}} : \mathbb{R}^r \to \prod_{\ell=1}^{L} \mathbb{R}^{m_\ell \times m_{\ell-1}}$ be an isometry:

$$B_{\mathbf{m}}^\top B_{\mathbf{m}} = I_r. \tag{82}$$

For an initialization $\vartheta_{\mathbf{m},0}$, define the two antithetic branches for every $\alpha \in \mathbb{R}^r$ by

$$\vartheta_{\mathbf{m}}^{\pm}(\alpha) = \vartheta_{\mathbf{m},0} \pm \frac{1}{\sqrt{2}} B_{\mathbf{m}}\alpha, \qquad \alpha \in \mathbb{R}^r. \tag{83}$$

Fix an ambient radius $A > 0$ and let $\overline{B}_A = \overline{B}(0,A)$. The geometry construction later selects $0 < R_{\mathrm{geom}} \le A/2$; write

$$\Theta_R := \overline{B}(0, R_{\mathrm{geom}}). \tag{84}$$

This closed ball is used only as a local geometry-certification tube. It is not the actor parameter space. The exact energy is

$$F_{\mathbf{m},\alpha}(s,a) = \frac{1}{\sqrt{2}} \Big[G_{\mathbf{m},\vartheta_{\mathbf{m}}^+(\alpha)}(s,a) - G_{\mathbf{m},\vartheta_{\mathbf{m}}^-(\alpha)}(s,a) \Big] . \tag{85}$$

This is an exact nonlinear network. It can be implemented as one wider block-diagonal network with a fixed antisymmetric readout. At initialization,

$$F_{\mathbf{m},0}(s,a) = 0, \qquad D_\alpha^2 F_{\mathbf{m},0}(s,a) = 0. \tag{86}$$

The reference $\pi_{\mathrm{ref}}(a \mid s) > 0$ is fixed and independent of x. The policy is its finite exponential tilt:

$$\frac{d\pi_{\mathbf{m},\alpha}(\cdot \mid s)}{d\pi_{\mathrm{ref}}(\cdot \mid s)}(a) = \exp\{F_{\mathbf{m},\alpha}(s,a) - A_{\mathbf{m},s}(\alpha)\}, \tag{87}$$

where

$$A_{\mathbf{m},s}(\alpha) = \log \sum_{a' \in \mathcal{A}} e^{F_{\mathbf{m},\alpha}(s,a')} \pi_{\mathrm{ref}}(a' \mid s). \tag{88}$$

Define the effective energy feature and policy score

$$Z_{\mathbf{m},\alpha}(s,a) := \nabla_\alpha F_{\mathbf{m},\alpha}(s,a), \tag{89}$$

$$\overline{Z}_{\mathbf{m},\alpha}(s,a) := Z_{\mathbf{m},\alpha}(s,a) - \mathbb{E}_{a' \sim \pi_{\mathbf{m},\alpha}(\cdot|s)}[Z_{\mathbf{m},\alpha}(s,a')] = \nabla_\alpha \log \pi_{\mathbf{m},\alpha}(a \mid s). \tag{90}$$

For a state law ν, define

$$\mathcal{F}_{\mathbf{m},\nu}(\alpha) := \mathbb{E}_{s\sim\nu,\, a\sim\pi_{\mathbf{m},\alpha}(\cdot|s)}[\overline{Z}_{\mathbf{m},\alpha}(s,a)\overline{Z}_{\mathbf{m},\alpha}(s,a)^\top]. \tag{91}$$

Write $\mathcal{F}_{x,\mathbf{m}}(\alpha)$ when $\nu = d_\alpha^\gamma$.

Construction requirements. The depth L, effective dimension r, and ambient radius A are independent of the hidden widths. The actor domain itself is $\mathbb{R}^r$. The input satisfies $\|\psi(s,a)\| \le B_\psi$. Each activation is bounded and belongs to $C^4(\mathbb{R})$, with

$$\left\|\sigma_\ell^{(k)}\right\|_\infty \le M_{\ell,k}, \qquad k = 0,1,2,3,4. \tag{92}$$

The initialized normalized layer operators satisfy

$$\frac{\left\|W^\ell_{\mathbf{m},0}\right\|_{\mathrm{op}}}{\sqrt{m_{\ell-1}}} \le \kappa_\ell, \qquad \ell = 1,\dots,L, \tag{93}$$

with width-independent constants. The chart map is an isometry as in (82). If it is selected from data, use data independent of the sample used to certify its covariance gap. Conditional analysis for a fixed accepted chart is also possible.

Chart spectral control and antithetic identities For a branch point $\vartheta^\pm_{\mathbf{m}}(\alpha)$, set

$$A^\ell_{\pm,\alpha} := \frac{1}{\sqrt{m_{\ell-1}}} W^\ell_{\pm,\alpha}. \tag{94}$$

Lemma C.1 (Uniform normalized spectral bound on the chart)**.** *Under Section C.1, on the ambient ball $\overline{B}_A$,*

$$\left\|A^\ell_{\pm,\alpha}\right\|_{\mathrm{op}} \le \kappa_\ell + \frac{A}{\sqrt{2m_{\ell-1}}} \le \bar\kappa_\ell := \kappa_\ell + \frac{A}{\sqrt{2}} \tag{95}$$

for every branch, chart point, and admissible width.

Proof. The layer-ℓ component of $B_{\mathbf{m}}\alpha$ has Frobenius norm at most $\|B_{\mathbf{m}}\alpha\|_{\mathrm{par}} = \|\alpha\| \le A$. Therefore its operator norm is at most A. Divide the triangle inequality for $W^\ell_{\mathbf{m},0} \pm 2^{-1/2}(B_{\mathbf{m}}\alpha)^\ell$ by $\sqrt{m_{\ell-1}}$ and use (93). □

Lemma C.2 (Antithetic initialization identities)**.** *The full two-branch chart is an isometry. Moreover,*

$$F_{\mathbf{m},0}(s,a) = 0, \tag{96}$$

$$Z_{\mathbf{m},0}(s,a) = B_{\mathbf{m}}^\top \nabla_\vartheta G_{\mathbf{m},\vartheta_{\mathbf{m},0}}(s,a). \tag{97}$$

More generally, for directions $u_1,\dots,u_k \in \mathbb{R}^r$,

$$D^k_\alpha F_{\mathbf{m},\alpha}[u_1,\dots,u_k] = 2^{-(k+1)/2}\Big(D^k_\vartheta G_{\vartheta^+(\alpha)}[B_{\mathbf{m}}u_1,\dots,B_{\mathbf{m}}u_k] - (-1)^k D^k_\vartheta G_{\vartheta^-(\alpha)}[B_{\mathbf{m}}u_1,\dots,B_{\mathbf{m}}u_k]\Big). \tag{98}$$

Proof. The map

$$u \longmapsto \left(2^{-1/2}B_{\mathbf{m}}u, -2^{-1/2}B_{\mathbf{m}}u\right)$$

has squared product norm

$$\frac{1}{2}\|B_{\mathbf{m}}u\|^2_{\mathrm{par}} + \frac{1}{2}\|B_{\mathbf{m}}u\|^2_{\mathrm{par}} = \|u\|^2.$$

The two branch values agree at $\alpha = 0$, proving (96). Differentiating once gives (97); repeated differentiation and the chain rule give (98). □

A width-independent mixed-derivative recursion For $\ell = 0,\dots,L$, define

$$q_0 := B_\psi/\sqrt{m_0}, \qquad q_\ell := M_{\ell,0} \quad (\ell \ge 1). \tag{99}$$

Set $H_{0,k} = 0$ for $k = 1,2,3,4$. Recursively, for $\ell = 1,\dots,L$, let

$$\widehat{H}_{\ell-1,0} := q_{\ell-1}, \qquad \widehat{H}_{\ell-1,j} := H_{\ell-1,j} \quad (j = 1,2,3), \tag{100}$$

and

$$Z_{\ell,k} := \bar\kappa_\ell H_{\ell-1,k} + k\widehat{H}_{\ell-1,k-1}, \qquad k = 1,2,3,4. \tag{101}$$

Define

$$H_{\ell,1} := M_{\ell,1} Z_{\ell,1}, \tag{102}$$

$$H_{\ell,2} := M_{\ell,1} Z_{\ell,2} + M_{\ell,2} Z_{\ell,1}^2, \tag{103}$$

$$H_{\ell,3} := M_{\ell,1} Z_{\ell,3} + 3M_{\ell,2} Z_{\ell,1} Z_{\ell,2} + M_{\ell,3} Z_{\ell,1}^3, \tag{104}$$

$$H_{\ell,4} := M_{\ell,1} Z_{\ell,4} + 4M_{\ell,2} Z_{\ell,1} Z_{\ell,3} + 3M_{\ell,2} Z_{\ell,2}^2 + 6M_{\ell,3} Z_{\ell,1}^2 Z_{\ell,2} + M_{\ell,4} Z_{\ell,1}^4. \tag{105}$$

All these numbers depend on fixed depth, the activation envelopes, the fixed ambient radius A, the input bound, and the normalized spectral envelopes, but not on the hidden widths.

Lemma C.3 (Exact mixed layer-derivative recursion)**.** *Fix a branch point and parameter directions $\Delta_1\vartheta, \dots, \Delta_k\vartheta$ with $\|\Delta_i\vartheta\|_{\mathrm{par}} \le 1$. For a nonempty subset $I \subseteq \{1, \dots, k\}$, write*

$$h_I^\ell = D_\vartheta^{|I|} h^\ell[\Delta_i\vartheta : i \in I],$$
$$z_I^\ell = D_\vartheta^{|I|} z^\ell[\Delta_i\vartheta : i \in I].$$

Then

$$z_I^\ell = A^\ell h_I^{\ell-1} + \frac{1}{\sqrt{m_{\ell-1}}} \sum_{i \in I} \Delta_i W^\ell h_{I\setminus\{i\}}^{\ell-1}, \tag{106}$$

where $h_\varnothing^{\ell-1} = h^{\ell-1}$. For $|I| = k \le 4$,

$$\left\|z_I^\ell\right\| \le Z_{\ell,k}, \qquad \left\|h_I^\ell\right\| \le H_{\ell,k}. \tag{107}$$

Proof. Because W^ℓ is affine in the parameter, at most one derivative can fall on it. Applying the product rule to $m_{\ell-1}^{-1/2} W^\ell h^{\ell-1}$ gives (106) exactly.

For the norm bound, first note that $\left\|\Delta_i W^\ell\right\|_{\mathrm{op}} \le \left\|\Delta_i W^\ell\right\|_F \le 1$. If $I \setminus \{i\} = \varnothing$, then

$$\frac{\left\|h^{\ell-1}\right\|}{\sqrt{m_{\ell-1}}} \le q_{\ell-1};$$

this follows from the input bound when $\ell = 1$ and from bounded activation coordinates when $\ell \ge 2$. If $I \setminus \{i\}$ is nonempty, induction gives $\left\|h_{I\setminus\{i\}}^{\ell-1}\right\| \le H_{\ell-1,k-1}$, and division by $\sqrt{m_{\ell-1}} \ge 1$ can only reduce the bound. Together with Lemma C.1, this proves $\left\|z_I^\ell\right\| \le Z_{\ell,k}$.

The coordinatewise Faà di Bruno formula is

$$h_I^\ell = \sum_{\mathcal{P} \in \mathfrak{P}(I)} \sigma_\ell^{(|\mathcal{P}|)}(z^\ell) \odot \prod_{B \in \mathcal{P}} z_B^\ell, \tag{108}$$

where $\mathfrak{P}(I)$ is the set of partitions of I and products are coordinatewise. For vectors $v_1, \dots, v_j$,

$$\|v_1 \odot \cdots \odot v_j\|_2 \le \prod_{i=1}^{j} \|v_i\|_2 . \tag{109}$$

The partitions of sets of sizes one through four have block-size multiplicities 1; $1, 1$; $1, 3, 1$; and $1, 4, 3, 6, 1$, respectively. Substituting the Z bounds into (108) gives exactly (102)–(105). Induction over ℓ proves (107). □

Theorem C.4 (Exact deep derivative bounds)**.** *Under Section C.1, define*

$$C_{k,L} := 2^{(1-k)/2} H_{L,k}, \qquad k = 1, 2, 3, 4. \tag{110}$$

Then, uniformly over all admissible hidden widths, points of $\overline{B}_A$, and state-action pairs,

$$\left\|D_\alpha^k F_{\mathbf{m},\alpha}(s,a)\right\|_{\mathrm{op}} \le C_{k,L}, \qquad k = 1, 2, 3, 4. \tag{111}$$

Consequently,

$$\|Z_{\mathbf{m},\alpha}(s,a)\| \le C_{1,L}, \tag{112}$$

$$\|Z_{\mathbf{m},\alpha}(s,a) - Z_{\mathbf{m},\beta}(s,a)\| \le C_{2,L} \|\alpha - \beta\| , \tag{113}$$

$$|F_{\mathbf{m},\alpha}(s,a) - F_{\mathbf{m},\beta}(s,a)| \le C_{1,L} \|\alpha - \beta\| . \tag{114}$$

Proof. Apply Lemma C.3 at either branch to unit product-norm directions. Since $\|b\| / \sqrt{m_L} = 1$,

$$\left|D_\vartheta^k G[\Delta_1\vartheta, \ldots, \Delta_k\vartheta]\right| \le H_{L,k}.$$

For unit $u_i \in \mathbb{R}^r$, the isometry gives $\|B_{\mathbf{m}} u_i\|_{\mathrm{par}} = 1$. Now use (98); the sum of the two branch bounds is $2 \cdot 2^{-(k+1)/2} H_{L,k} = C_{k,L}$. The first consequence is the case $k = 1$. The remaining two follow by integrating the second and first derivatives, respectively, along the segment from β to α. □

Corollary C.5 (Uniform energy and density envelope)**.** *For every point of the smaller geometry chart* $\Theta_R = \overline{B}_R$*,*

$$|F_{\mathbf{m},\alpha}(s,a)| \le C_{1,L} R, \tag{115}$$

$$e^{-2C_{1,L}R} \le \frac{d\pi_{\mathbf{m},\alpha}(\cdot \mid s)}{d\pi_{\mathrm{ref}}(\cdot \mid s)}(a) \le e^{2C_{1,L}R}, \tag{116}$$

$$\left\| \log \frac{d\pi_{\mathbf{m},\alpha}(\cdot \mid s)}{d\pi_{\mathrm{ref}}(\cdot \mid s)} \right\|_{\infty,a} \le 2C_{1,L} R. \tag{117}$$

Proof. Use $F_{\mathbf{m},0} = 0$ and (114). The log normalizer lies between the minimum and maximum of the energy, so both the numerator and normalizer comparisons contribute at most $C_{1,L}R$. □

Log-partition and log-policy derivatives For nonnegative numbers $c_1, \ldots, c_k$, define the complete exponential Bell polynomial

$$\mathsf{B}_k(c_1, \ldots, c_k) := \sum_{\substack{n_1,\ldots,n_k \ge 0 \\ \sum_{j=1}^k j n_j = k}} \frac{k!}{\prod_{j=1}^k n_j! (j!)^{n_j}} \prod_{j=1}^k c_j^{n_j}. \tag{118}$$

Set $\mathsf{B}_0 := 1$.

Lemma C.6 (Log-partition derivatives with explicit envelopes)**.** *Fix a state and a* C^4 *scalar path* $t \mapsto F_t(a)$*. Suppose* $\left\|\partial_t^j F_t\right\|_\infty \le c_j$ *for* $j = 1, 2, 3, 4$ *on an interval. Let*

$$Z(t) = \int e^{F_t(a)} \pi_{\mathrm{ref}}(da \mid s), \qquad A(t) = \log Z(t).$$

Define $b_j = \mathsf{B}_j(c_1, \ldots, c_j)$ *and recursively*

$$a_1 := b_1, \qquad a_k := b_k + \sum_{j=1}^{k-1} \binom{k-1}{j-1} a_j b_{k-j}, \quad k = 2, 3, 4. \tag{119}$$

Then $|A^{(k)}(t)| \le a_k$ *for* $k = 1, 2, 3, 4$*.*

Proof. Faà di Bruno's formula for $e^{F_t(a)}$ gives

$$\partial_t^k e^{F_t(a)} = e^{F_t(a)} \mathsf{B}_k(\partial_t F_t(a), \ldots, \partial_t^k F_t(a)).$$

After integration and division by $Z(t)$, $|Z^{(k)}(t)/Z(t)| \le b_k$. Differentiating the identity $Z' = A'Z$ $k-1$ additional times gives

$$\frac{Z^{(k)}}{Z} = \sum_{j=1}^{k} \binom{k-1}{j-1} A^{(j)} \frac{Z^{(k-j)}}{Z}. \tag{120}$$

Isolate the $j = k$ term and apply induction, the triangle inequality, and the bounds on $Z^{(q)}/Z$. This yields exactly (119). □

Lemma C.7 (Polarization bound)**.** *Let* T *be a continuous symmetric* k*-linear form on a real Hilbert space and let* $p_T(h) = T[h, \ldots, h]$*. Then*

$$\|T\|_{\mathrm{op}} \le \frac{k^k}{k!} \sup_{\|h\| \le 1} |p_T(h)|. \tag{121}$$

Proof. The real polarization identity is

$$T[h_1,\dots,h_k] = \frac{1}{2^k k!} \sum_{\epsilon\in\{-1,1\}^k} \left(\prod_{i=1}^k \epsilon_i\right) p_T\left(\sum_{i=1}^k \epsilon_i h_i\right).$$

For $\|h_i\| \le 1$, the argument of p_T has norm at most k. Homogeneity gives $|p_T(\sum_i \epsilon_i h_i)| \le k^k \sup_{\|h\|\le 1} |p_T(h)|$. Summing 2^k terms proves the claim. □

Lemma C.8 (Width-uniform deep log-policy derivatives)**.** *For $k = 1, 2, 3, 4$, let $b_{k,L}$ and $a_{k,L}$ be obtained from (118)–(119) with $c_j = C_{j,L}$, and define*

$$M_{\log,k,L} := \frac{k^k}{k!}\big(C_{k,L} + a_{k,L}\big). \tag{122}$$

Then, uniformly over widths, states, actions, and chart points,

$$\left\|D_\alpha^k \log \pi_{\mathbf{m},\alpha}(a \mid s)\right\|_{\mathrm{op}} \le M_{\log,k,L}, \qquad k = 1, 2, 3, 4. \tag{123}$$

For every fixed state law ν, the Fisher matrix is Lipschitz:

$$\|\mathcal{F}_{\mathbf{m},\nu}(\alpha) - \mathcal{F}_{\mathbf{m},\nu}(\beta)\|_{\mathrm{op}} \le L_{F,L} \|\alpha - \beta\|, \tag{124}$$

where

$$L_{F,L} := 2M_{\log,1,L} M_{\log,2,L} + M_{\log,1,L}^3. \tag{125}$$

Consequently, on the Fisher-conditioned region,

$$\left\|\mathcal{F}_{\mathbf{m},\nu}(\alpha)^{-1} - \mathcal{F}_{\mathbf{m},\nu}(\beta)^{-1}\right\|_{\mathrm{op}} \le L_{P,L} \|\alpha - \beta\|, \qquad L_{P,L} := \frac{L_{F,L}}{\underline{\lambda}_{F,L}^2}. \tag{126}$$

Proof. Fix a unit direction h and apply Lemma C.6 to $F_t(a) = F_{\mathbf{m},\alpha+th}(s,a)$. The deep derivative theorem gives $\left\|\partial_t^j F_t\right\|_\infty \le C_{j,L}$. Since $\log \pi_\alpha = F_\alpha - A_s(\alpha)$, the diagonal kth directional derivative has magnitude at most $C_{k,L} + a_{k,L}$. The derivative is a symmetric k-linear form, so Lemma C.7 gives (123).

Let $S_\alpha = \nabla_\alpha \log \pi_{\mathbf{m},\alpha}$. For a unit direction h, differentiation of a policy expectation gives

$$D_\alpha \mathbb{E}_{\nu,\pi_\alpha}[S_\alpha S_\alpha^\top][h] = \mathbb{E}_{\nu,\pi_\alpha}\big[(DS_\alpha[h])S_\alpha^\top + S_\alpha(DS_\alpha[h])^\top + S_\alpha S_\alpha^\top \langle S_\alpha, h\rangle\big].$$

The operator norm is at most $L_{F,L}$. Integrating this derivative along the segment from β to α proves (124). Finally use Lemma C.49 and the Fisher lower bound. □

Reference-KL envelopes. Use the fixed ambient actor ball. Put $\chi_\gamma = \gamma/(1-\gamma)$ and $v_0 = 2C_1 A$. Since $F_0 = 0$, $|F_\alpha| \le C_1 A$ and $|\log(\pi_\alpha/\pi_{\mathrm{ref}})| \le v_0$.

Let S_1, S_2, S_3 be uniform bounds on the first three derivatives of $\log \pi_\alpha$. They follow from differentiating the finite softmax and the energy bounds. One conservative choice is

$$S_1 = 2C_1, \qquad S_2 = 2C_2 + C_1^2, \qquad S_3 = 2C_3 + 6C_1C_2 + 8C_1^3.$$

Define

$$P_1 = S_1, \qquad P_2 = S_2 + S_1^2, \qquad P_3 = S_3 + 3S_1S_2 + S_1^3, \tag{127}$$

$$D_1 = \chi_\gamma P_1, \qquad D_2 = \chi_\gamma P_2 + 2\chi_\gamma^2 P_1^2, \qquad D_3 = \chi_\gamma P_3 + 6\chi_\gamma^2 P_1 P_2 + 6\chi_\gamma^3 P_1^3. \tag{128}$$

Here the norm for signed measures is the dual of the supremum norm, without a factor $1/2$.

Lemma C.9 (Occupancy and KL envelopes)**.** *The total-variation norms of the first three policy derivatives are at most P_i, and those of the normalized state occupancy derivatives are at most D_i. Let*

$$q_0 = v_0, \qquad q_1 = S_1 + P_1 v_0,$$
$$q_2 = S_2 + 2P_1S_1 + P_2 v_0, \qquad q_3 = S_3 + 3P_1S_2 + 3P_2S_1 + P_3 v_0.$$

Then

$$\left\|D_\alpha^3 \mathcal{K}\right\| \le L_K := q_3 + 3D_1 q_2 + 3D_2 q_1 + D_3 q_0. \tag{129}$$

In particular the KL Hessian is L_K-Lipschitz on actor segments in the ambient ball. All these constants are independent of τ, width, and target accuracy.

Proof. Differentiate $\pi = e^{\log \pi}$ and sum absolute values over actions to obtain P_i. The actor-induced Markov kernel T_α is a contraction in the supremum norm, and $\left\|D^i T_\alpha\right\| \le P_i$. The normalized occupancy is

$$d_\alpha = (1-\gamma)\rho(I - \gamma T_\alpha)^{-1}.$$

Differentiate its resolvent identity. The resolvent norm is at most $(1-\gamma)^{-1}$, while $\|d_\alpha\| = 1$. Product differentiation gives exactly the bounds D_i in (128); the coefficients count the ordered product-rule terms. The statewise KL is $k_\alpha(s) = \sum_a \pi_\alpha(a \mid s) \log(\pi_\alpha(a \mid s)/\pi_{\mathrm{ref}}(a \mid s))$. Product differentiation bounds its derivatives by q_i. Finally differentiate $\mathcal{K} = d_\alpha k_\alpha$ three times. The four groups of terms have bounds $q_3, 3D_1q_2, 3D_2q_1, D_3q_0$. All differentiations are justified by the finite action sums, derivative envelopes, and discounted resolvent. The argument does not require transition likelihood ratios for derivatives in α. □

Lemma C.10 (Exact KL Hessian at initialization)**.** *With $a_0 = (1-\gamma)\lambda_{\mathrm{eff},L}$, on the fixed actor chart,*

$$\mathcal{K}(0) = 0, \quad \nabla_\alpha \mathcal{K}(0) = 0, \quad \nabla^2_{\alpha\alpha}\mathcal{K}(0) = \mathcal{F}_{x,\mathbf{m}}(0) \succeq a_0 I, \tag{130}$$

where $\mathcal{F}_{x,\mathbf{m}}(0)$ is the occupancy-weighted policy Fisher matrix.

Proof. For every state s, $k_0(s) = 0$ and $Dk_0(s) = 0$. At zero the score is $Z_0(s,a) - \mathbb{E}_{\pi_{\mathrm{ref}}} Z_0(s,\cdot)$, so the information identity yields $D^2 k_0(s) = \mathrm{Cov}_{\pi_{\mathrm{ref}}}(Z_0(s,\cdot))$. Differentiating $\mathcal{K} = d_\alpha k_\alpha$ twice leaves only $d_0 D^2 k_0$ because the terms multiplying k_0 and Dk_0 vanish statewise. Finally $d_0^\gamma \succeq (1-\gamma)\rho$ and Assumption 4.2 give the lower bound. Normalization is essential: the bound contains $(1-\gamma)$, not its reciprocal. □

Lemma C.11 (Persistence of the restricted initialization covariance)**.** *Under Assumption 4.2 and Theorem C.4, for every $\alpha \in \Theta_R$,*

$$\mathbb{E}_{s\sim\rho}\,\mathrm{Cov}_{a\sim\pi_{\mathrm{ref}}(\cdot|s)}\big(Z_{\mathbf{m},\alpha}(s,a)\big) \succeq \big(\sqrt{\lambda_{\mathrm{eff},L}} - C_{2,L}\,\|\alpha\|\big)_+^2 I_r. \tag{131}$$

Proof. Fix a unit vector $v \in \mathbb{R}^r$ and let the joint reference law be $s \sim \rho$, $a \sim \pi_{\mathrm{ref}}(\cdot \mid s)$. Set $X = v^\top Z_{\mathbf{m},0}$ and $Y = v^\top Z_{\mathbf{m},\alpha}$. By (113),

$$\|X - Y\|_{L_2} \le C_{2,L}\,\|\alpha\|.$$

Conditional centering is a contraction, so

$$\begin{aligned}\|Y - \mathbb{E}[Y \mid s]\|_{L_2} &\ge \|X - \mathbb{E}[X \mid s]\|_{L_2} - \|X - Y\|_{L_2} \\ &\ge \sqrt{\lambda_{\mathrm{eff},L}} - C_{2,L}\,\|\alpha\|,\end{aligned}$$

where the final inequality is (12). Since the left-hand side is nonnegative, take the positive part of the right-hand side before squaring, and then take the infimum over unit v. This proves the claim. □

Lemma C.12 (Persistent discounted and common-state Fisher excitation)**.** *For the radius in (33), for every $\alpha \in \Theta_R$ and every state law $\nu \succeq (1-\gamma)\rho$,*

$$\underline{\lambda}_{F,L} I_r \preceq \mathcal{F}_{\mathbf{m},\nu}(\alpha) \preceq \overline{\lambda}_{F,L} I_r, \tag{132}$$

where

$$\underline{\lambda}_{F,L} = (1-\gamma)e^{-2C_{1,L}R}\lambda_{\mathrm{eff,L}}/4, \qquad \overline{\lambda}_{F,L} = 4C_{1,L}^2. \tag{133}$$

In particular, this holds for each discounted occupancy and for every common mixture law (154).

Proof. The upper bound follows from $\left\|\overline{Z}_{\mathbf{m},\alpha}\right\| \le 2C_{1,L}$. For the lower bound, Corollary C.5 gives, conditionally on each state,

$$\pi_{\mathbf{m},\alpha}(\cdot \mid s) \succeq e^{-2C_{1,L}R}\pi_{\mathrm{ref}}(\cdot \mid s).$$

Apply Lemma C.47 conditionally in the action variable to $Z_{\mathbf{m},\alpha}$, integrate with respect to $\nu \succeq (1-\gamma)\rho$, and then use Lemma C.11. For every unit v,

$$\begin{aligned}v^\top \mathcal{F}_{\mathbf{m},\nu}(\alpha) v &\ge (1-\gamma)e^{-2C_{1,L}R}\mathbb{E}_{s\sim\rho}\,\mathrm{Var}_{a\sim\pi_{\mathrm{ref}}}\big(v^\top Z_{\mathbf{m},\alpha}(s,a)\big) \\ &\ge \underline{\lambda}_{F,L}.\end{aligned}$$

The common mixture dominates $(1-\gamma)\rho$ because each of its two components does. □

Optional one-time initialization certificate.

Proposition C.13 (Independent pilot certification). *Fix a realized network and chart with* $\|Z_{\mathbf{m},0}\| \le G_Z := C_{1,L}$. *Draw independent pilot triples*

$$s_i \sim \rho, \qquad a_i, a_i' \overset{\text{i.i.d.}}{\sim} \pi_{\text{ref}}(\cdot \mid s_i), \qquad i = 1, \dots, N,$$

and define

$$\widehat{\mathcal{G}}_N = \frac{1}{2N} \sum_{i=1}^{N} (Z_i - Z_i')(Z_i - Z_i')^\top, \tag{134}$$

where $Z_i = Z_{\mathbf{m},0}(s_i, a_i)$ *and* $Z_i' = Z_{\mathbf{m},0}(s_i, a_i')$. *Then*

$$\mathbb{E}[\widehat{\mathcal{G}}_N \mid \mathbf{m}, \textit{initialization}] = \mathcal{G}_{\mathbf{m},0}. \tag{135}$$

Moreover, there is a universal constant C_{B} *such that, with probability at least* $1 - \delta$,

$$\left\| \widehat{\mathcal{G}}_N - \mathcal{G}_{\mathbf{m},0} \right\|_{\text{op}} \le C_{\mathrm{B}} G_Z^2 \left[\sqrt{\frac{\log(2r/\delta)}{N}} + \frac{\log(2r/\delta)}{N} \right] =: \epsilon_N. \tag{136}$$

Consequently,

$$\lambda_{\min}(\widehat{\mathcal{G}}_N) \ge \lambda_{\text{cert}} + \epsilon_N \tag{137}$$

certifies $\mathcal{G}_{\mathbf{m},0} \succeq \lambda_{\text{cert}} I_r$ *with probability at least* $1 - \delta$.

Proof. Conditionally on s_i, the two actions are independent and identically distributed, so

$$\frac{1}{2}\mathbb{E}[(Z_i - Z_i')(Z_i - Z_i')^\top \mid s_i] = \mathrm{Cov}_{a \sim \pi_{\text{ref}}(\cdot|s_i)}(Z_{\mathbf{m},0}(s_i, a)).$$

Averaging over s_i proves (135). Each summand in (134) is positive semidefinite and has operator norm at most $2G_Z^2$. Self-adjoint matrix Bernstein gives (136) after changing a universal numerical constant. Weyl's inequality then proves the certification statement. □

The chart must be fixed independently of the certification sample. Chart selection from data therefore requires a separate certificate split. The proposition bounds false acceptance by δ; it does not assert that every chart is accepted. For a fixed positive gap and slack, the sample size grows only logarithmically in $1/\delta$. Charge any collected reset/pilot data in (75). A2 can instead be assumed directly. The actor Fisher certificate is distinct from C1's critic feature Gram gap.

C.2 Preference losses and global mixed-derivative bounds

Bradley–Terry losses satisfy explicit global envelopes For two trajectories and a realized binary label Y, put

$$\Delta_x = \sum_{t=0}^{K-1} r_x(s_t^1, a_t^1) - \sum_{t=0}^{K-1} r_x(s_t^2, a_t^2), \qquad L = \log(1 + e^{\Delta_x}) - Y\Delta_x. \tag{138}$$

The conditional law of Y given the trajectories is fixed when differentiating. In particular, it is not a reward-model label distribution whose parameter dependence has been omitted.

Proposition C.14 (Samplewise preference-loss envelopes). *If* $|r_x| \le B_r$, $\|\nabla_x r_x\| \le G_r$, *and* $\left\|\nabla_x^2 r_x\right\|_{\text{op}} \le L_r$ *uniformly, then* (138) *obeys*

$$\begin{gathered} 0 \le L \le B_L := \log 2 + 2KB_r, \qquad \|\nabla_x L\| \le G_L := 2KG_r, \\ \left\|\nabla_x^2 L\right\|_{\text{op}} \le L_L := K^2 G_r^2 + 2KL_r. \end{gathered} \tag{139}$$

These bounds hold for each realized label, and $\Phi_\tau \ge 0$.

Proof. The corresponding bounds for Δ_x, its gradient, and its Hessian are $2KB_r$, $2KG_r$, and $2KL_r$. For $\sigma(z) = (1 + e^{-z})^{-1}$,

$$\begin{aligned} \nabla_x L &= (\sigma(\Delta_x) - Y)\nabla_x \Delta_x, \\ \nabla_x^2 L &= \sigma'(\Delta_x)\nabla_x \Delta_x \nabla_x \Delta_x^\top + (\sigma(\Delta_x) - Y)\nabla_x^2 \Delta_x. \end{aligned}$$

Use $|\sigma - Y| \le 1$, $0 \le \sigma' \le 1/4$, and $0 \le L \le \log 2 + |\Delta_x|$. Taking the policy expectation preserves nonnegativity. □

For example, $r_x(s,a) = B\tanh(x^\top \varphi(s,a))$ with $\sup_{s,a}\|\varphi(s,a)\| \le C_\varphi$ satisfies the assumptions on unrestricted x, with $B_r = B$, $G_r = BC_\varphi$, and $L_r = 2BC_\varphi^2$. Its higher derivatives also exist and are locally uniformly bounded. Bounded output alone does not give the derivative envelopes for a generic reward network with every weight trainable. For a fixed offline comparison dataset independent of α, this preference term instead has $f_\alpha = 0$; the policy-generated objective in the theorem retains the response-chain contribution.

The mixed derivatives used by the proof Let M_q be the width-independent actor log-policy derivative bounds from Lemma C.8, for $q = 1, \dots, 4$, on the ambient actor neighborhood. Define the likelihood-derivative polynomials

$$\begin{aligned} b_0(n) &= 1, \qquad b_1(n) = nM_1, \\ b_2(n) &= nM_2 + n^2M_1^2, \\ b_3(n) &= nM_3 + 3n^2M_1M_2 + n^3M_1^3, \\ b_4(n) &= nM_4 + 4n^2M_1M_3 + 3n^2M_2^2 + 6n^3M_1^2M_2 + n^4M_1^4. \end{aligned}$$

Set $A_q(\gamma) = (1-\gamma)\sum_{t\ge 0}\gamma^t b_q(t+1)$, $(R_0, R_1, R_2) = (B_r, G_r, L_r)$, and $(F_0, F_1, F_2) = (B_L, G_L, L_L)$.

Lemma C.15 (Uniform mixed-derivative bounds on unrestricted x)**.** *For $0 \le p \le 2$ and $0 \le q \le 4$, the separated mixed derivatives exist and satisfy*

$$\|D_x^p D_\alpha^q j(x,\alpha)\|_{\mathrm{op}} \le R_p A_q(\gamma), \tag{140}$$

$$\|D_x^p D_\alpha^q f(x,\alpha)\|_{\mathrm{op}} \le F_p b_q(IK), \tag{141}$$

uniformly in x and width. For $p \ge 1$, the corresponding derivatives of $g_\tau = -j + \tau\mathcal{K}$ equal those of $-j$; for $p = 0$, add τ times the pure actor KL envelope. In particular, with $M_0 = 2C_{1,L}A$ and $M_i = M_{\log,i,L}$ for $i \ge 1$,

$$\|D_\alpha^q \mathcal{K}(\alpha)\|_{\mathrm{op}} \le (1-\gamma)\sum_{t\ge 0}\gamma^t \sum_{j=0}^{q}\binom{q}{j} b_j(t+1)M_{q-j} < \infty, \quad 0 \le q \le 4. \tag{142}$$

No global bounds on pure third or fourth x derivatives are claimed.

Proof. The trajectory likelihood has log likelihood equal to a sum of n log-policy terms plus a parameter-independent term. Differentiating its exponential q times gives sums over set partitions; taking operator norms yields exactly $b_q(n)$ as a bound on the likelihood derivative divided by the likelihood. Fixed dynamics imply that x derivatives act only on r_x or L, while actor derivatives act only on likelihoods. The feedback kernel has no derivative. For the reward at time t there are $t+1$ action factors, and the upper tuple has IK action factors. The global bounds give domination, and each $A_q(\gamma)$ is finite because a geometric sequence times a fixed-degree polynomial is summable. This justifies differentiation under the expectations and the discounted sum. The same separated calculation gives mixed derivatives with $p + q > 4$ when needed, using only p primitive x derivatives and q actor derivatives; it does not assume joint global C^6 bounds. Finally, $\mathcal{K}$ is independent of x. For its time-t contribution, apply the product rule to the prefix likelihood and the terminal log-policy-ratio criterion. Assigning j derivatives to the likelihood gives $b_j(t+1)$; the remaining derivatives of that criterion are bounded by M_{q-j}. Summing the $\binom{q}{j}$ assignments and the discounted series proves (142), including the fourth actor derivative used by the tracking Hessian. □

For clarity, the largest derivative blocks used in the argument are listed below; their lower-order counterparts are also used.

Proof component	Derivative blocks
Curvature and responses	$g_{\alpha\alpha}, g_{\alpha\alpha\alpha}, g_{\alpha x}, f_\alpha, f_{\alpha\alpha}$
Target smoothness	$g_{\alpha xx}, g_{\alpha\alpha x}, g_{\alpha\alpha\alpha}$; all second derivatives of f
Response secant	$g_{x\alpha}, g_{x\alpha\alpha}, g_{\alpha\alpha\alpha}, f_{x\alpha}, f_{\alpha\alpha}, f_{\alpha\alpha\alpha}$
Full tracking Hessian	$g_{\alpha\alpha xx}, g_{\alpha\alpha\alpha x}, g_{\alpha\alpha\alpha\alpha}, f_{\alpha xx}, f_{\alpha\alpha x}, f_{\alpha\alpha\alpha}$
Common-KL update	Actor log-policy and KL derivatives through order four

A1's compact-set C^4 statement is not invoked as a global bound. The needed discounted derivative tails are bounded by $(1-\gamma)\sum_{t\geq H}\gamma^t R_p b_q(t+1)$. They therefore admit the logarithmic truncation schedule. The length K of the preference objective is fixed by its definition and introduces no discounted truncation bias.

C.3 Response and tracking-potential calculus

This subsection supplies the local-root and gap calculations used in Step B.1. Its upper-domain bounds use only the separated derivative blocks in Lemma C.15.

Lemma C.16 (Integral secants are uniform in the perturbation)**.** *On any segment in the analytic actor neighborhood,*

$$\frac{G(x,U+cS)-G(x,U)}{c}=\int_0^1 D_\alpha G(x,U+tcS)[S]\,dt. \tag{143}$$

If $D_\alpha G$ is bounded and Lipschitz in the variables being compared, then the secant is Lipschitz in those variables and in bounded S, uniformly for $0<c\leq c_0$. Its continuous value at zero is $D_\alpha G(x,U)[S]$. A Lipschitz constant L_G for $D_\alpha G$ also gives error at most $L_G c\|S\|^2/2$ from this zero-perturbation value.

Proof. Apply the fundamental theorem of calculus to $t\mapsto G(x,U+tcS)$. For two triples, subtract the integral expressions and add and subtract the derivative at one common point. Its bounded norm controls $S-S'$; its Lipschitz bound controls the remaining difference, using $\|U+tcS-U'-tcS'\|\leq \|U-U'\|+c_0\|S-S'\|$. Subtracting $D_\alpha G(x,U)[S]$ and integrating t proves the last bound. □

Lemma C.17 (Uniformly Lipschitz population residual)**.** *On the prospective actor neighborhood, R_c is globally Lipschitz in x and locally uniformly Lipschitz in (U,S), with constants independent of c and width. At the selected response pair,*

$$R_c(x,\alpha_0^*(x),S_c^*(x))=0. \tag{144}$$

Proof. Use Lemma C.16 for $G=g_\alpha$ and add $f_\alpha(x,U+cS)$. The derivative blocks $g_{\alpha\alpha x}$, $g_{\alpha\alpha\alpha}$, $f_{\alpha x}$ and $f_{\alpha\alpha}$ are uniformly bounded by Lemma C.15. Subtracting the base and perturbed response equations gives (144). The segment between any two admissible U values preserves the proof-ball inclusion by convexity. □

The scaled potential from (40) has the exact integral form

$$\begin{aligned}\mathcal{H}_c(x,U,S)=&\int_0^1(1-t)S^\top g_{\alpha\alpha}(x,U+tcS)S\,dt\\&+\int_0^1\langle f_\alpha(x,U+tcS),S\rangle\,dt.\end{aligned} \tag{145}$$

This formula defines its continuous extension at $c=0$ without a divergent additive constant.

Proposition C.18 (Interior root of the scaled residual)**.** *If $U+c\overline{B}_s\subset B_R$ and $G_f/\mu_\tau<s$, then $\mathcal{H}_c(x,U,\cdot)$ has a unique interior minimizer $\bar{S}_c(x,U)\in B_s$, satisfying*

$$R_c(x,U,\bar{S}_c)=0,\qquad \|\bar{S}_c(x,U)\|\leq G_f/\mu_\tau,\qquad \bar{S}_c(x,\alpha_0^*(x))=S_c^*(x). \tag{146}$$

These statements hold uniformly in x and extend to $c=0$.

Proof. Direct differentiation gives

$$\nabla_S\mathcal{H}_c=R_c,\qquad D_S R_c=h_{c,\alpha\alpha}(x,U+cS)\succeq\mu_\tau I. \tag{147}$$

At $S=0$, $R_c=f_\alpha(x,U)$. On the boundary $\|S\|=s$, $\langle R_c,S\rangle\geq\mu_\tau s^2-G_f s>0$. Thus a minimizer on the compact ball cannot lie on its boundary. Strong convexity gives uniqueness; comparing its zero gradient with $R_c(x,U,0)$ gives the norm bound. Equation (144) and uniqueness identify the response-pair value. All arguments follow from the continuous integral formula at $c=0$ as well. □

Lemma C.19 (Global gap regularity and self-bounds)**.** *Fix sufficiently large finite τ and any fixed $\lambda_S > 0$. On the prospective actor neighborhood, $\bar{S}_c$ is continuously differentiable, and $\mathcal{T}$ is twice continuously differentiable. The following bounds are uniform in x, c and width:*

$$\tfrac{1}{2}\mu_\tau\|U - \alpha_0^*\|^2 \le \Delta_0, \qquad \tfrac{1}{2}\mu_\tau\|S - \bar{S}_c\|^2 \le \Gamma_c, \tag{148}$$

$$\|a\|^2 \ge 2\mu_\tau\Delta_0, \qquad \|R_c\|^2 \ge 2\mu_\tau\Gamma_c, \qquad \|a\|^2 + \|R_c\|^2 \le C\mathcal{T}, \tag{149}$$

$$\|\nabla_{(x,U,S)}\mathcal{T}\|^2 \le C\mathcal{T}, \qquad \|D^2_{(x,U,S)}\mathcal{T}\|_{\mathrm{op}} \le C. \tag{150}$$

The gap values and their first derivatives are bounded on this neighborhood. These constants may depend on the fixed τ.

Proof. Actor Hessian upper bounds and Lemma C.50 give the distance and gradient inequalities for both gaps. The implicit-function theorem applied to $R_c(x, U, \bar{S}_c) = 0$ gives

$$D_y\bar{S}_c = -(D_S R_c)^{-1} D_y R_c, \qquad y = (x, U).$$

The inverse norm is at most μ_τ^{-1}. Two derivatives of (145) use only $g_{\alpha\alpha xx}, g_{\alpha\alpha\alpha x}, g_{\alpha\alpha\alpha\alpha}$ and $f_{\alpha xx}, f_{\alpha\alpha x}, f_{\alpha\alpha\alpha}$, together with lower-order blocks. Lemma C.15 supplies uniform bounds for exactly these derivatives. Hence $D_y\bar{S}_c$ and the full Hessian of $\mathcal{H}_c$ are bounded.

For $y = x, U$, the envelope identities are

$$D_y\Gamma_c = D_y\mathcal{H}_c(x, U, S) - D_y\mathcal{H}_c(x, U, \bar{S}_c), \qquad D_x\Delta_0 = g_x(x, U) - g_x(x, \alpha_0^*).$$

The first difference has norm at most $C\|S - \bar{S}_c\|$; the second at most $C\|U - \alpha_0^*\|$. Together with $D_U\Delta_0 = a$, $D_S\Gamma_c = R_c$ and the gap inequalities, these prove the gradient self-bound. Differentiating the envelope identities once uses only $D^2\mathcal{H}_c$, $D\bar{S}_c$ and $D_x\alpha_0^*$, whose bound is in Corollary B.2. This proves the full Hessian bound without differentiating the minimizing separation twice. Upper Hessian bounds times squared actor distances bound the gap values. No pure third or fourth global x derivative is used. □

Lemma C.20 (Regularization-uniform weight, distance control and initialization)**.** *Fix the actor and separation neighborhoods before τ, and let $p_- = \overline{\lambda}_{F,L}^{-1}$, $p_+ = \underline{\lambda}_{F,L}^{-1}$ and $K_{\max} = L_{P,L}s$. There are $d_*, E_* > 0$ and a choice $\lambda_S > 0$, independent of sufficiently large τ, such that*

$$\|U - \alpha_0^*(x)\|^2 + \|S - S_c^*(x)\|^2 \le \frac{E_*}{\mu_\tau}\mathcal{T}(x, U, S), \qquad \mathcal{T}(x, 0, 0) \le \frac{J_1^2 + \lambda_S G_f^2}{2\mu_\tau}. \tag{151}$$

One can also require $\|D_U\bar{S}_c\| \le d_*$ and $\|\nabla_U\Gamma_c\| \le d_*\|R_c\|$ on this neighborhood.

Proof. Choose a fixed $\tau_0 \ge 1$ above the curvature threshold in (34). Since $D^3_\alpha\mathcal{K}$ is bounded by L_K, differentiating the residual integral gives

$$\|D_U R_c\| \le (J_3 + \tau L_K)s + L_f \le D_*\tau, \qquad D_* = (J_3/\tau_0 + L_K)s + L_f/\tau_0.$$

Set $d_* = D_*/\kappa$. Implicit differentiation bounds $D_U\bar{S}_c$ by d_*. The envelope identity and strong monotonicity imply $\|\nabla_U\Gamma_c\| \le D_*\tau\|S - \bar{S}_c\| \le d_*\|R_c\|$. Choose

$$0 < \lambda_S \le \min\{1, p_-^2/[4(p_+d_* + K_{\max})^2]\}. \tag{152}$$

Omit the second restriction if its denominator is zero. This choice is independent of τ and makes the population direction descend as proved in Lemma C.27.

Integrating $D_U\bar{S}_c$ on the segment from U to α_0^* gives $\|\bar{S}_c(x, U) - S_c^*(x)\| \le d_*\|U - \alpha_0^*\|$. The proof-ball inclusion holds along that segment by convexity. The triangle inequality and (148) yield the first bound with $E_* = \max\{2(1 + 2d_*^2), 4/\lambda_S\}$. At $U = S = 0$, the KL gradient vanishes, so $\|a\| \le J_1$ and $\|R_c\| = \|f_\alpha(x, 0)\| \le G_f$. Apply the residual-to-gap inequalities in (149) to prove the initialization bound. This argument holds for every x and requires no warm start. □

Lemma C.21 (Tracked corrected-field error)**.** *Let $V_c^*(x) = V_c(x, \alpha_0^*(x), S_c^*(x))$. Then*

$$\|V_c(p) - \nabla\Phi(x)\|^2 \le C(\mathcal{T}(p) + c^2), \qquad \|V_c(p)\|^2 \le C(\|\nabla\Phi(x)\|^2 + \mathcal{T}(p) + c^2). \tag{153}$$

Proof. Lemma B.3 gives the Lipschitz bound for V_c in the actor pair and $\|V_c^* - \nabla\Phi\| \le C_{\mathrm{sec}}c$. Combining it with (151) gives the first inequality. Adding $\nabla\Phi$ and using the squared triangle inequality gives the second. □

C.4 COMMON-KL GEOMETRY AND FINITE LOCAL UPDATES

Both actors use the same stored empirical state measure $\overline{d}_t = \widehat{\rho}_n$ from Definition C.28. Its observable covariance certificate provides the same Fisher constants as the population law

$$\nu_t^{\mathrm{pop}} = \tfrac{1}{2} d_{U_t}^{\gamma} + \tfrac{1}{2} d_{U_t + cS_t}^{\gamma}. \tag{154}$$

The latter explains the original constants but is not evaluated by the algorithm. The metric data are drawn once, independently of the actor initialization and all subsequent training blocks, and their cost is included in (75). A rejected covariance certificate returns x_0. Conditional on acceptance, the following analytic reference maps apply to the actual empirical metric. Their finite implementation and its error propagation are proved below.

For a probability state law ν and actor points α, b, define

$$\mathcal{K}_{\mathbf{m},\nu}(\alpha \mid b) = \mathbb{E}_{s\sim\nu} D_{\mathrm{KL}}(\pi_{\mathbf{m},b}(\cdot \mid s) \| \pi_{\mathbf{m},\alpha}(\cdot \mid s)), \tag{155}$$

and the exact local map

$$\mathsf{M}_{\mathbf{m},\nu,\eta}(b, v) = \operatorname*{arg\,min}_{\alpha \in \overline{B}(b, r_{\mathrm{prox}})} \{\langle v, \alpha - b\rangle + \eta^{-1} \mathcal{K}_{\mathbf{m},\nu}(\alpha \mid b)\}. \tag{156}$$

Choose

$$0 < r_{\mathrm{prox}} \le \min\{R/8, \underline{\lambda}_{F,L}/(2M_{\log,3,L})\}, \tag{157}$$

omitting a zero-denominator restriction. If $G_{\mathrm{prox}}(H) \ge 1$ bounds both sampled actor directions, impose the deterministic local-step restriction

$$\eta_y G_{\mathrm{prox}}(H) \le m_{K,L} r_{\mathrm{prox}}/2, \qquad m_{K,L} = \underline{\lambda}_{F,L}/2. \tag{158}$$

The sampled bounds in Appendix C.6 make $G_{\mathrm{prox}}(H)$ polynomial in H. This condition is satisfied by the headline schedule for sufficiently small accuracy. The exact map is used only as a mathematical reference. Algorithm 1 returns the residual-controlled finite approximation in Definition C.32; its error is incorporated below. For brevity, write

$$k_{\mathbf{m},\nu}(\alpha, b) := \nabla_\alpha \mathcal{K}_{\mathbf{m},\nu}(\alpha \mid b). \tag{159}$$

Lemma C.22 (Derivatives of the oriented common KL)**.** *For every state law ν,*

$$k_{\mathbf{m},\nu}(\alpha, b) = -\mathbb{E}_{\nu,\pi_{\mathbf{m},b}}[\nabla_\alpha \log \pi_{\mathbf{m},\alpha}(a \mid s)], \tag{160}$$

$$k_{\mathbf{m},\nu}(b, b) = 0, \tag{161}$$

$$D_1 k_{\mathbf{m},\nu}(b, b) = \mathcal{F}_{\mathbf{m},\nu}(b), \tag{162}$$

$$D_2 k_{\mathbf{m},\nu}(b, b) = -\mathcal{F}_{\mathbf{m},\nu}(b). \tag{163}$$

On the fixed ambient actor ball, all mixed derivatives of $k_{\mathbf{m},\nu}$ of total order at most three are bounded uniformly in the widths and in the probability law ν.

Proof. Differentiate the first argument of the oriented KL under the integral to obtain (160). The conditional score identity $\mathbb{E}_{\pi_b} \nabla_b \log \pi_b = 0$ proves (161). Differentiating this zero-score identity in a direction h gives

$$\mathbb{E}_{\pi_b}[D^2 \log \pi_b[h] + \nabla \log \pi_b \, \langle \nabla \log \pi_b, h\rangle] = 0.$$

Hence $-\mathbb{E} D^2 \log \pi_b$ equals the Fisher matrix, proving (162). Differentiating the expectation in (160) with respect to the sampling parameter b gives the negative Fisher matrix at the diagonal, proving (163). Each further derivative is a finite sum of products of log-policy derivatives and score factors. Lemma C.8 bounds all terms through the required order. □

Lemma C.23 (Uniform local KL conditioning)**.** *Let ν be a probability law such that $\underline{\lambda}_{F,L} I_r \preceq \mathcal{F}_{\mathbf{m},\nu}(b) \preceq \overline{\lambda}_{F,L} I_r$, let $b \in \Theta_R$, and suppose the segment from b to α lies in the local geometry tube with $\|\alpha - b\| \le r_{\mathrm{prox}}$. Then*

$$m_{K,L} I_r \preceq \nabla^2_{\alpha\alpha} \mathcal{K}_{\mathbf{m},\nu}(\alpha \mid b) \preceq M_{K,L} I_r, \tag{164}$$

where

$$m_{K,L} := \frac{1}{2}\underline{\lambda}_{F,L}, \qquad M_{K,L} := \overline{\lambda}_{F,L} + M_{\log,3,L} r_{\mathrm{prox}}. \tag{165}$$

Consequently,

$$\frac{m_{K,L}}{2}\|\alpha-b\|^2 \le \mathcal{K}_{\mathbf{m},\nu}(\alpha \mid b) \le \frac{M_{K,L}}{2}\|\alpha-b\|^2. \tag{166}$$

Moreover,

$$\left|\mathcal{K}_{\mathbf{m},\nu}(b+\Delta \mid b) - \frac{1}{2}\langle \Delta, \mathcal{F}_{\mathbf{m},\nu}(b)\Delta\rangle\right| \le \frac{M_{\log,3,L}}{6}\|\Delta\|^3. \tag{167}$$

Proof. By Lemma C.22, the first-argument Hessian at the diagonal is the Fisher matrix. Since

$$\nabla^2_{\alpha\alpha}\mathcal{K}(\alpha \mid b) = -\mathbb{E}_{\nu,\pi_b} D^2_\alpha \log \pi_\alpha,$$

its derivative with respect to α has norm at most $M_{\log,3,L}$. Therefore

$$\left\|\nabla^2_{\alpha\alpha}\mathcal{K}(\alpha \mid b) - \mathcal{F}_{\mathbf{m},\nu}(b)\right\|_{\mathrm{op}} \le M_{\log,3,L}\|\alpha-b\|.$$

Use the stated Fisher bounds and (157) to obtain (164). The implemented empirical metric satisfies these bounds by Proposition C.30; its proof does not assume domination by a population state law. Since the KL and its first derivative vanish at the diagonal, integrating the Hessian twice along the segment gives (166). The integral third-order Taylor remainder gives (167). □

Lemma C.24 (Exact local proximal map and quantitative remainder)**.** *Let ν satisfy the Fisher premise of Lemma C.23, let b lie at distance at least r_{prox} from $\partial\Theta_R$, and let $\|v\| \le G_{\mathrm{prox}}(H)$. Under (158), the problem (156) has a unique interior minimizer $M = \mathsf{M}_{\mathbf{m},\nu,\eta}(b,v)$ satisfying*

$$k_{\mathbf{m},\nu}(M,b) + \eta v = 0, \qquad \|M-b\| \le \frac{\eta}{m_{K,L}}\|v\| \le \frac{1}{2}r_{\mathrm{prox}}. \tag{168}$$

Let

$$P_{\mathbf{m},\nu}(b) := \mathcal{F}_{\mathbf{m},\nu}(b)^{-1}. \tag{169}$$

There is a C^1 remainder $Q_{\mathbf{m},\nu}$ such that

$$\mathsf{M}_{\mathbf{m},\nu,\eta}(b,v) = b - \eta P_{\mathbf{m},\nu}(b)v + \eta^2 Q_{\mathbf{m},\nu}(b,v,\eta), \tag{170}$$

and

$$\|Q_{\mathbf{m},\nu}(b,v,\eta)\| \le C_{Q,L}\|v\|^2, \tag{171}$$

$$\|D_b Q_{\mathbf{m},\nu}(b,v,\eta)\|_{\mathrm{op}} \le C_{Q,L}\|v\|^2, \tag{172}$$

$$\|D_v Q_{\mathbf{m},\nu}(b,v,\eta)\|_{\mathrm{op}} \le C_{Q,L}\|v\|. \tag{173}$$

The constant is uniform in the widths, ν, and sufficiently small η; it is also independent of H under (158).

Proof. Strong convexity of the local objective gives uniqueness. For a constrained minimizer M, the variational inequality with comparison point b gives

$$\langle k(M,b) + \eta v, b - M\rangle \ge 0.$$

Since $k(b,b) = 0$ and $k(\cdot,b)$ is $m_{K,L}$-strongly monotone,

$$m_{K,L}\|M-b\|^2 \le \langle k(M,b), M-b\rangle \le -\eta\langle v, M-b\rangle \le \eta\|v\|\,\|M-b\|.$$

This proves the displacement bound. The tube and stepsize conditions make the solution strictly interior, so the first-order equation holds.

Write $M = b + \eta d$. For $\eta \ge 0$, define the continuously extended map

$$\Psi(d,b,v,\eta) := \int_0^1 D_1 k(b + t\eta d, b)[d]\,dt + v. \tag{174}$$

For $\eta > 0$, the fundamental theorem of calculus gives the exact identity

$$\Psi(d,b,v,\eta) = \eta^{-1}k(b+\eta d, b) + v, \tag{175}$$

while at $\eta = 0$ it becomes $\mathcal{F}_{\mathbf{m},\nu}(b)d + v = 0$. Differentiating (175) with respect to d gives, for $\eta > 0$,

$$D_d\Psi(d, b, v, \eta) = D_1 k(b + \eta d, b), \tag{176}$$

and the same formula holds at $\eta = 0$ by continuity with value $\mathcal{F}_{\mathbf{m},\nu}(b)$. Hence Lemma C.23 gives $D_d\Psi \succeq m_{K,L} I_r$ throughout the local tube. Since the derivatives of k through order three are uniformly bounded, the quantitative implicit-function theorem gives a uniformly C^2 solution $d = d(b, v, \eta)$ with $d(b, v, 0) = -P_{\mathbf{m},\nu}(b)v$ and $\|d\| \le m_{K,L}^{-1} \|v\|$.

Differentiate (174) with respect to η. Every term in $\partial_\eta\Psi$ contains two factors of d, so $\|\partial_\eta\Psi\| \le C_L \|v\|^2$. Uniform invertibility of $D_d\Psi$ gives $\|\partial_\eta d\| \le C_L \|v\|^2$. Differentiating this identity in b and v and using the bounded third derivatives of k gives

$$\|D_b\partial_\eta d\| \le C_L \|v\|^2, \qquad \|D_v\partial_\eta d\| \le C_L \|v\|.$$

Finally,

$$d(b, v, \eta) + P(b)v = \eta \int_0^1 \partial_\eta d(b, v, t\eta)dt.$$

Define Q by the integral on the right. Multiplying by η in $M = b + \eta d$ gives (170), and the three displayed derivative bounds follow from the preceding estimates. Only the displacement $\eta\|v\|$ is restricted to the fixed local ball; the displayed powers of $\|v\|$ account explicitly for the magnitude of the direction. Thus increasing the polynomial sample envelope with H does not enter the remainder constant. □

Lemma C.25 (c-uniform coupled expansion for the exact deep KL updates)**.** *Fix a common state law ν and let*

$$U^+ = \mathsf{M}_{\mathbf{m},\nu,\eta}(U, \widehat{A}),$$
$$U^+ + cS^+ = \mathsf{M}_{\mathbf{m},\nu,\eta}(U + cS, \widehat{A} + c\widehat{R}).$$

Set

$$P^0 = P_{\mathbf{m},\nu}(U), \qquad P^c = P_{\mathbf{m},\nu}(U + cS), \qquad K_c(U, S) = \frac{P^0 - P^c}{c}. \tag{177}$$

Then

$$U^+ - U = -\eta P^0 \widehat{A} + \rho^U, \tag{178}$$
$$S^+ - S = -\eta P^c \widehat{R} + \eta K_c(U, S)\widehat{A} + \rho^S, \tag{179}$$
$$\|K_c(U, S)\|_{\mathrm{op}} \le L_{P,L} \|S\|, \tag{180}$$

where

$$\|\rho^U\| \le C_{Q,L}\eta^2 \left\|\widehat{A}\right\|^2, \tag{181}$$
$$\|\rho^S\| \le C_{Q,L}\eta^2 \left[\|S\| \left(\left\|\widehat{A}\right\| + c\left\|\widehat{R}\right\|\right)^2 + \left(\left\|\widehat{A}\right\| + c\left\|\widehat{R}\right\|\right)\left\|\widehat{R}\right\|\right]. \tag{182}$$

Every constant is uniform as $c \downarrow 0$ and over the hidden widths.

Proof. Apply (170) to both updates and subtract before dividing by c:

$$S^+ - S = -\eta \frac{P^c(\widehat{A} + c\widehat{R}) - P^0\widehat{A}}{c} + \eta^2 \frac{Q(U + cS, \widehat{A} + c\widehat{R}, \eta) - Q(U, \widehat{A}, \eta)}{c}.$$

The inverse-Fisher Lipschitz estimate gives (180). For the nonlinear remainder, apply the fundamental theorem of calculus along $\theta \mapsto (U + \theta cS, \widehat{A} + \theta c\widehat{R})$ and use (172)–(173). This proves (182); the base estimate is immediate from (171). □

Lemma C.26 (Deterministic prospective actor increments)**.** *On the safe actor set, for sufficiently small absolute steps,*

$$\|U^+ - U\| + \|S^+ - S\| \le C\eta_y \,\mathrm{poly}(H).$$

The constants are uniform in c, width and critic scale.

Proof. Lemma C.25 bounds the virtual exact increments by linear sampled directions plus quadratic local-map remainders. The directions are polynomial in H by Lemma C.43, and $\|S\| \le s$, $c \le c_0$, with uniformly bounded inverse metrics. Lemma C.35 adds at most $c\eta_y^2$ to the base increment and $2\eta_y^2$ to the normalized increment. The displayed bound therefore holds for the finite implementation as well when $c, \eta_y \le 1$. Selecting accuracy small enough gives $\|\Delta U\| < r_{\mathrm{prox}}/2$ and $\|\Delta S\| < s/4$. For all interpolation points used in the paired remainder proof, the base point is $U + tcS$ and its direction is $\widehat{a} + tc\widehat{R}$, $0 \le t \le 1$. They obey the same buffered actor inclusion and direction bound, so the local map exists along the entire interpolation before containment is invoked. □

The population direction. Let $p_- = \overline{\lambda}_{F,L}^{-1}$, $p_+ = \underline{\lambda}_{F,L}^{-1}$ and $K_{\max} = L_{P,L}s$. The constant d_* and the weight λ_S are fixed in Lemma C.20.

Lemma C.27 (Population coupled direction is descending)**.** *Let $A = a(p) = g_\alpha(x, U)$ and $R = R_c(p)$. Let*

$$d_U = -P^0 A,$$
$$d_S = -P^c R + K_c(U, S)A,$$

where P^0, P^c are inverse common-state Fisher matrices. Then

$$\langle \nabla_U \mathcal{T}, d_U \rangle + \langle \nabla_S \mathcal{T}, d_S \rangle \le -\rho_{0,L}\mathcal{T}, \qquad \rho_{0,L} := p_- \min\{\mu_\tau, \mu_\tau\}. \tag{183}$$

Proof. Since $p_- I \preceq P^0, P^c \preceq p_+ I$ and $\|K_c\| \le K_{\max}$, the left-hand side is at most

$$-p_- \|A\|^2 - \lambda_S p_- \|R\|^2 + \lambda_S(p_+ d_* + K_{\max}) \|A\| \|R\|.$$

Let $D = p_+ d_* + K_{\max}$. Young's inequality gives

$$\lambda_S D \|A\| \|R\| \le \frac{p_-}{4} \|A\|^2 + \frac{\lambda_S^2 D^2}{p_-} \|R\|^2.$$

The choice (152) makes the last coefficient at most $\lambda_S p_-/4$. Hence the directional derivative is at most $-(p_-/2)\|A\|^2 - (\lambda_S p_-/2)\|R\|^2$. The PL inequalities for the base gap and Γ give $\|A\|^2 \ge 2\mu_\tau \Delta_0$ and $\|R\|^2 \ge 2\mu_\tau \Gamma$, proving (183). □

Throughout this subsection, suppress width subscripts and write $Z_\alpha = \nabla_\alpha F_{\mathbf{m},\alpha}$, $G_Z = C_{1,L}$, $\lambda_0 = \lambda_{\mathrm{eff,L}}$ and $L_3 = M_{\log,3,L}$. The actor bounds in Appendix C.1 and the radius (33) give

$$\|Z_\alpha(s,a)\| \le G_Z, \qquad \|Z_\alpha(s,a) - Z_b(s,a)\| \le C_{2,L}\|\alpha - b\|, \tag{184}$$
$$\mathcal{G}_0 := \mathbb{E}_\rho \operatorname{Cov}_{\pi_{\mathrm{ref}}}(Z_0) \succeq \lambda_0 I_r, \qquad C_{2,L}R \le \sqrt{\lambda_0}/2, \tag{185}$$
$$\pi_\alpha(a \mid s) \ge e^{-2G_Z R}\pi_{\mathrm{ref}}(a \mid s), \qquad \|D_\alpha^3 \log \pi_\alpha(a \mid s)\|_{\mathrm{op}} \le L_3. \tag{186}$$

All these inequalities hold on $\overline{B}_R$. In particular, the population constants already used in the exact reference-map analysis are

$$\underline{\lambda}_{F,L} = (1-\gamma)e^{-2G_Z R}\lambda_0/4, \qquad \overline{\lambda}_{F,L} = 4G_Z^2. \tag{187}$$

The finite metric below preserves these constants, so it does not change the fixed regularization, the actor buffers or the domain required by C2. It is used only for update geometry: the population objectives, selected response and all trajectory laws for evaluation and TD retain their original definitions. Actor probabilities and derivatives are computed on finitely many states with full finite-action sums, in the same real-arithmetic model as the other neural calculations.

C.4.1 An empirical metric with the existing Fisher constants

Definition C.28 (One initial metric sample)**.** Fix the realized actor initialization and retained chart before drawing metric data. For a chosen $\delta_{\mathrm{met}} \in (0,1)$, independently draw $s_1, \dots, s_n$ from ρ, with

$$n \ge \left\lceil \frac{8G_Z^2}{\gamma^2 \lambda_0} \log \frac{r}{\delta_{\mathrm{met}}} \right\rceil. \tag{188}$$

Define

$$\widehat{\rho}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{s_i}, \tag{189}$$

$$B(s) = \mathrm{Cov}_{a\sim\pi_{\mathrm{ref}}(\cdot|s)}(Z_0(s,a)), \qquad \widehat{\mathcal{G}}_n = \frac{1}{n}\sum_{i=1}^{n} B(s_i). \tag{190}$$

Compute each covariance by summing over all actions. Accept the metric if

$$\widehat{\mathcal{G}}_n \succeq (1-\gamma/2)\lambda_0 I_r. \tag{191}$$

If the certificate fails, return the prescribed fallback x_0. Upon acceptance, use $\bar{d}_t = \widehat{\rho}_n$ for both endpoint updates at every iteration. All subsequent training blocks are generated using the specified sampling protocol, independently of the metric data conditional on the old parameters and history.

This is the actual initialization covariance under $\widehat{\rho}_n$. The specified optional pilot based on two sampled actions estimates a population covariance and is a different matrix. The full action sums in (190) directly certify the empirical KL measure used here.

Proposition C.29 (Certificate acceptance probability)**.** *Conditional on any fixed initialization and chart satisfying A1–A2, Definition C.28 rejects with probability at most* δ_{met}.

Proof. For every unit vector v,

$$0 \le v^\top B(s) v = \mathrm{Var}_{\pi_{\mathrm{ref}}}(v^\top Z_0(s,\cdot)) \le \mathbb{E}_{\pi_{\mathrm{ref}}}(v^\top Z_0(s,\cdot))^2 \le G_Z^2.$$

Thus $0 \preceq B(s_i) \preceq G_Z^2 I_r$, and $\mathbb{E}B(s_i) = \mathcal{G}_0 \succeq \lambda_0 I_r$. Let $\mu_{\min} = \lambda_{\min}(\sum_i \mathbb{E}B(s_i)) \ge n\lambda_0$. The matrix Chernoff lower-tail inequality (Tropp, 2012, Corollary 5.2 and Remark 5.3) gives

$$\mathbb{P}\left(\lambda_{\min}\Big(\sum_i B(s_i)\Big) \le (1-t)\mu_{\min}\right) \le r\exp\left(-\frac{t^2\mu_{\min}}{2G_Z^2}\right), \qquad 0<t<1.$$

The rejection event in (191) is contained in this event with $t = \gamma/2$. Hence

$$\mathbb{P}(\text{metric rejection}) \le r\exp\left(-\frac{n\gamma^2\lambda_0}{8G_Z^2}\right) \le \delta_{\mathrm{met}}. \tag{192}$$

□

Proposition C.30 (Uniform empirical Fisher certificate)**.** *On the observable acceptance event, simultaneously for every* $\alpha \in \overline{B}_R$,

$$\underline{\lambda}_{F,L} I_r \preceq \widehat{\mathcal{F}}_n(\alpha) := \mathcal{F}_{\mathbf{m},\widehat{\rho}_n}(\alpha) \preceq \overline{\lambda}_{F,L} I_r. \tag{193}$$

These are the constants in (187). The inverse $P_n(\alpha) = \widehat{\mathcal{F}}_n(\alpha)^{-1}$ *is uniformly bounded and Lipschitz, with bounds independent of* n.

Proof. Fix a unit v, and use the joint law $s \sim \widehat{\rho}_n$, $a \sim \pi_{\mathrm{ref}}(\cdot \mid s)$. Conditional centering is an orthogonal projection in L^2. Applying the triangle inequality to centered $v^\top Z_\alpha$ and $v^\top Z_0$, followed by (184), gives

$$\left(\mathbb{E}_{\widehat{\rho}_n}\mathrm{Var}_{\pi_{\mathrm{ref}}}(v^\top Z_\alpha)\right)^{1/2} \ge \left(\mathbb{E}_{\widehat{\rho}_n}\mathrm{Var}_{\pi_{\mathrm{ref}}}(v^\top Z_0)\right)^{1/2} - C_{2,L}\,\|\alpha\|\,.$$

On acceptance, the first term on the right is at least $\sqrt{(1-\gamma/2)\lambda_0}$. The right-hand side is positive on $\overline{B}_R$ by (185).

For any real-valued action function Y, the density lower bound in (186) implies

$$\mathrm{Var}_{\pi_\alpha}(Y) = \inf_{a_0\in\mathbb{R}} \mathbb{E}_{\pi_\alpha}(Y-a_0)^2 \ge e^{-2G_Z R}\inf_{a_0\in\mathbb{R}} \mathbb{E}_{\pi_{\mathrm{ref}}}(Y-a_0)^2.$$

Therefore

$$v^\top \widehat{\mathcal{F}}_n(\alpha) v \ge e^{-2G_Z R}\left(\sqrt{(1-\gamma/2)\lambda_0} - C_{2,L}R\right)^2. \tag{194}$$

Set $a = \sqrt{1-\gamma}$. The root-mean-square inequality gives

$$\sqrt{1-\gamma/2} = \sqrt{\frac{1+a^2}{2}} \geq \frac{1+a}{2}.$$

Combining this with (185) proves the lower bound in (193). The centered feature has norm at most $2G_Z$, proving the upper bound.

The specified log-policy derivative bounds give, for every fixed probability law ν,

$$\|\mathcal{F}_{\mathbf{m},\nu}(\alpha) - \mathcal{F}_{\mathbf{m},\nu}(\beta)\|_{\mathrm{op}} \leq L_F \|\alpha - \beta\|,$$

where L_F is independent of ν. Apply this to $\widehat{\rho}_n$, and use the identity $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$, to obtain

$$\|P_n(\alpha)\|_{\mathrm{op}} \leq \underline{\lambda}_{F,L}^{-1}, \qquad \|P_n(\alpha) - P_n(\beta)\|_{\mathrm{op}} \leq \frac{L_F}{\underline{\lambda}_{F,L}^2} \|\alpha - \beta\|. \tag{195}$$

□

Remark C.31 (Why one sample works throughout training). The accepted covariance certificate implies (193) for all points of the actor ball, including later data-dependent iterates. No union bound over iterations or covering argument over actor parameters is needed. An atomic empirical measure need not dominate $(1-\gamma)\rho$; the proof above establishes its Fisher bound directly. Nor is an approximation of the current population occupancy required: the existing update analysis uses the measure as a predictable, well-conditioned common metric.

The threshold $(1-\gamma/2)\lambda_0$ is chosen to reproduce the existing constants. A half-gap threshold also gives positive curvature with suitable constants, but is not used for that purpose here. The γ^{-2} factor in (188) is an explicit sufficient cost of this choice; no optimal dependence on γ is claimed.

C.4.2 An explicit finite local solver

Fix accepted metric data, a buffered old actor b, and a frozen sampled direction h, with $\|h\| \leq G_{\mathrm{prox}}(H)$. Set

$$\widehat{\mathcal{K}}_n(\alpha \mid b) = \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}\big(\pi_b(\cdot \mid s_i) \,\|\, \pi_\alpha(\cdot \mid s_i)\big), \tag{196}$$

$$\ell_{b,h}(\alpha) = \widehat{\mathcal{K}}_n(\alpha \mid b) + \eta_y \langle h, \alpha - b\rangle. \tag{197}$$

This scaled objective has the same minimizer as the specified linear cost plus η_y^{-1} times KL. Every derivative in the numerical solver is with respect to α, holding b, h, and the metric sample fixed.

The gradient is the finite sum

$$\nabla \ell_{b,h}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \big(\pi_\alpha(a \mid s_i) - \pi_b(a \mid s_i)\big) Z_\alpha(s_i, a) + \eta_y h. \tag{198}$$

Indeed, the derivative of KL is $-\sum_a \pi_b(a \mid s)\nabla_\alpha \log \pi_\alpha(a \mid s)$; substitute the centered-feature identity from the softmax actor definition. The calculation uses actor first derivatives and finite action sums. It requires neither transition probabilities nor a population-state expectation.

Retain the local radius and step restrictions

$$0 < r_{\mathrm{prox}} \leq \min\left\{\frac{R}{8}, \frac{\underline{\lambda}_{F,L}}{2L_3}\right\}, \qquad \eta_y G_{\mathrm{prox}}(H) \leq \frac{m_{K,L} r_{\mathrm{prox}}}{2}, \tag{199}$$

omitting the second radius restriction if $L_3 = 0$, where

$$m_{K,L} = \frac{\underline{\lambda}_{F,L}}{2}, \qquad M_{K,L} = \overline{\lambda}_{F,L} + L_3 r_{\mathrm{prox}}, \qquad \kappa_K = \frac{M_{K,L}}{m_{K,L}}. \tag{200}$$

Assume the local ball $\mathcal{B}_b = \overline{B}(b, r_{\mathrm{prox}})$ lies in $\overline{B}_R$. The specified observable safe-set test provides this buffer for both endpoint bases and their interpolation.

By Lemmas C.23 and C.24, $\ell_{b,h}$ is $m_{K,L}$-strongly convex and $M_{K,L}$-smooth on $\mathcal{B}_b$, and its reference minimizer $M(b,h) = \mathsf{M}_{\mathbf{m},\widehat{\rho}_n,\eta_y}(b,h)$ is interior, with $\nabla \ell_{b,h}(M) = 0$ and $\|M - b\| \leq \eta_y \|h\|/m_{K,L} \leq r_{\mathrm{prox}}/2$. These are analysis points; the algorithm computes the finite routine below.

Definition C.32 (Residual-controlled projected gradient descent)**.** Take $0 < c \le \min\{c_0, 1\}$ and set

$$d_{\mathrm{sol}} = c\eta_y^2, \qquad t_{\mathrm{res}} = m_{K,L} c\eta_y^2. \tag{201}$$

Initialize $\alpha^{(0)} = b$. At numerical iteration j, compute $g^{(j)} = \nabla \ell_{b,h}(\alpha^{(j)})$ from (198). If

$$\left\| g^{(j)} \right\| \le t_{\mathrm{res}}, \tag{202}$$

return $\widetilde{\mathsf{M}}(b,h) = \alpha^{(j)}$. Otherwise set

$$\alpha^{(j+1)} = \Pi_{\mathcal{B}_b}\left(\alpha^{(j)} - \frac{g^{(j)}}{2M_{K,L}} \right) \tag{203}$$

and continue. The Euclidean projection is explicit:

$$\Pi_{\mathcal{B}_b}(y) = b + (y - b)\min\{1, r_{\mathrm{prox}} / \|y - b\|\},$$

with value b when $y = b$. The same frozen metric sample is reused at every numerical iteration.

Proposition C.33 (Finite termination and certified endpoint accuracy)**.** *Definition C.32 returns an endpoint satisfying*

$$\left\| \widetilde{\mathsf{M}}(b,h) - M(b,h) \right\| \le c\eta_y^2. \tag{204}$$

Define $\log_+(u) = \max\{0, \log u\}$ *for* $u > 0$*. The number of projected updates is at most*

$$J_{\max} = \left\lceil 2\kappa_K \log_+ \left(\frac{M_{K,L} G_{\mathrm{prox}}(H)}{m_{K,L}^2 c\eta_y} \right) \right\rceil. \tag{205}$$

At most $J_{\max} + 1$ *finite-sum gradients are evaluated.*

Proof. Let $M = M(b,h)$. For $\alpha \in \mathcal{B}_b$, the integral Hessian along the segment from M to α has eigenvalues in $[m_{K,L}, M_{K,L}]$. Since $\nabla \ell_{b,h}(M) = 0$, the unprojected gradient map with step $1/(2M_{K,L})$ contracts distance to M by at most

$$q_K = 1 - \frac{m_{K,L}}{2M_{K,L}}.$$

Projection onto the convex ball is nonexpansive and fixes M. Consequently,

$$\left\| \alpha^{(j)} - M \right\| \le q_K^j \|b - M\| \le e^{-j/(2\kappa_K)} \frac{\eta_y G_{\mathrm{prox}}(H)}{m_{K,L}}. \tag{206}$$

Local smoothness gives

$$\left\| \nabla \ell_{b,h}(\alpha^{(j)}) \right\| \le M_{K,L} \left\| \alpha^{(j)} - M \right\| \le \frac{M_{K,L} \eta_y G_{\mathrm{prox}}(H)}{m_{K,L}} e^{-j/(2\kappa_K)}.$$

This is at most $m_{K,L} c\eta_y^2$ by $j = J_{\max}$. Finally, strong monotonicity and the zero gradient at M imply

$$m_{K,L} \|\alpha - M\|^2 \le \langle \nabla \ell_{b,h}(\alpha), \alpha - M \rangle \le \|\nabla \ell_{b,h}(\alpha)\| \, \|\alpha - M\|.$$

At the stopping point, this yields (204). All iterates stay in the local ball by projection, so the proof does not presuppose actor containment beyond the current buffered step. □

Remark C.34 (Normalization and arithmetic). The stopping test is for the scaled objective (197). If implemented using $\langle h, \alpha - b \rangle + \eta_y^{-1} \widehat{\mathcal{K}}_n(\alpha \mid b)$, the equivalent gradient-residual threshold is $m_{K,L} c\eta_y$. The parameter tolerance $c\eta_y^2$ is not an objective-value tolerance. With $c = \Theta(\sqrt{\epsilon})$ and $\eta_y = \Theta(\epsilon)$, it is $O(\epsilon^{5/2})$; geometric convergence gives a logarithmic iteration dependence on this accuracy. Neither the stopping rule nor the count above assumes an accurate optimization oracle. Gradient evaluations are the explicit finite neural computations in (198).

C.4.3 PAIRED NUMERICAL ERRORS AND THE SAMPLING HISTORY

At an old iterate $p = (x, U, S)$, use the specified frozen lower estimates $(\widehat{a}, \widehat{R})$, and form $\widehat{C} = \widehat{a} + c\widehat{R}$. The two bases and directions are

$$b_0 = U, \quad h_0 = \widehat{a}, \qquad b_c = U + cS, \quad h_c = \widehat{a} + c\widehat{R}.$$

Let $M_0 = M(b_0, h_0)$, $M_c = M(b_c, h_c)$. These are mathematical reference points used only in the proof. The implemented updates are

$$U^+ = \widetilde{\mathsf{M}}(b_0, h_0), \qquad \alpha_c^+ = \widetilde{\mathsf{M}}(b_c, h_c), \qquad S^+ = \frac{\alpha_c^+ - U^+}{c}. \tag{207}$$

Lemma C.35 (Samplewise paired error)**.** *At the same old iterate and with the same sampled directions, define the virtual exact update $\bar{U}^+ = M_0$, $\bar{S}^+ = (M_c - M_0)/c$. Then, samplewise,*

$$e_U := U^+ - \bar{U}^+, \qquad \|e_U\| \le c\eta_y^2, \tag{208}$$

$$e_S := S^+ - \bar{S}^+, \qquad \|e_S\| \le 2\eta_y^2. \tag{209}$$

In particular, for $c \le 1$,

$$\|(e_U, e_S)\| \le \sqrt{5}\,\eta_y^2. \tag{210}$$

No differentiability of the stopping rule or coupling of numerical stopping times is required.

Proof. Set $e_0 = \widetilde{\mathsf{M}}(b_0, h_0) - M_0$ and $e_c = \widetilde{\mathsf{M}}(b_c, h_c) - M_c$. Proposition C.33 gives $\|e_0\|, \|e_c\| \le c\eta_y^2$. The identities $e_U = e_0$ and $e_S = (e_c - e_0)/c$ prove the first two bounds. Their squared sum is at most $(c^2 + 4)\eta_y^4$. □

For completeness, the exact reference-map expansion from Lemma C.25 therefore yields the implemented expansion

$$U^+ - U = -\eta_y P^0 \widehat{a} + \rho_U + e_U, \tag{211}$$

$$S^+ - S = -\eta_y P^c \widehat{R} + \eta_y K_c \widehat{a} + \rho_S + e_S, \tag{212}$$

$$P^0 = P_n(U), \qquad P^c = P_n(U + cS), \qquad K_c = \frac{P^0 - P^c}{c}.$$

The original remainders ρ_U, ρ_S have conditional first moments $O(\eta_y^2)$, second moments $O(\eta_y^4)$, and fourth moments $O(\eta_y^8)$. Adding (e_U, e_S) preserves these orders: its corresponding bounds follow immediately from (210). The normalized inverse-metric difference is still bounded by $\|K_c\|_{\mathrm{op}} \le L_F \underline{\lambda}_{F,L}^{-2} \|S\|$.

Lemma C.36 (Filtration and prospective-step validity)**.** *Include the entire initial metric sample in the pre-draw history $\mathcal{F}_t$. Use old parameters for all three training blocks, and run both local solvers solely with the frozen lower block and metric data. Then the simultaneous (x, U, S) increment remains conditionally independent of the current TD block. The conditional movement estimates of Appendix C.7 retain their orders. For sufficiently small accuracy in the headline schedule, the implemented prospective step lies in the existing larger analytic actor neighborhood, before containment has been established.*

Proof. Each numerical iterate and its stopping index are measurable functions of the old history and lower block. The outer update uses its independent old-parameter block. The TD block is independent of these two blocks conditional on $\mathcal{F}_t$, as stipulated in the existing protocol. Hence the conditional centering argument for moving-target TD applies. The mean and centered numerical perturbations have orders $O(\eta_y^2)$, with second and fourth moment bounds supplied by (210); these preserve the movement estimates.

At a safe old point, both local balls and the corresponding balls along the endpoint interpolation lie in $\overline{B}_R$. The projected inner iterations stay in their respective balls. The virtual exact coupled increments have deterministic bound $C\eta_y \,\mathrm{poly}(H)$ by the specified paired expansion. The implemented increments add at most $\sqrt{5}\eta_y^2$. With logarithmic H and $\eta_y = \Theta(\epsilon)$, both tend to zero. In particular, the existing prospective restrictions $\|\Delta U\| < r_{\mathrm{prox}}/2$ and $\|\Delta S\| < s/4$ hold for sufficiently small ϵ. The virtual and implemented points, and the segment between them, lie in the same convex buffered analytic neighborhood. This argument uses only the current safe point and therefore introduces no circular appeal to future containment. □

C.4.4 TRANSFER TO TRACKING, CONTAINMENT, AND STATIONARITY

Use the specified actor potential $\mathcal{T}(p) = \Delta_0(x, U) + \lambda_S \Gamma_c(x, U, S)$. To distinguish the moving critic target from the actor feature Z_α, write that target as $\mathcal{Z}(p)$, and set

$$\mathcal{C}(p, z) = \|z - \mathcal{Z}(p)\|_W^2, \qquad X(p) = \|\nabla\Phi(x)\|^2, \qquad \omega = c^2 + \mathcal{A}_Q + b_{\mathrm{ev}}^2 + b_{\mathrm{TD}}^2.$$

The established bounds on the prospective neighborhood include

$$\|\nabla_p \mathcal{T}\| \le L_{\mathcal{T}}, \qquad \|\mathcal{Z}(p) - \mathcal{Z}(\bar{p})\|_W \le L_{\mathcal{Z}} \|p - \bar{p}\|, \qquad \mathcal{Z}(p), z \in \mathcal{K}_Q. \tag{213}$$

These constants are global in the upper parameter and uniform in the critic scale and c. Put $D_Q = \mathrm{diam}_W(\mathcal{K}_Q) < \infty$.

Proposition C.37 (One-step transfer without new accuracy assumptions)**.** *At a safe old iterate, compare the implemented step with the virtual exact-proximal step at the same old iterate, using identical lower, outer, and TD blocks. The outer and critic updates themselves agree, and the next actor/upper tuples $p^+, \bar{p}^+$ satisfy*

$$x^+ = \bar{x}^+, \qquad z^+ = \bar{z}^+, \qquad \left\|p^+ - \bar{p}^+\right\| \le \sqrt{5}\eta_y^2.$$

Consequently,

$$|\mathcal{T}(p^+) - \mathcal{T}(\bar{p}^+)| \le \sqrt{5} L_{\mathcal{T}} \eta_y^2, \tag{214}$$

$$|\mathcal{C}(p^+, z^+) - \mathcal{C}(\bar{p}^+, z^+)| \le 2\sqrt{5} D_Q L_{\mathcal{Z}} \eta_y^2, \tag{215}$$

$$\Phi(x^+) - \Phi(\bar{x}^+) = 0. \tag{216}$$

The existing conditional actor/critic tracking recursions and outer descent inequality thus retain their form, with enlarged second-order remainder constants.

Proof. Both the outer and TD routines use the old parameters, so using the same sampled blocks gives identical x^+ and z^+. Lemma C.35 bounds the remaining difference. Integrating the actor-potential gradient along the segment supplied by Lemma C.36 gives (214). For the critic error, both $z^+ - \mathcal{Z}(p^+)$ and $z^+ - \mathcal{Z}(\bar{p}^+)$ have W-norm at most D_Q. The difference of their squared norms is at most

$$2D_Q \left\|\mathcal{Z}(p^+) - \mathcal{Z}(\bar{p}^+)\right\|_W,$$

which proves (215). The last identity is immediate.

The specified exact-map inequalities are one-step inequalities uniform over safe old iterates. They can therefore be applied to the virtual step at each implemented iterate; a separate exact-solver trajectory is unnecessary. Take conditional expectations in the displayed samplewise bounds. Each additional actor or critic term is $O(\eta_y^2)$, already present in the corresponding recurrence. The outer descent inequality is unchanged. □

In particular, the transferred recursions have the existing orders

$$\mathbb{E}_t \Delta\mathcal{C}_t \le -k\beta\mathcal{C}_t + C\frac{\eta_y^2}{\beta}\mathcal{T}_t + C\frac{\eta_x^2}{\beta}X_t + C\beta\omega + C(\beta^2 + \eta_y^2 + \eta_x^2), \tag{217}$$

$$\mathbb{E}_t \Delta\mathcal{T}_t \le -u\eta_y\mathcal{T}_t + v\eta_x X_t + w\eta_y\mathcal{C}_t + C\eta_y\omega + C(\eta_y^2 + \eta_x^2), \tag{218}$$

$$\mathbb{E}_t \Delta\Phi_t \le -a\eta_x X_t + b\eta_x\mathcal{T}_t + d\eta_x\mathcal{C}_t + C\eta_x\omega + C\eta_x^2. \tag{219}$$

Here the positive drift constants and sufficiently small fixed step-ratio choices are those furnished by the specified proof; the numerical perturbation enters the indicated remainder terms. In particular, it does not create a new c^{-1} error floor.

Corollary C.38 (Containment and augmented descent)**.** *Let $\eta_y = r_Q\beta$, $\eta_x = q\eta_y$, with the sufficiently small fixed positive ratios selected in the analysis. The contracting potential $\mathcal{D} = \mathcal{T} + \lambda_D r_Q \mathcal{C}$ and stationarity potential*

$$\Psi = \Phi - \underline{\Phi} + a_\Psi q\mathcal{T} + b_\Psi q r_Q \mathcal{C}$$

retain their required drift and innovation bounds. In particular, the new contribution to the one-step drift of Ψ is $O(\eta_x\eta_y)$, and the added contribution to the drift of $\mathcal{D}$ is $O(\eta_y^2)$.

Proof. For Ψ, multiply (214) by $a_\Psi q$ and (215) by $b_\Psi q r_Q$. Since $r_Q \le 1$ and $q\eta_y = \eta_x$, the extra term is bounded by $Cq\eta_y^2 = C\eta_x\eta_y$. It is absorbed in the existing $C\eta_x(\omega + \beta + \eta_y + \eta_x)$ remainder. For $\mathcal{D}$, the additional term is $C\eta_y^2$. After division by η_y, it is $C\eta_y \le C\beta$, again within the existing contracting floor.

The difference between the implemented next $\mathcal{D}$ and the virtual exact next $\mathcal{D}$ is samplewise $O(\eta_y^2)$. Centering at most doubles this bound. The original innovation bound $|N_{t+1}| \le \eta_y B_D(H)$ therefore remains valid after enlarging its polynomial envelope. Together with Lemma C.36, the same stopped martingale maximum argument and first-exit contradiction apply. □

C.5 Critic calculus and auxiliary TD estimates

The critic in (6) trains a fixed-dimensional chart θ. Increasing m_Q changes its nonlinear scale, not its number of independent trainable coordinates. For integer m_Q, tied tanh units give this realization exactly.

Lemma C.39 (Nonlinear critic derivatives and linearization)**.** *On $\|\theta\| \le R_Q$, the following bounds hold uniformly over inputs:*

$$\begin{aligned} |Q_{m_Q,\theta}| &\le F_Q R_Q, & \left\|\nabla Q_{m_Q,\theta}\right\| &\le F_Q, \\ \left\|Q_{m_Q,\theta} - \varphi_Q^\top \theta\right\|_\infty &\le F_Q^3 R_Q^3/(3m_Q), & \left\|\nabla Q_{m_Q,\theta} - \varphi_Q\right\|_\infty &\le F_Q^3 R_Q^2/m_Q, \\ \left\|\nabla^2 Q_{m_Q,\theta}\right\|_\infty &\le 2F_Q^3 R_Q/m_Q, & \left\|\nabla^3 Q_{m_Q,\theta}\right\|_\infty &\le 2F_Q^3/m_Q. \end{aligned} \tag{220}$$

Fourth derivatives are also uniformly bounded for $m_Q \ge 1$.

Proof. Put $t = \varphi_Q^\top \theta/\sqrt{m_Q}$. Direct differentiation gives $\nabla Q = \operatorname{sech}^2(t)\varphi_Q$ and $\nabla^2 Q = -2\operatorname{sech}^2(t)\tanh(t)\varphi_Q^{\otimes 2}/\sqrt{m_Q}$. Use $|\tanh t| \le |t|$, $|\tanh t - t| \le |t|^3/3$, and $1 - \operatorname{sech}^2 t \le t^2$. The third derivative is $(4\operatorname{sech}^2 t - 6\operatorname{sech}^4 t)\varphi_Q^{\otimes 3}/m_Q$, whose scalar coefficient has magnitude at most two. The fourth derivative is a bounded polynomial in $\tanh t$ and $\operatorname{sech} t$, multiplied by $\varphi_Q^{\otimes 4}/m_Q^{3/2}$. These formulas prove all claims. □

Lemma C.40 (Anisotropic derivative bounds on unrestricted x)**.** *Regard (U, α, θ) as independent variables and denote this group by y. On the indicated neighborhoods, every derivative block*

$$D_x^a D_y^b \mathsf{H}(x, U, \alpha, \theta), \qquad 0 \le a \le 2, \quad a + b \le 3, \tag{221}$$

is uniformly bounded. Derivatives of $\widehat{\mathsf{H}}$ at a fixed transition, holding its sampling law fixed, have uniform envelopes of the corresponding orders. Derivatives of ν_U are treated separately by likelihood differentiation and geometric moments; their prefix likelihood factors have polynomial length envelopes. Pure $D_x^3\mathsf{H}$ is not needed or claimed uniformly bounded.

Proof. The base occupancy is independent of x and α. Its U derivatives through order three are controlled by differentiating the prefix likelihood under π_U. Bounded derivatives of the log-policy give degree-at-most-three polynomials in prefix length. The sums $\sum_{k\ge 0}(1-\gamma)\gamma^k(k+1)^j$ are finite. This proves differentiation under the discounted sum in total variation and supplies uniform bounds for those derivatives.

For the TD integrand, an x derivative acts only on $q = -r_x + \tau\log(\pi_\alpha/\pi_{\mathrm{ref}})$. Thus $D_x^a q = -D_x^a r_x$ for $a = 1, 2$, and $D_x^a D_\alpha^b q = 0$ when $a, b \ge 1$. Actor derivatives act on K_α, the entropy criterion, or the base likelihood. Three derivatives in θ of the TD integrand use derivatives of Q through order four, uniformly bounded by (220) and its fourth-order analogue. Product differentiation and (14) give every block in (221).

For example, at independent (U, α, θ), $\mathsf{H}_{xx} = \mathbb{E}_{\nu_U}[r_{xx}\nabla Q]$, $\mathsf{H}_{xx\alpha} = 0$, and $\mathsf{H}_{xx\theta} = \mathbb{E}_{\nu_U}[r_{xx}\nabla^2 Q]$; H_{xxU} differentiates only the base law in the first expression. These are the potentially delicate blocks needed after taking the normalized endpoint difference. No pure third x derivative enters. The same domination proves continuity of these derivatives, not merely a norm bound. □

C.6 SAMPLING ROUTINES AND PAIRED ORACLE ESTIMATES

All blocks use the pre-draw history and old parameters specified at the start of Appendix B, including the stored accepted metric sample. The reset states used for that metric are independent of these training blocks and are reused only for the finite KL computations. Write $M_1 = M_{\log,1,L}$. Define the sample TD residual at a common transition by

$$\begin{aligned}\widehat{\mathsf{H}}(x,U,\alpha,\theta) &= \Big[Q_{m_Q,\theta}(s,a) - q_{x,\alpha}(s,a) \\ &\qquad - \gamma \sum_{a'} \pi_\alpha(a' \mid s') Q_{m_Q,\theta}(s',a')\Big] \nabla_\theta Q_{m_Q,\theta}(s,a), \\ F^{\mathrm{samp}}(p,z;s,a,s') &= \begin{pmatrix} \widehat{\mathsf{H}}(x,U,U,u) \\ [\widehat{\mathsf{H}}(x,U,U+cS,u+cv) - \widehat{\mathsf{H}}(x,U,U,u)]/c \end{pmatrix}. \end{aligned} \tag{222}$$

Here $q_{x,\alpha}$ is the regularized lower criterion in (15). Both endpoints use the same base-law transition; the next-action expectations are exact finite sums. With $\widehat{F}_{t,H}$ the truncated sample average, update

$$z_{t+1} = \Pi^W_{\mathcal{K}_Q}(z_t - \beta \widehat{F}_{t,H}), \qquad \mathcal{K}_Q = \overline{B}_{R_u} \times \overline{B}_{R_v}, \qquad \|(u,v)\|^2_W = M\|u\|^2 + \|v\|^2. \tag{223}$$

The critic construction in Step B.2 supplies fixed R_u, R_v, M and an admissible critic scale. Projection is a specified critic update; critic accuracy is proved. Initialize any deterministic $z_0 \in \mathcal{K}_Q$, without an accuracy or burn-in requirement. Both algorithms use the single actor test (65), initially and after every completed iteration.

Explicit evaluation samplers. Write $\alpha_0 = U$, $\alpha_c = U + cS$, $\theta_0 = u$, $\theta_c = u + cv$, and $Q_j = Q_{m_Q,\theta_j}$. The critic represents the unnormalized discounted action value, whereas g_τ is normalized by $1-\gamma$. Consequently the following score-times-value actor estimator has no additional $(1-\gamma)^{-1}$ factor. For a common-start mixture prefix $\xi = (s_0, a_0, \dots, s_n, a_n)$ with $s_0 \sim \rho$, set

$$\ell_j(\xi) = \prod_{k=0}^{n} \pi_{\alpha_j}(a_k \mid s_k), \qquad w_j(\xi) = \frac{2\ell_j(\xi)}{\ell_0(\xi) + \ell_c(\xi)}, \qquad j \in \{0, c\}. \tag{224}$$

The terminal action is included. Shared reset and transition factors cancel in this ratio. At the same terminal state-action pair define

$$\begin{aligned} G_j(\xi) &= \nabla_\alpha \log \pi_{\alpha_j}(a_n \mid s_n) Q_j(s_n,a_n) + \partial_\alpha q_{x,\alpha_j}(s_n,a_n), \\ a(\xi) &= w_0 G_0, \qquad d(\xi) = \frac{w_c G_c - w_0 G_0}{c}, \qquad e(\xi) = -\frac{w_c - w_0}{c} \nabla_x r_x(s_n,a_n). \end{aligned} \tag{225}$$

The direct criterion derivative $\partial_\alpha q$ differentiates the policy-ratio criterion, not the trained critic coordinates. For a fresh preference tuple and label at the old perturbed actor,

$$P_\alpha(\boldsymbol{\xi}, Y) = L(x;\boldsymbol{\xi},Y) \sum_{i=1}^{I} \sum_{k=0}^{K-1} \nabla_\alpha \log \pi_{\alpha_c}(a^i_k \mid s^i_k), \qquad P_x(\boldsymbol{\xi},Y) = \nabla_x L(x;\boldsymbol{\xi},Y). \tag{226}$$

These are unbiased for f_α, f_x and satisfy $\|P_\alpha\| \le B_L I K M_1$, $\|P_x\| \le G_L$. Neither is divided by c.

In Algorithm 2, a capped geometric draw means drawing N independently with $\Pr(N = n) = (1-\gamma)\gamma^n$ and using a zero sample if $N \ge H$. For $N < H$, an equal-mixture prefix chooses $J \sim \mathrm{Unif}\{0, c\}$, resets from ρ, and follows π_{α_J} for N transitions and a terminal action. This convention does not condition on $N < H$; such conditioning would change the field normalization. The three blocks below define LowerPair, OuterPair, and TDPair in Algorithm 1.

Algorithm 2 One iteration: explicit sampling and persistent paired TD

Require: Frozen $p = (x, U, S)$, $z = (u, v)$, stored certified $\overline{d} = \widehat{\rho}_n$; H, $B_H, B_V, B_C, B_{f,a}, B_{f,x}$ and the main algorithm's steps and radii.
Ensure: New (x^+, U^+, S^+, z^+) or failure; all oracle arguments are old.
Set $\alpha_0 = U$, $\alpha_c = U + cS$, $\theta_0 = u$, $\theta_c = u + cv$.
for $b = 1, \dots, B_H$ **do**
  Independently draw geometric N_b, with $\Pr(N_b = n) = (1-\gamma)\gamma^n$.
  if $N_b \geq H$ **then**
    Set $(a_b, d_b) = (0, 0)$.
  else
    Draw an equal-mixture reset prefix ξ_b with N_b transitions and terminal action.
    Compute $(a_b, d_b) = (a(\xi_b), d(\xi_b))$ using both old critics.
  end if
end for
Independently draw $B_{f,a}$ preference tuples and labels at old α_c; let $\widehat{P}_\alpha$ average P_α.
$\widehat{a} = B_H^{-1} \sum_b a_b$, $\widehat{R} = B_H^{-1} \sum_b d_b + \widehat{P}_\alpha$; $\widehat{C} = \widehat{a} + c\widehat{R}$.
Use Definition C.32 to compute $U^+ = \widetilde{\mathsf{M}}(U, \widehat{a})$; $\alpha_c^+ = \widetilde{\mathsf{M}}(\alpha_c, \widehat{C})$; $S^+ = (\alpha_c^+ - U^+)/c$.
for $b = 1, \dots, B_V$ **do**
  Independently draw a fresh capped geometric equal-mixture prefix at the old actor pair.
  Set $e_b = 0$ for a capped draw and $e_b = e(\xi_b)$ otherwise.
end for
Independently draw $B_{f,x}$ preference tuples and labels at old α_c; let $\widehat{P}_x$ average P_x.
$\widehat{V} = B_V^{-1} \sum_b e_b + \widehat{P}_x$; $x^+ = x - \eta_x \widehat{V}$.
for $b = 1, \dots, B_C$ **do**
  Independently of both evaluation blocks, draw geometric N_b.
  if $N_b \geq H$ **then**
    Set $F_b = 0$.
  else
    Reset from ρ and follow π_U for N_b transitions to s; draw $a \sim \pi_U(\cdot \mid s)$, $s' \sim P(\cdot \mid s, a)$.
    Evaluate $F_b = F^{\mathrm{samp}}(p, z; s, a, s')$ from (222) using old p, z.
  end if
end for
$\widehat{F} = B_C^{-1} \sum_b F_b$; $z^+ = \Pi^W_{\mathcal{K}_Q}(z - \beta \widehat{F})$.
if $\mathsf{Safe}(U^+, S^+)$ fails **then**
  return failure; Algorithm 1 returns the fixed x_0.
end if
return (x^+, U^+, S^+, z^+).

Mean identities and truncation. If a scalar test function v is integrable, geometric prefix sampling gives $\mathbb{E}[v(s_N, a_N)] = \mathbb{E}_{\nu_\alpha} v$. Differentiating the discounted Bellman equation for Q^{true} and multiplying by the normalized reset resolvent yields

$$\begin{aligned} g_\alpha(x,\alpha) &= \mathbb{E}_{\nu_\alpha}[\nabla_\alpha \log \pi_\alpha \, Q^{\mathrm{true}}_{x,\alpha} + \partial_\alpha q_{x,\alpha}], \\ g_x(x,\alpha) &= \mathbb{E}_{\nu_\alpha}[-\nabla_x r_x]. \end{aligned} \tag{227}$$

Indeed, differentiating $Q = q + \gamma K_\alpha Q$ first accounts for the change in the current action law, the criterion and all future action laws. The resolvent collects these score terms over discounted occupancy; the lower normalization cancels the reset-resolvent factor. Fixed dynamics make the second identity immediate. No derivative is taken through learned critic coordinates. With zero output for $N \geq H$, the TD tail is

$$\|\mathbb{E}_t \widehat{F}_{t,H} - F(p_t, z_t)\|_W \leq C_{\mathrm{TD}} \gamma^H = b_{\mathrm{TD}}. \tag{228}$$

This is truncation, not conditioning on $N < H$.

Common-prefix balance weights. For fixed prefix length n (including $n+1$ action factors), let p^0, p^c have reset law ρ, common dynamics and endpoint policies; let $q = (p^0 + p^c)/2$ and $w_j = dp^j/dq$. Selecting the endpoint once per trajectory implements precisely this mixture.

Lemma C.41 (Balance-weight identities)**.**

$$0 \leq w_0, w_c \leq 2, \qquad w_0 + w_c = 2 \quad q\text{-a.s.}, \qquad \mathbb{E}_q[w_j Z] = \mathbb{E}_{p^j}[Z], \quad j \in \{0, c\}. \tag{229}$$

Proof. Since $q \geq p^j/2$, $w_j \leq 2$. The identity $w_0 + w_c = 2$ follows from $p^0 + p^c = 2q$, and the expectation identity is the Radon–Nikodym change-of-measure formula. □

Lemma C.42 (Divided balance-weight difference is uniform in c)**.** *Let* $r(\xi) = \log(dp^c/dp^0)(\xi)$. *Then*

$$w_c - w_0 = 2\tanh(r/2) \tag{230}$$

and

$$\left| \frac{w_c(\xi) - w_0(\xi)}{c} \right| \leq (n+1) M_{\log,1,L} \, \|S\| \, . \tag{231}$$

Proof. Writing $z = dp^c/dp^0 = e^r$ gives $w_c = 2z/(1+z)$ and $w_0 = 2/(1+z)$, hence (230). The common initial law and common environment dynamics cancel from the path likelihood ratio, leaving

$$r(\xi) = \sum_{h=0}^{n} \left[\log \pi_{U+cS}(a_h \mid s_h) - \log \pi_U(a_h \mid s_h) \right].$$

The integral mean-value formula and Lemma C.8 give $|r| \leq c(n+1) M_{\log,1,L} \, \|S\|$. Since $|\tanh z| \leq |z|$, (231) follows. □

Lemma C.43 (Paired evaluation errors, tails and moments)**.** *For* $z^* = Z(p)$ *and* $e_j = Q_{m_Q,\theta_j} - Q^{\mathrm{true}}_{x,\alpha_j}$,

$$\|e_0\|_\infty^2 + \|(e_c - e_0)/c\|_\infty^2 \leq C\big(\|z - z^*\|_W^2 + \mathcal{A}_Q\big). \tag{232}$$

The base and normalized lower evaluation biases are therefore bounded in squared norm by $C(\|z - z^*\|_W^2 + \mathcal{A}_Q + b_{\mathrm{ev}}^2)$. *The outer bias is bounded by* b_{ev} *without critic error. For some fixed integer* ℓ, $b_{\mathrm{ev}} \leq C(H+1)^\ell \gamma^H$. *Every fixed-order conditional moment of the sampled directions is bounded independently of* H, c, m_Q, x*; their samplewise envelopes are polynomial in* H.

Proof. Use the integral

$$\frac{Q_{m_Q,u+cv} - Q_{m_Q,u}}{c} = \int_0^1 \nabla Q^\top_{m_Q,u+tcv} v \, dt.$$

The first two critic derivative bounds make this map Lipschitz in (u, v), uniformly in c on $\mathcal{K}_Q$. The base value is also Lipschitz. Subtract their values at z^* and use (54) to obtain (232).

For lower score-times-value errors, set $T_j = \nabla \log \pi_{\alpha_j} e_j$. Both the scores and their normalized endpoint differences are bounded. Decompose before dividing:

$$\frac{w_c T_c - w_0 T_0}{c} = w_c \frac{T_c - T_0}{c} + \frac{w_c - w_0}{c} T_0,$$
$$\frac{T_c - T_0}{c} = \nabla \log \pi_{\alpha_c} \frac{e_c - e_0}{c} + \frac{\nabla \log \pi_{\alpha_c} - \nabla \log \pi_{\alpha_0}}{c} e_0.$$

Lemmas C.41–C.42 bound this by $C(n+1)(\|z - z^*\|_W + \sqrt{\mathcal{A}_Q})$. The direct criterion derivatives introduce no critic error. Taking geometric expectations proves the untruncated bias bound. The outer sample in (225) uses only r_x derivatives.

On the projection set, base learned values and normalized differences are bounded even when inaccurate. Thus each untruncated direction is bounded by $C(n+1)^\ell$ for a fixed degree. Every moment $\sum_{n\geq 0}(1-\gamma)\gamma^n(n+1)^j$ is finite. The removed tail is at most $C(H+1)^\ell \gamma^H$, by writing $n = H + k$ and summing the geometric polynomial. Capping supplies the samplewise polynomial envelope; it does not worsen any fixed-order moment. Preference terms are bounded by Proposition C.44. For TD, the integral identity (52) bounds the sample uniformly, and its omitted tail proves (228). □

Direct preference samples.

Proposition C.44 (Unbiased bounded tuple gradients)**.** *For a tuple and label sampled at any frozen actor α, use P_x, P_α from (226), with α_c replaced by α. Conditionally on every pre-sampling history that fixes (x, α),*

$$\mathbb{E}[P_x \mid x, \alpha] = f_x(x, \alpha), \qquad \mathbb{E}[P_\alpha \mid x, \alpha] = f_\alpha(x, \alpha). \tag{233}$$

For every realized sample,

$$\|P_x\| \leq G_L, \qquad \|P_\alpha\| \leq B_L I K M_1. \tag{234}$$

Averages $\widehat{F}_x, \widehat{F}_\alpha$ of respectively $B_{f,x}, B_{f,a}$ independent tuple–label samples have conditional variance at most $G_L^2/B_{f,x}$ and $(B_L I K M_1)^2/B_{f,a}$.

Proof. The joint tuple–label law factors into the I trajectory laws and $\mathcal{D}$. Fixed initial law and dynamics and the parameter-independent label kernel imply that its actor log derivative is the score sum $\sum_{i,k} \nabla_\alpha \log \pi_\alpha(a_k^i \mid s_k^i)$; its x derivative is zero. Differentiating under the finite expectation is justified by the displayed uniform envelopes. Since the sample loss has no explicit actor dependence, this gives (233). The triangle inequality gives a score-sum bound of IKM_1, proving (234). Conditional independence and centering divide the second-moment bounds by the batch size. These statements concern each sampled label, rather than only a label-averaged population loss. □

Batch sizes and timing. Fix positive preference batch sizes $B_{f,a}, B_{f,x}$ independently of accuracy. One may take any positive TD/evaluation batch sizes B_C, B_H, B_V bounded by a fixed polynomial in H, including fixed sizes: the geometric moment bounds already give the uniform moments through order eight used in the proof. Averaging independent samples preserves those bounds and reduces centered second moments. Conditional on $\mathcal{F}_t$, each evaluation stage has its own independent tuple/label block, and TD is independent of both evaluation stages. The exact identity $\widehat{C} = \widehat{a} + c\widehat{R}$ is samplewise. After conditioning further on $\mathcal{G}_t$, the outer block still has its old-parameter conditional mean. No lower labels are reused for the outer estimator, and the critic update is first used at iteration $t+1$. The full interaction ledger is (75).

C.7 Conditional tracking estimates

The results here use the derived trained-critic bias from Lemma B.8, the residual-controlled finite solves, and the pre-draw history $\mathcal{F}_t$ defined in Appendix B. They do not introduce an independent oracle-accuracy assumption.

Lemma C.45 (Conditional means and moments of the simultaneous increment)**.** *Write $\Delta p_t = p_{t+1} - p_t = d_t + w_{t+1}$, where $d_t = \mathbb{E}_t \Delta p_t$ and $\mathbb{E}_t w_{t+1} = 0$. For the prospective step,*

$$\|d_t\|^2 \leq C\{(\eta_y^2 + \eta_x^2)(\mathcal{T}_t + \mathcal{C}_t + \omega) + \eta_x^2 X_t + (\eta_y^2 + \eta_x^2)^2\},$$
$$\mathbb{E}_t\|w_{t+1}\|^2 \leq C(\eta_y^2 + \eta_x^2), \qquad \mathbb{E}_t\|\Delta p_t\|^4 \leq C(\eta_y + \eta_x)^4. \tag{235}$$

The combined reference-map and finite-solver remainders have conditional first, second and fourth moments $O(\eta_y^2)$, $O(\eta_y^4)$ and $O(\eta_y^8)$, respectively. The entire Δp_t is conditionally independent of the current TD block. All constants are uniform over subsequently smaller step ratios.

Proof. For each sampled direction decompose, for example, $\widehat{a} = a + e_a + \zeta_a$, where $e_a = \mathbb{E}_t\widehat{a} - a$ and $\mathbb{E}_t\zeta_a = 0$; use the same decomposition for $\widehat{R}, \widehat{V}$. The inverse metrics P^0, P^c and correction K_c and the full metric sample are fixed under $\mathcal{F}_t$. Lemma C.25 and the samplewise bounds in Lemma C.35 give

$$\begin{aligned}\Delta U &= -\eta_y P^0\widehat{a} + \rho^U + e_U,\\ \Delta S &= -\eta_y P^c\widehat{R} + \eta_y K_c\widehat{a} + \rho^S + e_S,\\ \Delta x &= -\eta_x\widehat{V}.\end{aligned}$$

The reference-map remainders are η_y^2 times polynomials of degree at most two in the directions, with bounded S, c. Direction moments through order eight from Lemma B.8 bound their means, second moments and fourth moments. Moreover, $\|(e_U, e_S)\| \le \sqrt{5}\eta_y^2$ samplewise, so their means and centered parts preserve these moment orders. Centered first-order terms have zero conditional mean. Thus the population field bounds (43) and the squared bias bound (56) give the first line of (235). The bounded second moments of all directions and the remainder estimates give its second line; the same argument using fourth moments of the increment gives the last bound. No independence between the base and scaled lower estimates is needed.

The two finite actor updates, including their stopping indices, are measurable functions of the lower block and stored metric sample, while the upper update is a function of its separate outer block. Both use old parameters. The TD block is independent of this pair of blocks under $\mathcal{F}_t$. Hence revealing TD first preserves the centering of w_{t+1}, exactly as used in the moving-target critic proof. Conditional fourth moments are finite without any bound on upper iterates, because the reward and loss envelopes are global. □

Lemma C.46 (Conditional Taylor transfer for actor tracking)**.** *Let $d_U^{\mathrm{pop}} = -P^0 a$ and $d_S^{\mathrm{pop}} = -P^c R_c + K_c a$. On a prospective step,*

$$\begin{aligned}\mathbb{E}_t\Delta\mathcal{T}_t \le{}& \eta_y\langle\nabla_{U,S}\mathcal{T}_t, (d_U^{\mathrm{pop}}, d_S^{\mathrm{pop}})\rangle\\ &- \eta_x\langle\nabla_x\mathcal{T}_t, V_c(p_t)\rangle + C(\eta_y+\eta_x)\sqrt{\mathcal{T}_t}\sqrt{\mathcal{C}_t+\omega} + C(\eta_y^2+\eta_x^2).\end{aligned} \tag{236}$$

Proof. Apply the quadratic Taylor formula in all variables (x, U, S) along the segment from p_t to $p_t + \Delta p_t$. Lemma C.19 bounds the full Hessian there, including mixed actor–upper blocks. Lemma C.45 therefore bounds the expected quadratic remainder by $C(\eta_y^2 + \eta_x^2)$. Substitute the conditional mean of each update from its paired expansion. Centered first-order noise vanishes because the old gradient and metric are $\mathcal{F}_t$-measurable. The conditional means of the combined reference-map and numerical remainders cost $C\eta_y^2$, since the old potential gradient is bounded. The remaining bias products are bounded by $C(\eta_y+\eta_x)\|\nabla\mathcal{T}_t\|\sqrt{\mathcal{C}_t + \mathcal{A}_Q + b_{\mathrm{ev}}^2}$. The gradient self-bound gives the displayed inequality. This calculation also covers the lower-first implementation order: all fields remain functions of the old parameters, and fresh outer data retain their conditional mean after revealing $\mathcal{G}_t$. No conditional expectation is taken with respect to a future non-exit event. □

C.8 Analytic and probabilistic tools

All segmentwise statements below are used on the larger prospective actor neighborhood. Short Young and quadratic Taylor inequalities are applied directly in the main proofs.

Lemma C.47 (Covariance transfer under domination)**.** *If $P \succeq qQ$ for probability measures P, Q and $q \in (0, 1]$, then for every square-integrable vector X,*

$$\mathrm{Cov}_P(X) \succeq q\,\mathrm{Cov}_Q(X).$$

Proof. For a vector v, write $Y = v^\top X$. The variational identity $\mathrm{Var}_P(Y) = \inf_{a\in\mathbb{R}}\mathbb{E}_P[(Y-a)^2]$ gives

$$\mathrm{Var}_P(Y) \ge q\inf_a \mathbb{E}_Q[(Y-a)^2] = q\,\mathrm{Var}_Q(Y).$$

Apply the scalar conclusion to every v. □

Lemma C.48 (Conditional centering is an L_2 contraction)**.** *On a joint state-action probability space, let* $\Pi X = X - \mathbb{E}[X \mid s]$*. Then*

$$\|\Pi X\|_{L_2} \le \|X\|_{L_2}, \qquad \|\Pi X - \Pi Y\|_{L_2} \le \|X - Y\|_{L_2}.$$

Proof. Conditional expectation is the orthogonal projection onto the closed subspace of s-measurable random variables. Hence Π is the complementary orthogonal projection and has operator norm one. □

Lemma C.49 (Inverse perturbation)**.** *If A and B are invertible matrices with $\left\|A^{-1}\right\|_{\mathrm{op}}, \left\|B^{-1}\right\|_{\mathrm{op}} \le \mu^{-1}$, then*

$$\left\|A^{-1} - B^{-1}\right\|_{\mathrm{op}} \le \mu^{-2} \|A - B\|_{\mathrm{op}}.$$

Proof. Use $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$. □

Lemma C.50 (Gradient bounds from strong convexity and smoothness)**.** *Let H be differentiable, μ-strongly convex, and have L_H-Lipschitz gradient on a convex set. Suppose its minimizer $z^\star$ is interior, so $\nabla H(z^\star) = 0$. Then*

$$\frac{\mu}{2}\|z - z^\star\|^2 \le H(z) - H(z^\star), \tag{237}$$

$$2\mu\big(H(z) - H(z^\star)\big) \le \|\nabla H(z)\|^2, \tag{238}$$

$$\|\nabla H(z)\|^2 \le \frac{2L_H^2}{\mu}\big(H(z) - H(z^\star)\big). \tag{239}$$

Proof. Strong convexity at $z^\star$ gives (237). Strong convexity at z gives

$$H(z^\star) \ge H(z) + \langle \nabla H(z), z^\star - z\rangle + \frac{\mu}{2}\|z - z^\star\|^2.$$

Maximizing the right-hand upper bound on $H(z) - H(z^\star)$ over the displacement yields $H(z) - H(z^\star) \le \|\nabla H(z)\|^2/(2\mu)$, proving (238). Finally, $\|\nabla H(z)\| \le L_H\|z - z^\star\|$ and (237) give (239). □

Lemma C.51 (Bounded innovations despite critic projection)**.** *There is a polynomial envelope $B_D(H)$ such that, at pre-exit steps including the prospective first exit,*

$$|N_{t+1}| \le \eta_y B_D(H), \qquad N_{t+1} = \mathcal{D}_{t+1} - \mathbb{E}_t\mathcal{D}_{t+1}. \tag{240}$$

No differentiability or inactivity of the critic projection is needed.

Proof. Since $z_t \in \mathcal{K}_Q$, nonexpansiveness gives $\|z_{t+1} - z_t\|_W \le \beta\left\|\widehat{F}_{t,H}\right\|_W \le C\beta$. Uniform target derivatives and (44) give $\|Z(p_{t+1}) - Z(p_t)\|_W \le C\eta_y\,\mathrm{poly}(H)$. Both old and new error vectors have bounded norm because the critic iterate and prospective target lie in $\mathcal{K}_Q$. Hence

$$|\mathcal{C}_{t+1} - \mathcal{C}_t| \le C[\beta + \eta_y\,\mathrm{poly}(H)].$$

Multiplication by λr_Q gives $C\eta_y\,\mathrm{poly}(H)$, using $r_Q\beta = \eta_y$ and $r_Q \le 1$. Uniform first derivatives of the actor potential similarly give $|\mathcal{T}_{t+1} - \mathcal{T}_t| \le C\eta_y\,\mathrm{poly}(H)$. Centering doubles the resulting deterministic bound. Relative to the virtual exact step, Proposition C.37 changes the next $\mathcal{D}$ by at most $C\eta_y^2$ samplewise, and centering changes its innovation by at most $2C\eta_y^2$. This is absorbed by enlarging $B_D(H)$, without a c^{-1} factor. The argument is valid for all sufficiently small absolute steps on the reserved prospective buffers. □

C.9 Assumption verification and the PL counterexample

A1's actor derivative requirement is verified by Theorem C.4; the fixed bounded input, normalized initialization and fixed chart are part of that construction. A2 is an identifiability requirement with the optional independent pilot in Proposition C.13. Its local persistence does not imply global nondegeneracy for unrestricted neural parameters. For G, Proposition C.14 and the bounded tanh reward example in Appendix C.2 give explicit global envelopes. Neither requires a correctly specified fitted preference model.

For C1, a feature Gram gap under the base reset law ρ_U is sufficient by $\nu_U \succeq (1-\gamma)\rho_U$. This concerns fixed critic features, not the actor tangent features. A sufficient structural certificate for C2 is that the feature span contains r_x and the log-policy-ratio criterion separately and is invariant under every admitted K_α. The Neumann series for $(I-\gamma K_\alpha)^{-1}$ then stays in this finite-dimensional closed span. Finite state–action indicator features provide one such realization when their occupancy Gram matrix has a uniform gap. For continuous states the same span invariance and Gram gap are substantive restrictions. The theorem requires C2 at the selected fixed τ; increasing τ or m_Q alone does not establish it. Even under exact C2, the finite nonlinear critic retains the derived $O(m_Q^{-2})$ squared paired approximation floor in (54).

What generic policy smoothness does not imply.

Proposition C.52 (Smooth actors and exact critics do not imply uniform PL)**.** *Consider a one-state self-loop MDP with actions* $0, 1$*, rewards* $r(1) = 1$ *and* $r(0) = 0$*, and the actor*

$$\pi_\theta(1) = \tfrac{1}{2} + \tfrac{1}{4}\sin\theta, \qquad \pi_\theta(0) = 1 - \pi_\theta(1), \qquad \theta \in [-\pi, \pi]. \tag{241}$$

The actor probabilities are uniformly positive and all required score derivatives are bounded. A tabular critic can represent the action-value function exactly. Nevertheless the normalized discounted return fails every uniform maximization PL inequality with a positive constant.

Proof. The one state is visited at every step, so the normalized return is

$$J(\theta) = (1-\gamma)\sum_{t\geq 0}\gamma^t \pi_\theta(1) = \tfrac{1}{2} + \tfrac{1}{4}\sin\theta.$$

Its maximum is $J^* = 3/4$ at $\theta = \pi/2$, whereas

$$J'(-\pi/2) = \tfrac{1}{4}\cos(-\pi/2) = 0, \qquad J^* - J(-\pi/2) = \tfrac{1}{2}.$$

Thus $\frac{1}{2}\left\|\nabla J(\theta)\right\|^2 \geq \mu(J^* - J(\theta))$ fails for every $\mu > 0$. Equivalently, the minimization PL inequality fails for $g = -J$. Since both action probabilities lie in $[1/4, 3/4]$, their log derivatives are smooth and bounded on the compact domain. The ordinary discounted action values are $Q_\theta(a) = r(a) + \gamma\pi_\theta(1)/(1-\gamma)$, which an exact two-entry tabular critic represents. □

The example establishes that generic actor smoothness, positive probabilities, and critic realizability do not *imply* uniform PL. It does not contradict a theorem that imposes PL separately. It is not a witness that our entire actor/assumption class strictly contains Gaur et al.'s class.

C.10 KL-regularized counterexamples and degenerate neural policies

The example in Proposition C.52 concerns the unregularized return. We now give an additional smooth policy with a spurious strict local minimum for every positive KL-regularization level, and explain a related obstruction in multilayer neural parameterizations.

C.10.1 Setting and the PL condition

Consider a discounted MDP with finite action set $\mathcal{A}$, a strictly positive reference policy π_{ref}, and a parametric policy family $\{\pi_\theta : \theta \in \Theta\}$. For an upper parameter x and $\tau \geq 0$, the normalized lower cost is

$$g_\tau(x,\theta) = (1-\gamma)\,\mathbb{E}_{\pi_\theta}\sum_{t\geq 0}\gamma^t\Big[-r_x(s_t,a_t) + \tau\log\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\mathrm{ref}}(a_t \mid s_t)}\Big], \tag{242}$$

as in (2), so the regularizer is $D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})$ averaged over the discounted occupancy. Write $g_\tau^\star(x) = \inf_{\theta\in\Theta} g_\tau(x,\theta)$.

A uniform lower PL condition with constant $\mu > 0$ on Θ requires, for every x in the upper domain and every $\theta \in \Theta$,

$$\tfrac{1}{2}\,\|\nabla_\theta g_\tau(x,\theta)\|^2 \;\geq\; \mu\,\big(g_\tau(x,\theta) - g_\tau^\star(x)\big). \tag{243}$$

Consequently, a single pair $(x, \bar\theta)$ with $\nabla_\theta g_\tau(x,\bar\theta) = 0$ and $g_\tau(x,\bar\theta) > g_\tau^\star(x)$ rules out (243) for every $\mu > 0$.

C.10.2 A SPURIOUS LOCAL MINIMUM FOR EVERY REGULARIZATION LEVEL

One-state reduction. Take one state with a self-loop, $\mathcal{A} = \{0,1\}$, $\pi_{\mathrm{ref}} = (\frac{1}{2},\frac{1}{2})$, and reward $r_x(1) = u(x)$, $r_x(0) = 0$. Write $p = \pi_\theta(1)$ and

$$k(p) = D_{\mathrm{KL}}\big(\mathrm{Bern}(p)\,\|\,\mathrm{Bern}(\tfrac{1}{2})\big) = p\log(2p) + (1-p)\log(2(1-p)).$$

The state is visited at every step, so (242) reduces to

$$g_\tau(x,\theta) = h_{u,\tau}(p(\theta)), \qquad h_{u,\tau}(p) = -u\,p + \tau\,k(p). \tag{244}$$

For $\tau > 0$, $h_{u,\tau}$ is strictly convex on $(0,1)$, with $h'_{u,\tau}(p) = -u + \tau\log\frac{p}{1-p}$.

The policy family. For $\theta \in \mathbb{R}$ (equivalently, on the circle), let

$$p(\theta) = \frac{1}{2} + \frac{\cos\theta + \frac{1}{2}\cos 2\theta}{6}. \tag{245}$$

Lemma C.53 (Properties of the family)**.** *The map p is C^∞ and 2π-periodic, and its range is exactly $[\frac{3}{8},\frac{3}{4}]$. Within each period it has a strict global maximum $p(0) = \frac{3}{4}$, strict global minima $p(\pm\frac{2\pi}{3}) = \frac{3}{8}$, and a strict local maximum $p(\pi) = \frac{5}{12}$. All log-policy derivatives of every order are bounded.*

Proof. Put $\varphi(\theta) = \cos\theta + \frac{1}{2}\cos 2\theta$. Then $\varphi'(\theta) = -\sin\theta\,(1 + 2\cos\theta)$, so the critical points in $(-\pi,\pi]$ are $\theta \in \{0,\pi,\pm\frac{2\pi}{3}\}$, with values $\varphi(0) = \frac{3}{2}$, $\varphi(\pi) = -\frac{1}{2}$ and $\varphi(\pm\frac{2\pi}{3}) = -\frac{3}{4}$. Since $\varphi''(\theta) = -\cos\theta - 2\cos 2\theta$, we get $\varphi''(0) = -3$, $\varphi''(\pi) = -1$ and $\varphi''(\pm\frac{2\pi}{3}) = \frac{3}{2}$. This gives the classification and the range $\varphi \in [-\frac{3}{4},\frac{3}{2}]$, hence $p \in [\frac{3}{8},\frac{3}{4}]$. Both action probabilities lie in $[\frac{1}{4},\frac{5}{8}] \cup [\frac{3}{8},\frac{3}{4}] \subset [\frac{1}{4},\frac{3}{4}]$, so $\log\pi_\theta$ is a smooth function of a trigonometric polynomial bounded away from the singularities of log, and all of its derivatives are bounded. □

Proposition C.54 (Spurious local minimum for every τ)**.** *Let $u \geq 0$ and $\tau > 0$. In the setting (244)–(245):*

(i) $\bar\theta = \pi$ is a strict local minimizer of $g_\tau(x,\cdot)$ with $\nabla_\theta g_\tau(x,\pi) = 0$;

(ii) its suboptimality gap satisfies

$$g_\tau(x,\pi) - g_\tau^\star(x) \;\geq\; \frac{u}{12} + \tau\,k\big(\tfrac{5}{12}\big) \;>\; 0; \tag{246}$$

(iii) consequently, (243) fails for every $\mu > 0$ on any parameter set containing π and a point θ with $p(\theta) = \frac{1}{2}$.

Here $k(\frac{5}{12}) = \frac{5}{12}\log\frac{5}{6} + \frac{7}{12}\log\frac{7}{6} \approx 0.01395$.

Proof. (i) By the chain rule, $\nabla_\theta g_\tau(x,\theta) = h'_{u,\tau}(p(\theta))\,p'(\theta)$, and $p'(\pi) = 0$ by Lemma C.53. At $p = \frac{5}{12}$, $h'_{u,\tau}(\frac{5}{12}) = -u + \tau\log\frac{5}{7} < 0$, because $u \geq 0$, $\tau > 0$ and $\log\frac{5}{7} < 0$. Differentiating once more at a critical point of p gives

$$\partial_\theta^2 g_\tau(x,\pi) = h''_{u,\tau}(p(\pi))\,p'(\pi)^2 + h'_{u,\tau}(p(\pi))\,p''(\pi) = h'_{u,\tau}(\tfrac{5}{12})\cdot\big(-\tfrac{1}{6}\big) > 0,$$

so π is a strict local minimizer.

(ii) The range of p contains $\frac{1}{2}$, so $g^\star_\tau(x) \le h_{u,\tau}(\frac{1}{2})$. Since $k(\frac{1}{2}) = 0$,

$$h_{u,\tau}(\tfrac{5}{12}) - h_{u,\tau}(\tfrac{1}{2}) = u\big(\tfrac{1}{2} - \tfrac{5}{12}\big) + \tau\big(k(\tfrac{5}{12}) - 0\big) = \frac{u}{12} + \tau\, k(\tfrac{5}{12}).$$

This quantity is positive because $k(\frac{5}{12}) > 0$ and $\tau > 0$.

(iii) Combine (i) and (ii) with the observation following (243). □

Remark C.55 (What the proposition adds). (a) *Regularization does not repair PL.* The lower bound in (246) is positive for every $\tau > 0$ and increases with τ. For $u > 0$ it is also bounded below by $u/12$ uniformly over $\tau > 0$. The bad point is a genuine local minimizer, not a local maximizer, so gradient methods can converge to it from a neighborhood.

(b) *Uniformity in the upper parameter.* The conclusion holds for every x with $u(x) \ge 0$, so it applies throughout any upper domain on which the reward scale is nonnegative, for example $u = \tanh$ on $[0, \infty)$.

(c) *Exact value representation.* Write $q(a) = -r_x(a) + \tau \log(\pi_\theta(a)/\pi_{\mathrm{ref}}(a))$ and $\bar{q} = \sum_a \pi_\theta(a) q(a)$. The regularized cost action values are $Q(a) = q(a) + \gamma\bar{q}/(1-\gamma)$, so a two-entry tabular critic represents them exactly, as in Proposition C.52.

(d) *The original example for every τ.* For the earlier family $p(\theta) = \frac{1}{2} + \frac{1}{4}\sin\theta$, the same reduction gives $\nabla_\theta g_\tau(x, \pm\frac{\pi}{2}) = 0$ for every τ. By symmetry of k about $\frac{1}{2}$, $g_\tau(x, -\frac{\pi}{2}) - g_\tau(x, \frac{\pi}{2}) = u/2$ for every τ. That example therefore also extends to all regularization levels when $u > 0$, but its bad point is a local maximizer.

C.10.3 Degenerate policy Jacobians and multilayer networks

Lemma C.56 (Vanishing policy Jacobian forces a critical point). *Let $\mathcal{A}$ be finite, the rewards bounded, and π_{ref} strictly positive. Suppose $R_\theta = \pi_\theta/\pi_{\mathrm{ref}}$ is differentiable at $\bar{\theta}$ as a map into bounded functions with the supremum norm, and that $0 < c \le R_\theta(s,a) \le C < \infty$ on a neighborhood of $\bar{\theta}$. If $D_\theta \pi_{\bar{\theta}}(a \mid s) = 0$ for every (s,a), then $\nabla_\theta g_\tau(x, \bar{\theta}) = 0$ for every x and finite $\tau \ge 0$. If this point is suboptimal for some x, uniform lower PL fails on any parameter set containing it and a strictly better point. The differentiability and ratio bounds hold for finite state spaces with smooth positive policies and for smooth bounded-input neural exponential tilts on bounded parameter neighborhoods.*

Proof. The pointwise derivative condition implies that the derivative of R_θ at $\bar{\theta}$ is the zero bounded function. As a function of R, the state transition operator is affine and bounded in the supremum norm. Its discounted resolvent $(I - \gamma T_\pi)^{-1}$ is Fréchet differentiable, by the resolvent identity and $\|(I - \gamma T_\pi)^{-1}\| \le (1-\gamma)^{-1}$. The expected one-step cost is

$$\sum_a \pi_{\mathrm{ref}}(a \mid s) R(s,a)[-r_x(s,a) + \tau \log R(s,a)],$$

which is differentiable in the same norm because R stays in $[c, C]$. Thus the normalized discounted objective is a differentiable function of R, and the chain rule gives a zero parameter gradient. A strictly positive objective gap then contradicts (243). □

Proposition C.57 (Zero hidden weights are degenerate for depth at least two). *Let σ be continuously differentiable with $\sigma(0) = 0$ (e.g. $\tanh$). For depth $L \ge 2$, arbitrary widths, bounded input $\psi(s,a)$ and fixed readout b, consider*

$$h^0 = \psi, \qquad h^\ell = \sigma\big(W_\ell h^{\ell-1}\big), \quad \ell = 1, \dots, L,$$
$$F_\theta(s,a) = b^{\mathsf{T}} h^L, \qquad \pi_\theta(a \mid s) \propto \pi_{\mathrm{ref}}(a \mid s)\, e^{F_\theta(s,a)}.$$

with $\theta = (W_1, \dots, W_L)$, and let $\bar{\theta}$ have $W_\ell = 0$ for all ℓ. Then $D_\theta F_{\bar{\theta}}(s,a) = 0$ and $D_\theta \pi_{\bar{\theta}}(a \mid s) = 0$ for all (s,a), and $\pi_{\bar{\theta}} = \pi_{\mathrm{ref}}$. The same holds for $L = 1$ if the readout b is also trainable and set to zero. Any normalization constants such as $m_{\ell-1}^{-1/2}$ do not change the conclusion.

Proof. At $\bar{\theta}$, $h^1 = \sigma(0) = 0$, and inductively $h^\ell = 0$ for all $\ell \ge 1$. Backpropagation gives $\partial F/\partial W_\ell = \delta_\ell\, (h^{\ell-1})^{\mathsf{T}}$, where $\delta_L = b \odot \sigma'(0)$ and $\delta_\ell = (W_{\ell+1}^{\mathsf{T}} \delta_{\ell+1}) \odot \sigma'(0)$ for $\ell < L$. For

$\ell \geq 2$, $h^{\ell-1} = 0$. For $\ell = 1$, δ_1 contains the factor $W_2^\top = 0$, which uses $L \geq 2$. Hence $D_\theta F_{\bar\theta} = 0$. Since $F_{\bar\theta} = 0$, $\pi_{\bar\theta} = \pi_{\mathrm{ref}}$, and $D_\theta \pi_{\bar\theta} = D_F \pi \circ D_\theta F_{\bar\theta} = 0$. For $L = 1$ with trainable $b = 0$, $\partial F/\partial b = h^1 = 0$ and $\partial F/\partial W_1 = (b \odot \sigma'(0))\psi^\top = 0$. □

Corollary C.58 (Uniform PL fails on domains containing zero hidden weights)**.** *In the setting of Proposition C.57, suppose that for some upper parameter x and some $\theta \in \Theta$, $g_\tau(x,\theta) < g_\tau(x,\bar\theta)$. Then uniform lower PL (243) fails on Θ for every $\mu > 0$ and every $\tau \geq 0$ for which this strict inequality holds.*

Proof. Combine Proposition C.57 with Lemma C.56. □

A concrete instance. Take a one-state bandit with $\mathcal{A} = \{-1, +1\}$, $\pi_{\mathrm{ref}} = \frac{1}{2}$, $r_x(a) = u(x)\, a$ with $u(x) > 0$, input $\psi(a) = a$, depth $L = 2$, width one, $\tanh$ activations and readout $b = 1$. Then $F_\theta(a) = \tanh(w_2 \tanh(w_1 a))$, which is odd in a, so

$$\pi_\theta(+1) = \sigma_{\log}\big(2\tanh(w_2 \tanh w_1)\big), \qquad \sigma_{\log}(v) = (1 + e^{-v})^{-1}.$$

Writing $q = \pi_\theta(+1)$, the cost is $\tilde h(q) = -u(2q-1) + \tau\, k(q)$, with $\tilde h'(\frac{1}{2}) = -2u < 0$. Taking $w_1 = w_2 = w > 0$ small gives $q \in (\frac{1}{2}, \frac{1}{2} + \delta)$ for small δ, hence $g_\tau(x,\theta) < \tilde h(\frac{1}{2}) = g_\tau(x,\bar\theta)$. So $\bar\theta = (0,0)$ is a critical point that is suboptimal for every x with $u(x) > 0$ and every $\tau \geq 0$. When $u < \tau$, the regularized optimum $q^\star = \sigma_{\log}(2u/\tau)$ is representable, and the gap equals $\tau \log\cosh(u/\tau)$.

Remark C.59 (Continuous actions). Lemma C.56 carries over to Gaussian policies $\mathcal{N}(\mu_\theta(s), \Sigma)$ with fixed covariance, provided the occupancy and cost are differentiable in the mean map. If the mean network has all hidden weights zero and depth at least two, then $D_\theta \mu_\theta = 0$ by the argument of Proposition C.57, and the same conclusion follows. This gives the analogous obstruction for fixed-covariance continuous-action policies under the stated differentiability condition.

C.10.4 Relation to the local assumptions

Remark C.60 (Why the local analysis avoids these points). (a) A vanishing policy Jacobian forces the centered score $Z - \mathbb{E}_\pi Z$ to vanish, so its conditional covariance is zero. For the zero-hidden-weight construction, $Z = D_\theta F$ itself also vanishes. Either case prevents an A2 certificate at that point. Assumption 4.2 requires A2 at the reference point and localizes the analysis within the radius on which the covariance gap persists (Appendix C.1). Degenerate points are therefore excluded from the certified neighborhood by construction.

(b) This also explains why the two-hidden-layer construction in Remark C.68 uses a nonzero second-layer initialization: all-zero hidden weights would give the degeneracy of Proposition C.57.

(c) In Proposition C.54, a neighborhood of a point with $p(\theta_0) = \frac{1}{2}$ and $p'(\theta_0) \neq 0$ is the analogue of the certified local region. For sufficiently large τ the unique local response lies near $p = \frac{1}{2}$ and away from $\theta = \pi$.

Remark C.61 (Scope of the claim). These results show that uniform lower PL is not implied by smoothness, positivity, exact value representation, or reference-KL regularization at any level. For the neural architectures in Proposition C.57, PL also fails on domains containing zero hidden weights whenever that point is suboptimal. A theorem that assumes PL separately, such as Gaur et al. (2025), is not contradicted by these examples. They do not establish containment between that assumption class and ours; the obstruction concerns domains that actually contain a suboptimal degenerate point.

C.11 A concrete continuous-state instance with a nontrivial response correction

We give a joint satisfiability example for A1, A2, G, and C1–C2 in a class with finitely many policy contexts. It has continuous state variation in rewards and transition probabilities, a nonlinear tied actor, and five fixed critic features. An attainable preference teacher produces a finite interior stationary point of the response-induced upper objective, at which the response-chain contribution is nonzero. The population illustration below evaluates this objective independently. Appendix D additionally implements sampled single-loop training on the same instance, using its two-hidden-layer actor specialization and the persistent nonlinear paired critic.

C.11.1 A CLASS WITH FINITE POLICY CONTEXTS AND BELLMAN CLOSURE

Proposition C.62 (A sufficient structural certificate for C1–C2)**.** *Let $\mathcal{S}$ be compact, $\mathcal{A}$ finite, and $\{\mathcal{S}_j\}_{j=1}^{M}$ a measurable partition. Write $j(s)$ for the cell index. Suppose the following conditions hold.*

(i) The fixed transition kernel has a finite-rank mixture $P(ds' \mid s, a) = \sum_{\ell=1}^{M'} w_\ell(s,a)\mu_\ell(ds')$, where the μ_ℓ are probability measures and the measurable weights are nonnegative and sum to one.

(ii) The tied actor uses the bounded, fixed input $\psi(s,a) = \psi_0(j(s),a)$, with fixed hidden depth and the normalized initialization, activation, and chart conditions of Appendix C.1. The reference policy depends only on $(j(s),a)$ and is strictly positive on every such pair.

(iii) The fixed feature map $\varphi_Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_Q}$ is bounded. Its span contains all cell–action indicators, all w_ℓ, and every bounded reward r_x.

(iv) For some $\lambda_0 > 0$, $\mathbb{E}_{\rho(ds)\pi_{\mathrm{ref}}(da|s)}[\varphi_Q\varphi_Q^{\mathsf{T}}] \succeq \lambda_0 I_{d_Q}$.

Then C2 holds for every finite $\tau > 0$, every x, and every finite actor parameter. On an ambient ball $\|\alpha\| \le A$, C1 holds with

$$\lambda_Q = (1-\gamma)c_A\lambda_0, \qquad c_A = \exp(-2C_{F,1}A),$$

where $C_{F,1}$ is a width-uniform bound on $\|D_\alpha F\|$ there and the tied construction has $F_0 = 0$.

Proof. At fixed parameters the neural score is constant on each cell–action pair. The finite-action log normalizer is therefore constant within each state cell. Both terms of $\log(\pi_\alpha/\pi_{\mathrm{ref}})$ belong to the cell-indicator span. The regularized cost $q_{x,\alpha} = -r_x + \tau\log(\pi_\alpha/\pi_{\mathrm{ref}})$ thus belongs to $V = \operatorname{span}\varphi_Q$.

For every bounded v,

$$K_\alpha v(s,a) = \sum_{\ell=1}^{M'} w_\ell(s,a)\int\sum_{a'}\pi_\alpha(a' \mid s')v(s',a')\,\mu_\ell(ds') \in V.$$

The Neumann series $\sum_{k\ge0}\gamma^k K_\alpha^k q_{x,\alpha}$ converges uniformly and stays in the closed finite-dimensional space V. This proves C2. On the ambient ball, $|F_\alpha| \le C_{F,1}A$ and $\pi_\alpha/\pi_{\mathrm{ref}} \ge c_A$. Therefore $\rho_U \ge c_A\rho\pi_{\mathrm{ref}}$, and $\nu_U \ge (1-\gamma)\rho_U$ transfers the reference Gram bound to C1. □

The bounded-feature condition is needed for C1; compactness alone does not bound an arbitrary measurable feature map. The normalized network conditions in the proposition supply A1(iii). Reward and preference regularity and A2 remain separate requirements, verified for the instance below. This proposition is a sufficient certificate within our broader assumption-based theorem.

Proposition C.63 (A log-normalizer obstruction under finite-rank dynamics)**.** *Let $\mathcal{S} = [-1,1]$, $\mathcal{A} = \{-1,+1\}$, $\pi_{\mathrm{ref}} = 1/2$, and $F_\alpha(s,a) = a\alpha s$, for α in an open interval containing zero. No fixed finite-dimensional subspace of bounded functions contains $\log(\pi_\alpha/\pi_{\mathrm{ref}})$ for every such α. If, in addition, the transition kernel has a fixed finite-rank mixture as in Proposition C.62, then, for every fixed $\tau > 0$ and fixed bounded reward, no fixed finite feature space can represent all corresponding regularized values.*

Proof. The log-policy ratio equals $a\alpha s - \log\cosh(\alpha s)$. If these functions belonged to a finite-dimensional space V, uniform finite differences and closedness would put all their α-derivatives at zero in V. Uniform convergence of these differences follows from smoothness on a compact neighborhood in (α, s).

Every even Taylor coefficient of $\log\cosh y$ is nonzero. For completeness, write $\tanh y = \sum_{k\ge1}(-1)^{k-1}a_k y^{2k-1}$ near zero. The identity $(\tanh y)' = 1 - \tanh^2 y$ gives $a_1 = 1$ and $(2k-1)a_k = \sum_{j=1}^{k-1} a_j a_{k-j} > 0$ for $k \ge 2$. Integration gives nonzero coefficients $(-1)^{k-1}a_k/(2k)$ in $\log\cosh y$. Thus V would contain every s^{2k}, $k \ge 1$, contradicting their linear independence.

For the value claim let $W = \operatorname{span}\{w_\ell\}_{\ell=1}^{M'}$. The finite-rank kernel implies $K_\alpha v \in W$ for every bounded v. If all Q_α belonged to a fixed finite-dimensional V, the Bellman equation would imply

$$\tau \log(\pi_\alpha/\pi_{\mathrm{ref}}) = Q_\alpha - \gamma K_\alpha Q_\alpha + r \in V + W + \operatorname{span}\{r\},$$

contradicting the first part. □

Remark C.64 (The obstruction is not a necessity theorem for finite contexts). The value conclusion above uses the finite-rank transition condition; failure of log-ratio membership alone does not imply failure of C2 for an arbitrary policy-dependent Bellman operator. Nor does this binary-logit example characterize every finite-action policy family. For instance, with three actions, let $q(s) = 1/2 + s/4$ and $\pi_{\mathrm{ref}}(s) = (q(s)/2, (1-q(s))/2, 1/2)$. Take $v = (1, 1, -1)$ and $F_\alpha(s, a) = h(\alpha)v_a$. The log normalizer is $\log\cosh h(\alpha)$, independent of s, so the log-policy ratio lies in $\operatorname{span}\{1, v\}$ although the policy varies continuously with s. With reset dynamics $P(\cdot \mid s, a) = \rho$, reward $\tanh(x)sv_a$, and features $(1, v_a, sv_a)$, C2 follows immediately. Under uniform ρ the reference Gram matrix is $\operatorname{diag}(1, 1, 1/3)$, so C1 also holds on bounded actor neighborhoods. Finite policy contexts are consequently a useful sufficient construction, not an additional necessary restriction on the general theorem.

Remark C.65 (Retained dimension and exact A2 calculation). For a cell-constant initialization feature Z_0, each conditional covariance has rank at most $|\mathcal{A}| - 1$. Hence

$$\operatorname{rank} \mathcal{G}_0 \le M(|\mathcal{A}| - 1), \qquad \text{A2 implies } r \le M(|\mathcal{A}| - 1).$$

When the reset cell masses are known, the certificate is the explicit finite sum

$$\mathcal{G}_0 = \sum_{j=1}^{M} \rho(\mathcal{S}_j) \operatorname{Cov}_{a \sim \pi_{\mathrm{ref}}(\cdot|j)}(Z_0(j, a)).$$

No pilot samples are required to evaluate this sum for the example below. If the cell masses are unknown, they must instead be bounded or estimated; finite encoding alone does not remove that statistical requirement. The example attains the dimension bound: $r = 2 = M(|\mathcal{A}| - 1)$.

C.11.2 The controlled process and preference model

Fix any discount factor $\gamma \in (0, 1)$. Set

$$\mathcal{S} = [-1, 1], \qquad \mathcal{A} = \{-1, +1\}, \qquad \rho = \mathrm{Unif}[-1, 1].$$

Use the Borel sigma algebra on $\mathcal{S}$. Define

$$I_- = [-1, 0), \qquad I_+ = [0, 1], \qquad i(s) = \begin{cases} -1, & s < 0, \\ +1, & s \ge 0, \end{cases}$$

and let μ_- and μ_+ be the uniform probability measures on I_- and I_+, respectively. The endpoint convention is immaterial for these measures and fixes all functions pointwise. Define the transition kernel

$$P(ds' \mid s, a) = \left(\frac{1}{2} + \frac{as}{4}\right) \mu_+(ds') + \left(\frac{1}{2} - \frac{as}{4}\right) \mu_-(ds'). \tag{247}$$

Both mixture weights lie in $[1/4, 3/4]$. Thus this is a measurable probability kernel with continuous state distributions. Its mixture weights depend continuously on s and depend on the action. The reset law is $\rho = (\mu_- + \mu_+)/2$.

The upper parameter is $x \in \mathbb{R}$. Write $u(x) = \tanh x$ and set

$$r_x(s, a) = u(x)\, as. \tag{248}$$

Only the reward model and the fitted preference loss depend on x; the transition kernel, reset law, and reference policy remain fixed.

For the upper objective, independently sample two trajectories ξ^1, ξ^2, each with $K = 2$ state–action pairs, using the current actor, the reset law, and (247). Put

$$D(\boldsymbol{\xi}) = \sum_{t=0}^{1} (a_t^1 s_t^1 - a_t^2 s_t^2), \qquad \Delta_x = u(x) D(\boldsymbol{\xi}),$$

$$\mathbb{P}(Y = 1 \mid \boldsymbol{\xi}) = \sigma(\kappa D(\boldsymbol{\xi})), \qquad \sigma(v) = (1 + e^{-v})^{-1}, \tag{249}$$

$$L(x; \boldsymbol{\xi}, Y) = \log(1 + e^{\Delta_x}) - Y \Delta_x. \tag{250}$$

Fix $\kappa \in (0,1)$, with $\kappa = 1/2$ in the numerical illustration. Here the teacher reward is the fixed function $r^\dagger(s,a) = \kappa as$. Conditional on the trajectories, the label law has no dependence on x or on the actor parameter. The loss has no additional explicit actor dependence. Fresh tuples and labels can be drawn independently.

C.11.3 A TIED NEURAL ACTOR WITH AN EXPLICIT INITIALIZATION GAP

Use $\pi_{\mathrm{ref}}(a \mid s) = 1/2$. For every integer $n \geq 1$ and retained parameter $\alpha = (\alpha_-, \alpha_+) \in \mathbb{R}^2$, define

$$h_n(t) = \sqrt{n}\,\tanh\left(\frac{t}{2\sqrt{n}}\right), \qquad F_{n,\alpha}(s,a) = a h_n(\alpha_{i(s)}), \tag{251}$$

$$\pi_{n,\alpha}(a \mid s) = \frac{\exp\{a h_n(\alpha_{i(s)})\}}{2\cosh(h_n(\alpha_{i(s)}))}. \tag{252}$$

Write $\pi_{n,\alpha}(a \mid i)$ for this region-constant probability. The two regions have independently trainable action preferences. In particular, the policy is state dependent whenever $h_n(\alpha_-) \neq h_n(\alpha_+)$.

Lemma C.66 (Realization by our tied architecture)**.** *The score in (251) is an exact one-hidden-layer specialization of the normalized antithetic actor, with input dimension 2, hidden width $2n$, and retained dimension 2. It satisfies the actor part of A1 uniformly in n and satisfies A2 with $\lambda_{\mathrm{eff},L} = 1/8$.*

Proof. Let e_-, e_+ be the coordinate vectors in $\mathbb{R}^2$ and use the bounded input $\psi(s,a) = a e_{i(s)}$. Index the $2n$ hidden units by (i,k), where $i \in \{-1,+1\}$ and $k = 1,\dots,n$. For a $2n$-by-2 weight matrix W, use the normalized branch

$$G_{n,W}(s,a) = \frac{1}{\sqrt{2n}} \mathbf{1}_{2n}^{\mathsf{T}} \tanh\left(\frac{W\psi(s,a)}{\sqrt{2}}\right).$$

The activation acts coordinatewise and the readout is fixed. Initialize $W_0 = 0$ and define the fixed chart B_n by

$$(B_n\alpha)_{(i,k),j} = \frac{\alpha_i}{\sqrt{n}}\,\mathbf{1}_{\{j=i\}}.$$

Then $\|B_n\alpha\|_F^2 = \alpha_-^2 + \alpha_+^2$, so B_n is an isometry. Take the two branches $W^\pm(\alpha) = \pm B_n\alpha/\sqrt{2}$ and their antithetic score $F = (G_{n,W^+} - G_{n,W^-})/\sqrt{2}$. At an input in region i, exactly n units have nonzero preactivation, equal to $\pm a\alpha_i/(2\sqrt{n})$. Oddness of $\tanh$ gives

$$G_{n,W^\pm}(s,a) = \pm a\sqrt{\frac{n}{2}}\,\tanh\left(\frac{\alpha_i}{2\sqrt{n}}\right),$$

and therefore gives exactly (251). The initialization satisfies the normalized spectral requirement, and all hidden weights vary through the specified fixed chart.

For $k = 1,\dots,4$,

$$h_n^{(k)}(t) = 2^{-k} n^{(1-k)/2} \tanh^{(k)}\left(\frac{t}{2\sqrt{n}}\right). \tag{253}$$

All derivatives of $\tanh$ in this display are bounded on $\mathbb{R}$. Since $n^{(1-k)/2} \leq 1$, the required operator norms of $D_\alpha^k F_{n,\alpha}$ are bounded independently of n, states, actions, and points in any fixed ambient ball. Also $F_{n,0} = 0$, $D_\alpha^2 F_{n,0} = 0$, and $h_n'(0) = 1/2$. Thus the initialization feature is

$$Z_{n,0}(s,a) = D_\alpha F_{n,0}(s,a) = \frac{a}{2} e_{i(s)}.$$

Its conditional mean under the uniform reference policy is zero, and its conditional covariance is $e_i e_i^{\mathsf{T}}/4$ in region i. Both regions have reset probability $1/2$, whence

$$\mathcal{G}_{n,0} := \mathbb{E}_{s\sim\rho}\,\mathrm{Cov}_{a\sim\pi_{\mathrm{ref}}(\cdot|s)}(Z_{n,0}(s,a)) = \frac{1}{8} I_2. \tag{254}$$

□

Lemma C.67 (Reward and upper-loss regularity)**.** *The reward and sample loss satisfy the remaining parts of A1 and all of G on the unrestricted upper domain $\mathbb{R}$, with*

$$B_r = 1, \quad G_r = 1, \quad L_r = 2, \qquad B_L = \log 2 + 4, \quad G_L = 4, \quad L_L = 12. \tag{255}$$

Proof. The bounds $|u| \le 1$, $|u'| \le 1$, and $|u''| \le 2$, together with $|as| \le 1$, give the reward constants. The reward is C^∞ in x, with bounded derivatives through order four uniformly in (s, a).

Since $|D| \le 4$, we have $|\Delta_x| \le 4$, $|\partial_x \Delta_x| \le 4$, and $|\partial_x^2 \Delta_x| \le 8$. For each realized $Y \in \{0, 1\}$,

$$\begin{aligned} 0 \le L &\le \log 2 + |\Delta_x|, \\ \partial_x L &= (\sigma(\Delta_x) - Y)\partial_x \Delta_x, \\ \partial_x^2 L &= \sigma'(\Delta_x)(\partial_x \Delta_x)^2 + (\sigma(\Delta_x) - Y)\partial_x^2 \Delta_x. \end{aligned}$$

Use $|\sigma - Y| \le 1$ and $0 \le \sigma' \le 1/4$ to obtain (255). Composition with the smooth softplus function, with $|D| \le 4$, also supplies the samplewise C^4 bounds required by A1. Equation (249) provides G's parameter-independent conditional feedback law; the sampling construction provides its fixed tuple size, finite trajectory length, and fresh sample access. □

Remark C.68 (A concrete two-hidden-layer variant). Cell constancy is preserved at every fixed hidden depth, but A2 must still be checked. In particular, initializing all hidden weights at zero with $\tanh$ gives zero parameter gradients at depth at least two, and therefore fails A2.

Here is an explicit alternative that trains both hidden layers and has a gap independent of width. Give each hidden layer width $2n$, with n units for each region. Use the same input and fixed all-ones readout as above. Set $W_0^1 = 0$ and let W_0^2 be block diagonal, each n-by-n block having every entry $\sqrt{2/n}$. Then $\|W_0^2/\sqrt{2n}\|_{\mathrm{op}} = 1$. Define a product-space chart by

$$(B_n^1 \alpha)_{(i,k),j} = \frac{\alpha_i}{\sqrt{2n}} \mathbf{1}_{\{j=i\}}, \qquad (B_n^2 \alpha)_{(i,k),(j,\ell)} = \frac{\alpha_i}{\sqrt{2}\, n} \mathbf{1}_{\{j=i\}}.$$

Each chart component has squared Frobenius norm $\|\alpha\|^2/2$, so the combined chart is an isometry. Use $W^{\ell,\pm} = W_0^\ell \pm B_n^\ell \alpha/\sqrt{2}$, our layer normalizations, and its antithetic readout. Writing $t_i = \alpha_i/(2\sqrt{2n})$, direct substitution gives

$$F_{n,\alpha}^{(2)}(s, a) = a H_n(\alpha_{i(s)}), \qquad H_n(\alpha_i) = \frac{\sqrt{n}}{2} \left[\tanh((1 + t_i)\tanh t_i) + \tanh((1 - t_i)\tanh t_i)\right]. \tag{256}$$

Thus $H_n'(0) = 1/(2\sqrt{2})$ and $\mathcal{G}_{n,0}^{(2)} = I_2/16$, for every n. The normalized initialization and the isometric chart supply the width-uniform derivative bounds. By uniform continuity of the first derivative, H_n is strictly increasing on a common neighborhood of zero. Choosing τ sufficiently large puts the target logits $i\tanh(x)/(2\tau)$ in its local image. The local response is then $\alpha_i^*(x) = H_n^{-1}(i\tanh(x)/(2\tau))$. The same five critic features and all Bellman-closure arguments remain valid. The induced response policy, and hence the upper objective and the numerical computations below, are identical to those for the one-hidden-layer parameterization whenever this local response is used.

C.11.4 Five fixed critic features

Define

$$z(s) = \sqrt{12}\left(s - \frac{i(s)}{2}\right), \qquad \varphi_Q(s, a) = \begin{pmatrix} e_{(i(s),a)} \\ a z(s) \end{pmatrix} \in \mathbb{R}^5, \tag{257}$$

where the first four coordinates are ordered as $(-1,-1), (-1,+1), (+1,-1), (+1,+1)$. These features are fixed independently of x, α, n, the critic scale m_Q, and target accuracy. On either region,

$$|z(s)| \le \sqrt{3}, \qquad \mathbb{E}_{\mu_i} z = 0, \qquad \mathbb{E}_{\mu_i} z^2 = 1. \tag{258}$$

Consequently $\|\varphi_Q(s, a)\|_2 \le 2$.

For a fixed actor width, suppress n in the transition-operator notation and set

$$(K_\alpha v)(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \sum_{a' \in \mathcal{A}} \pi_{n,\alpha}(a' \mid s') v(s', a'),$$

$$\rho_U(ds, da) = \rho(ds)\pi_{n,U}(da \mid s), \qquad \nu_U = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k \rho_U K_U^k. \tag{259}$$

Lemma C.69 (C1: uniform feature excitation)**.** *For any fixed ambient actor radius $A > 0$, C1 holds throughout $\|U\|_2 \le A$, uniformly in $n \ge 1$, with*

$$d_Q = 5, \qquad F_Q = 2, \qquad \lambda_Q = \frac{1-\gamma}{2(1+e^A)}. \tag{260}$$

Proof. The inequality $|\tanh t| \le |t|$ gives $|h_n(U_i)| \le |U_i|/2 \le A/2$, so

$$\pi_{n,U}(a \mid i) \ge p_A := \frac{1}{1+e^A}.$$

Within each region the policy is constant in s. Therefore the zero mean in (258) makes the cross moments between the first four features and the fifth feature vanish. The reset Gram matrix is exactly

$$\mathbb{E}_{\rho_U}[\varphi_Q \varphi_Q^\top] = \operatorname{diag}\Big(\big\{\tfrac{1}{2}\pi_{n,U}(a \mid i)\big\}_{i,a}, 1\Big) \succeq \frac{p_A}{2} I_5.$$

Since $\nu_U \ge (1-\gamma)\rho_U$ as measures, integration of the positive semidefinite matrix $\varphi_Q \varphi_Q^\top$ gives (260). This covers every smaller base-actor neighborhood used in Theorem 4.6. □

Lemma C.70 (C2: exact realizability of the regularized value)**.** *For every finite $\tau > 0$, $x \in \mathbb{R}$, $n \ge 1$, and $\alpha \in \mathbb{R}^2$, the regularized lower action value*

$$q_{x,\alpha}(s,a) = -r_x(s,a) + \tau \log \frac{\pi_{n,\alpha}(a \mid s)}{\pi_{\mathrm{ref}}(a \mid s)},$$
$$Q^{\mathrm{true}}_{x,\alpha} = (\mathrm{Id} - \gamma K_\alpha)^{-1} q_{x,\alpha} \tag{261}$$

belongs to $\operatorname{span} \varphi_Q$. *On $\|\alpha\|_2 \le A$,*

$$\|q_{x,\alpha}\|_\infty \le 1 + \tau A, \qquad \|Q^{\mathrm{true}}_{x,\alpha}\|_\infty \le \frac{1+\tau A}{1-\gamma}, \tag{262}$$

uniformly in x and n.

Proof. First,

$$as = \frac{a\, i(s)}{2} + \frac{a z(s)}{\sqrt{12}}. \tag{263}$$

The first term is constant on each region–action pair and the second is a multiple of the fifth feature. Thus $r_x \in \operatorname{span} \varphi_Q$. Also

$$\log \frac{\pi_{n,\alpha}(a \mid s)}{\pi_{\mathrm{ref}}(a \mid s)} = a h_n(\alpha_{i(s)}) - \log\cosh(h_n(\alpha_{i(s)})), \tag{264}$$

which is constant on each region–action pair. In particular $q_{x,\alpha} = \varphi_Q^\top c$ with coefficients

$$c_{(i,a)} = -\frac{u(x) a i}{2} + \tau\big[a h_n(\alpha_i) - \log\cosh(h_n(\alpha_i))\big], \qquad c_5 = -\frac{u(x)}{\sqrt{12}}.$$

For any bounded measurable v, define

$$m_i(v,\alpha) = \int_{I_i} \sum_{a' \in \mathcal{A}} \pi_{n,\alpha}(a' \mid s') v(s',a')\, \mu_i(ds').$$

The transition kernel gives the exact identity

$$(K_\alpha v)(s,a) = \frac{m_+(v,\alpha) + m_-(v,\alpha)}{2} + \frac{as}{4}\big[m_+(v,\alpha) - m_-(v,\alpha)\big]. \tag{265}$$

Both 1 and as belong to $\operatorname{span} \varphi_Q$, so this span is invariant under K_α. The Markov operator is a contraction in the supremum norm. Hence $\sum_{k \ge 0} \gamma^k K_\alpha^k q_{x,\alpha}$ converges uniformly. Every partial sum lies in the finite-dimensional, and therefore closed, span of φ_Q. Its limit is the value in (261), which proves C2.

Finally, on the ambient ball, $|h_n(\alpha_i)| \le A/2$ and $0 \le \log\cosh(h_n(\alpha_i)) \le |h_n(\alpha_i)|$. Thus the absolute log-policy ratio is at most A. The reward bound and the discounted resolvent bound give (262). □

The Bellman-closure step is the same structural principle used for value representation in linear MDPs Jin et al. (2020). Here the additional membership of the log-policy ratio in the fixed feature span is explicitly verified in (264). The unregularized transition has rank-two features $(w_+, w_-) = (1/2+as/4, 1/2-as/4)$ and measure vector (μ_+, μ_-). The five-feature augmentation also represents the actor's KL cost. The hypotheses of Proposition C.62 specialize to

$$\begin{aligned} &M = M' = 2, \quad P = w_+\mu_+ + w_-\mu_-, \\ &\psi(s,a) = a e_{i(s)}, \quad \pi_{\mathrm{ref}}(a \mid i) = 1/2, \\ &\|\varphi_Q\| \le 2, \quad 1,\ as,\ \{\mathbf{1}_{\{(i(s),a)=(j,b)\}}\}_{j,b} \in \operatorname{span}\varphi_Q, \\ &w_\pm,\ r_x \in \operatorname{span}\varphi_Q, \quad \mathbb{E}_{\rho\pi_{\mathrm{ref}}}\varphi_Q\varphi_Q^{\mathsf{T}} = \operatorname{diag}(1/4, 1/4, 1/4, 1/4, 1). \end{aligned}$$

Corollary C.71 (A fixed coefficient envelope)**.** *The coefficient vector $w_{x,\alpha}$ in $Q^{\mathrm{true}}_{x,\alpha} = \varphi_Q^{\mathsf{T}} w_{x,\alpha}$ is unique and, on the ambient ball, satisfies*

$$\|w_{x,\alpha}\|_2 \le \frac{2(1+\tau A)}{1-\gamma}.$$

Proof. At the reference policy, the reset Gram matrix is $\operatorname{diag}(1/4, 1/4, 1/4, 1/4, 1) \succeq I_5/4$. This proves linear independence of the features. Applying the same Gram inequality to $Q^{\mathrm{true}} = \varphi_Q^{\mathsf{T}} w$ yields $\|w\|_2^2/4 \le \mathbb{E}_{\rho_0}(Q^{\mathrm{true}})^2 \le \|Q^{\mathrm{true}}\|_\infty^2$. □

C.11.5 An explicit nonconstant lower response

Use the normalized lower objective (2),

$$g_{\tau,n}(x,\alpha) = (1-\gamma)\mathbb{E}_{n,\alpha}\sum_{t=0}^{\infty}\gamma^t q_{x,\alpha}(s_t,a_t). \tag{266}$$

Proposition C.72 (Selected response and true critic in closed form)**.** *If $\tau > 1/2$, then, for every $x \in \mathbb{R}$ and $n \ge 1$,* (266) *has the unique global minimizer*

$$\alpha_i^*(x) = 2\sqrt{n}\ \operatorname{arctanh}\left(\frac{i\,u(x)}{2\tau\sqrt{n}}\right), \qquad i \in \{-1,+1\}. \tag{267}$$

Its value and true regularized action value are

$$g_{\tau,n}(x,\alpha^*(x)) = -\tau \log\cosh\left(\frac{u(x)}{2\tau}\right), \tag{268}$$

$$Q^{\mathrm{true}}_{x,\alpha^*(x)}(s,a) = -\frac{u(x)}{\sqrt{12}}\,az(s) - \frac{\tau}{1-\gamma}\log\cosh\left(\frac{u(x)}{2\tau}\right). \tag{269}$$

Moreover,

$$\sup_{n\ge1,x\in\mathbb{R}}\|\alpha^*(x)\|_2 \le 2\sqrt{2}\ \operatorname{arctanh}\left(\frac{1}{2\tau}\right), \qquad \partial_x\alpha_i^*(x)|_{x=0} = \frac{i}{\tau}. \tag{270}$$

Proof. Every state law generated from ρ is a mixture of μ_- and μ_+, because this is true initially and every transition in (247) has that form. The normalized discounted state occupancy therefore has the form $d_\alpha = d_-(\alpha)\mu_- + d_+(\alpha)\mu_+$, where $d_- + d_+ = 1$ and each $d_i \ge (1-\gamma)/2 > 0$.

Write $h_i = h_n(\alpha_i)$ and $p_b(a) = e^{ab}/(2\cosh b)$. Under p_{h_i}, $\mathbb{E}a = \tanh h_i$. Since $\mathbb{E}_{\mu_i} s = i/2$, the average one-step cost in region i is

$$\ell_i(x,h_i) = -\frac{i\,u(x)}{2}\tanh h_i + \tau\big[h_i\tanh h_i - \log\cosh h_i\big].$$

Let $b_i = i\,u(x)/(2\tau)$. Direct calculation of binary relative entropy gives

$$\ell_i(x,h_i) = \tau\,\mathrm{KL}(p_{h_i}\|p_{b_i}) - \tau\log\cosh\left(\frac{u(x)}{2\tau}\right).$$

Consequently

$$g_{\tau,n}(x,\alpha) = \tau \sum_{i\in\{-1,+1\}} d_i(\alpha)\ \mathrm{KL}(p_{h_i}\|p_{b_i}) - \tau \log\cosh\left(\frac{u(x)}{2\tau}\right). \tag{271}$$

The dependence of d_i on the actor is retained in this identity. Both weights are strictly positive. Thus equality with the lower bound occurs exactly when $h_i = b_i$ for both regions. Since h_n is strictly increasing with range $(-\sqrt{n}, \sqrt{n})$ and $|b_i| < 1 \le \sqrt{n}$ when $\tau > 1/2$, these equalities have the unique solution (267). This also proves (268).

At this response, (263) and (264) give

$$q_{x,\alpha^*}(s,a) = -\frac{u(x)}{\sqrt{12}} az(s) - \tau \log\cosh\left(\frac{u(x)}{2\tau}\right).$$

The function $az(s)$ has zero mean under each $\mu_i(ds)\pi_{n,\alpha^*}(da \mid i)$. Hence $K_{\alpha^*}(az) = 0$ and $K_{\alpha^*}1 = 1$. The discounted resolvent therefore gives (269) directly.

For $0 \le t < 1$, the power series for inverse hyperbolic tangent gives

$$\sqrt{n}\ \mathrm{arctanh}(t/\sqrt{n}) = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)n^k} \le \mathrm{arctanh}\, t.$$

Apply this with $t = |u(x)|/(2\tau)$ to obtain the norm bound. Differentiating (267) gives

$$\partial_x \alpha_i^*(x) = \frac{i\ \mathrm{sech}^2 x}{\tau\big(1 - u(x)^2/(4\tau^2 n)\big)},$$

which proves the final claim. □

For any fixed target radius $R_0 > 0$, the bound (270) is smaller than R_0 whenever

$$\tau > \frac{1}{2\tanh(R_0/(2\sqrt{2}))}.$$

Thus τ can be chosen large enough to meet our geometry and initialization requirements while placing the entire response family inside any prescribed smaller actor buffer. This same τ is then fixed before selecting critic constants. For $x \neq 0$ the response gives different policies in the two regions, and (269) varies with the continuous state within each region.

C.11.6 An interior upper stationary point and a nonzero chain term

Let $f_n(x,\alpha) = \mathbb{E}L(x;\boldsymbol{\xi},Y)$ and $\Phi_\tau(x) = f_n(x,\alpha^*(x))$. The conditional mean sample loss is

$$\ell_\kappa(u,d) = \log(1+e^{ud}) - ud\,\sigma(\kappa d) = \log(2\cosh(ud/2)) - \frac{ud}{2}\tanh(\kappa d/2). \tag{272}$$

It is even in d. For arbitrary actor coordinates, the exact score is

$$\partial_{\alpha_j} \log \pi_{n,\alpha}(a \mid s) = \mathbf{1}_{\{i(s)=j\}} h_n'(\alpha_j)[a - \tanh(h_n(\alpha_j))].$$

Since the dynamics and conditional feedback law are fixed,

$$f_\alpha = \mathbb{E}\left[L \sum_{\ell=1}^{2}\sum_{t=0}^{1} \nabla_\alpha \log \pi_{n,\alpha}(a_t^\ell \mid s_t^\ell)\right]. \tag{273}$$

Lemma C.73 (An exact one-dimensional quadrature formula)**.** *At the lower response put $u = \tanh x$ and $p = \sigma(u/\tau)$. Define*

$$B_4(y) = \frac{1}{6}\sum_{j=0}^{4}(-1)^j \binom{4}{j}(y-j)_+^3, \qquad 0 \le y \le 4,$$

$$J_k(u) = \int_0^4 B_4(y)\,\ell_\kappa(u, y-4+k)\,dy, \qquad k = 0,\dots,4. \tag{274}$$

Here B_4 is the density of the sum of four independent uniform variables on $[0,1]$. Write $v = p(1-p)$ and

$$(\omega_0, \omega_1, \omega_2, \omega_3, \omega_4) = (v^2,\ 2v - 4v^2,\ 1 - 4v + 6v^2,\ 2v - 4v^2,\ v^2). \tag{275}$$

Then the exact population objective and gradient decomposition are

$$\begin{aligned}
\Phi_\tau(x) &= \sum_{k=0}^{4} \omega_k(p) J_k(u), \\
\mathsf{A}_\tau(x) := f_x(x, \alpha^*(x)) &= (1-u^2) \sum_{k=0}^{4} \omega_k(p) J_k'(u), \\
\mathsf{C}_\tau(x) := D_x \alpha^*(x)^\mathsf{T} f_\alpha(x, \alpha^*(x)) &= \frac{(1-u^2)p(1-p)}{\tau} \sum_{k=0}^{4} \omega_k'(p) J_k(u), \\
\Phi_\tau'(x) &= \mathsf{A}_\tau(x) + \mathsf{C}_\tau(x).
\end{aligned} \tag{276}$$

In the derivative of J_k, the integrand is $[\sigma(ud) - \sigma(\kappa d)]d$, with $d = y - 4 + k$.

Proof. At the response the policy logits are $h_n(\alpha_i^*) = iu/(2\tau)$. Write each state as $s_t = i_t R_t$, where $R_t = |s_t|$, and define $\epsilon_t = a_t i_t$. At reset, R_0 is uniform on $[0,1]$ and independent of i_0. Every subsequent transition first selects a region and then draws a fresh uniform magnitude independent of the past. Conditional on every region and magnitude, the response policy has $\mathbb{P}(\epsilon_t = +1) = p$, independent of that region, magnitude, and the past. Thus the products $a_t s_t = \epsilon_t R_t$ are independent across time, with independent uniform R_t and independent signs of probability p.

This independence concerns the products $a_t s_t$, not the raw states conditioned on an entire future region path. It accounts for the dependence of the transition weights on the preceding state and action. For two independent length-two trajectories,

$$D = \epsilon_1 R_1 + \epsilon_2 R_2 - \epsilon_3 R_3 - \epsilon_4 R_4.$$

Conditional on the four signs in this sum, if k signs are positive then D has the law of a sum of four independent $[0,1]$ uniforms minus $4-k$. Replacing a negatively signed uniform $-R$ by $(1-R)-1$ proves this claim. The probability generating polynomial of k is $[(1-p) + pz]^2[p + (1-p)z]^2$; its coefficients are exactly (275). This proves the objective formula. Differentiating $u(x)$ and $p(x)$ gives $u' = 1 - u^2$ and $p' = (1-u^2)p(1-p)/\tau$. Differentiation of the integrals is justified on their compact domains. The derivative through u is the direct partial derivative, and the derivative through p is exactly the trajectory-law contribution in (273). □

Proposition C.74 (A finite interior minimizer with a nonzero correction)**.** *For every $\kappa \in (0,1)$ and $\tau > 1/2$, let $x^\dagger = \operatorname{arctanh} \kappa$. Then*

$$\Phi_\tau'(0) < 0, \qquad \Phi_\tau'(x^\dagger) > 0.$$

Consequently Φ_τ has an interior minimizer on $[0, x^\dagger]$. At every stationary point $\bar{x} \in (0, x^\dagger)$,

$$\mathsf{A}_\tau(\bar{x}) < 0, \qquad \mathsf{C}_\tau(\bar{x}) = -\mathsf{A}_\tau(\bar{x}) > 0, \qquad \frac{|\mathsf{C}_\tau(\bar{x})|}{|\mathsf{A}_\tau(\bar{x})|} = 1. \tag{277}$$

The chain term vanishes at $x = 0$.

Proof. At $x = 0$, the loss equals $\log 2$ for every tuple and every label, so $f_\alpha(0, \alpha) = 0$ and $\mathsf{C}_\tau(0) = 0$. The direct derivative equals $\mathbb{E}[(1/2 - \sigma(\kappa D))D]$, which is strictly negative: the integrand is negative whenever $D \neq 0$, an event of probability one here.

At $x = x^\dagger$, $u = \kappa$ and the direct derivative is zero. The conditional mean loss is the binary entropy $H(\sigma(\kappa d))$, an even function strictly decreasing in $|d| > 0$. To compare the integrals in (274), let X be the sum of four independent uniforms minus two. Its density b is symmetric and decreasing in $|X|$. For example, $B_4(y) = y^3/6$ on $[0,1]$ and $B_4(y) = [y^3 - 4(y-1)^3]/6$ on $[1,2]$ is increasing there, and symmetry supplies the other half of the density. For any $c > 0$, differentiating the convolution with the entropy function H_κ gives

$$\frac{d}{dc} \mathbb{E} H_\kappa(X + c) = \int_0^\infty H_\kappa'(z)[b(z-c) - b(z+c)]\, dz < 0.$$

x	$\mathsf{A}_2(x)$	$\mathsf{C}_2(x)$	$\Phi_2'(x)$	$\lvert\mathsf{C}_2\rvert/\lvert\mathsf{A}_2\rvert$
0	−0.15561048	0	−0.15561048	0
0.25	−0.07068856	0.00057120	−0.07011736	0.008081
0.50	−0.00814360	0.00116961	−0.00697399	0.143623
$\bar{x}$	−0.00120485	0.00120485	≈ 0	1
$x^\dagger$	0	0.00120937	0.00120937	—
0.75	0.02048602	0.00113742	0.02162344	0.055522
1.00	0.02633981	0.00079848	0.02713829	0.030315

Table 2: Population gradient decomposition for $\kappa = 1/2, \tau = 2$. The ratio at $x^\dagger$ is undefined because its denominator is zero.

Indeed $H'_\kappa(z) < 0$ for $z > 0$ and the bracket is nonnegative, and strictly positive on a set of positive measure. It follows that $J_2(\kappa) > J_1(\kappa) > J_0(\kappa)$ and $J_{4-k}(\kappa) = J_k(\kappa)$.

Holding $u = \kappa$ fixed, write the objective as a polynomial in $v = p(1-p)$. Its derivative with respect to v is

$$4v[J_0 - J_2] + (4 - 16v)[J_1 - J_2] < 0.$$

Here $0 < v < 1/4$, $p > 1/2$, and $dv/dp = 1 - 2p < 0$. Since $p'(x^\dagger) > 0$, the response contribution is strictly positive. This proves the second endpoint sign.

A minimum on the compact interval exists, and the strict endpoint derivative signs exclude either endpoint. At every interior stationary point, $0 < u < \kappa$, so $[\sigma(uD) - \sigma(\kappa D)]D < 0$ almost surely. Thus the direct term is strictly negative and stationarity forces the positive, nonzero cancellation in (277). □

The existence proof holds for every $\tau > 1/2$, including any larger fixed τ selected by our geometry and containment constants. The threshold $1/2$ is sufficient for the explicit response formula; it does not replace that operative algorithmic choice of τ.

Remark C.75 (Preference-risk stationarity and the teacher objective). Let $p^\dagger = \sigma(\kappa D)$, $p_x = \sigma(\tanh(x)D)$, and let H_{B} denote binary entropy. Conditional cross-entropy gives

$$\Phi_\tau(x) = \mathbb{E}_{\boldsymbol{\xi} \sim (\mathbb{P}^2_{\alpha^*(x)})^{\otimes 2}} \left[H_{\mathrm{B}}(p^\dagger) + \mathrm{KL}\big(\mathrm{Ber}(p^\dagger) \| \, \mathrm{Ber}(p_x)\big) \right]. \tag{278}$$

At $x^\dagger = \operatorname{arctanh} \kappa$, the divergence and its derivative vanish, but the expected entropy can vary through the response-dependent trajectory law. Proposition C.74 gives $\Phi'_\tau(x^\dagger) > 0$ and an interior minimizer $\bar{x} \in (0, x^\dagger)$.

The normalized return under the teacher reward $r^\dagger(s,a) = \kappa a s$ is, at the selected response,

$$J^\dagger(x) := (1-\gamma)\mathbb{E}_{\alpha^*(x)} \sum_{t \geq 0} \gamma^t r^\dagger(s_t, a_t) = \frac{\kappa}{2} \tanh\left(\frac{\tanh x}{2\tau}\right). \tag{279}$$

Indeed, conditional on either region, the state is uniform and $\mathbb{E}[a \mid i] = i \tanh(\tanh(x)/(2\tau))$, so every time step has the displayed mean reward. This return is strictly increasing in x; thus $J^\dagger(\bar{x}) < J^\dagger(x^\dagger)$. The example establishes a nonzero response correction and also separates optimization of the stated on-policy risk from reward identification or teacher-return maximization.

A fully specified population illustration. Take $\gamma = 1/2$, $\kappa = 1/2$, $n = 1$, and $\tau = 2$. Deterministic quadrature of (274) gives an interior minimizer

$$\bar{x} \approx 0.5414613800, \qquad x^\dagger \approx 0.5493061443, \qquad \Phi_2(\bar{x}) \approx 0.6559732584. \tag{280}$$

At $\bar{x}$, the direct and response terms are approximately -0.001204848558 and $+0.001204848558$, respectively. Table 2 reports both absolute magnitudes and their ratio. The ratio equals one at a stationary point by cancellation; this identity alone should not be interpreted as a large correction throughout the domain.

The density B_4 is polynomial between consecutive integers. Gauss–Legendre quadrature with 12, 24, and 48 nodes on each of its four pieces agreed at the checked points to less than 10^{-13} in the

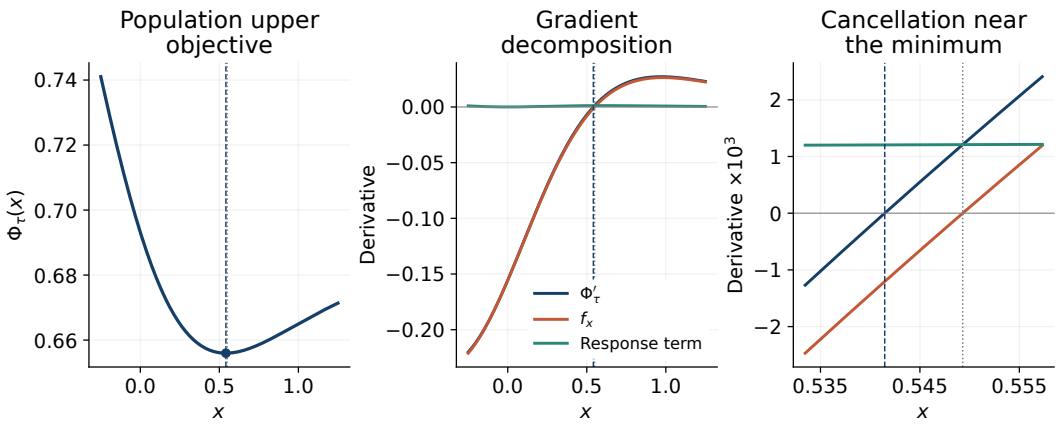

Figure 2: Population quantities evaluated by deterministic quadrature. Left: upper objective. Middle: full, direct, and response derivatives. Right: the same derivatives near the minimum, with the vertical scale multiplied by 10^3. The dashed line is $\bar{x}$ and the dotted line is $x^\dagger$. These curves illustrate the response mechanism; they are not stochastic convergence results.

τ	Interior minimizer $\bar{x}$	$\mathsf{C}_\tau(\bar{x})$
2	0.5414613800	$1.204848558 \times 10^{-3}$
4	0.5473047717	$3.064773816 \times 10^{-4}$
10	0.5489840685	$4.927812863 \times 10^{-5}$
100	0.5493029200	$4.932401058 \times 10^{-7}$

Table 3: The finite stationary point and its nonzero correction persist as τ increases, with $\kappa = 1/2$.

objective and both gradient terms. An independent four-dimensional tensor quadrature with explicit enumeration of the 16 sign patterns agreed at $\bar{x}$ to the same tolerance. These are numerical consistency checks, rather than formal interval-error certificates; the existence and nonzero-chain claims are established analytically above.

For larger fixed regularization the correction remains nonzero, but becomes smaller, as Table 3 illustrates. In this symmetric example $\mathsf{C}_\tau = O(\tau^{-2})$ on compact x intervals: $p - 1/2 = O(\tau^{-1})$ and $\omega'_k(p) = O(p - 1/2)$ in (276). The $\tau = 2$ plot is a population illustration, not a certification of every conservative constant in the stochastic containment theorem.

C.11.7 Joint conclusion and exact critic targets

Proposition C.76 (Joint continuous-state instance)**.** *Fix an ambient actor radius $A > 0$ and $\gamma \in (0, 1)$. Use (247)–(250), the actor (251)–(252), and the feature map (257). Choose sufficiently large finite τ as required by our actor geometry and containment construction, and then keep it fixed. The resulting instance satisfies A1, A2, G, C1, and C2 simultaneously, uniformly over actor widths $2n$ and over the prescribed actor neighborhoods contained in $\|\alpha\|_2 \le A$. The reward and upper domains are unrestricted in $x \in \mathbb{R}$. The actor has retained dimension 2, the critic has fixed feature dimension 5, and the constants in (254), (255), and (260) do not depend on critic scale or target accuracy.*

Proof. Combine Lemmas C.66, C.67, C.69, and C.70. The critic verifications hold on the full ambient ball, so they also cover base and perturbed actors and the local update segments used by the theorem. Proposition C.72 shows that the selected response can be placed strictly inside the chosen geometry neighborhood. No assumption about an independently accurate critic initialization is used in these verifications. □

Remark C.77 (Scope and the nonlinear trained critic). The example has a finite-rank transition kernel and a policy that uses a two-region state encoding. State differentiability is not required by A1; the measurable encoding $i(s)$ is therefore allowed. The transition probabilities and rewards retain

within-region continuous state dependence. This establishes joint satisfiability within the stated model class.

C2 is a statement about the fixed linear span $\operatorname{span}\varphi_Q$. We train

$$Q_{m_Q,\theta}(s,a) = \sqrt{m_Q}\tanh\left(\frac{\varphi_Q(s,a)^\mathsf{T}\theta}{\sqrt{m_Q}}\right),$$

whose finite nonlinear scale need not yield exact representation of the true value. For example, on $\|\theta\|_2 \le R_Q$ the feature bound above gives the elementary estimate

$$\left\|Q_{m_Q,\theta} - \varphi_Q^\mathsf{T}\theta\right\|_\infty \le \frac{8R_Q^3}{3m_Q}.$$

Our paired-TD stability and tracking arguments are still used to obtain its derived squared paired approximation floor and interaction bound. Exact feature-span membership is the primitive realizability condition verified by this example.

Exact value coefficients for arbitrary actors. The one-dimensional response formula is not needed to compute the true value away from the response. Let the cell order be that of (257). For $j \in \{-1,+1\}$ define coefficient vectors $v_j, m_j(\alpha) \in \mathbb{R}^5$ by

$$(v_j)_{(i,a)} = \frac{1}{2} + \frac{jai}{8}, \qquad (v_j)_5 = \frac{j}{4\sqrt{12}}, \qquad (m_j)_{(i,a)} = \mathbf{1}_{\{i=j\}}\pi_{n,\alpha}(a \mid i), \qquad (m_j)_5 = 0.$$

Then $w_j(s,a) = \varphi_Q(s,a)^\mathsf{T} v_j$ and $\mathbb{E}_{\mu_j\pi_\alpha}\varphi_Q = m_j(\alpha)$. Define

$$T_\alpha = \sum_{j\in\{-1,+1\}} v_j m_j(\alpha)^\mathsf{T}, \qquad K_\alpha(\varphi_Q^\mathsf{T} b) = \varphi_Q^\mathsf{T} T_\alpha b. \tag{281}$$

If $c_{x,\alpha}$ is the coefficient vector for q in the proof of Lemma C.70, the true-value coefficients are exactly

$$w_{x,\alpha} = (I_5 - \gamma T_\alpha)^{-1} c_{x,\alpha}. \tag{282}$$

The inverse exists because the discounted Bellman operator is invertible on the invariant feature span and the features are linearly independent.

The finite nonlinear TD target is a different root. For a base actor U, the state occupancy weights can themselves be computed from a two-by-two matrix. Its rows are

$$\overline{P}_U(i,+) = \frac{1}{2} + \frac{i\tanh(h_n(U_i))}{8}, \qquad \overline{P}_U(i,-) = 1 - \overline{P}_U(i,+).$$

With row-vector convention,

$$d_U = (1-\gamma)(1/2,1/2)(I_2 - \gamma\overline{P}_U)^{-1},$$
$$\nu_U(ds,da) = \sum_i d_U(i)\mu_i(ds)\pi_{n,U}(a \mid i).$$

For finite critic scale, the relevant mean field remains

$$\mathsf{H}_{m_Q}(x,U,\alpha,\theta) = \mathbb{E}_{\nu_U}\left[(Q_{m_Q,\theta} - q_{x,\alpha} - \gamma K_\alpha Q_{m_Q,\theta})\nabla_\theta Q_{m_Q,\theta}\right]. \tag{283}$$

All integrals in this field reduce to finite action sums and one-dimensional within-region integrals. Our interior root $\theta^*_{m_Q}(x,U,\alpha)$ can therefore be computed by a five-dimensional nonlinear root solve with controlled quadrature. For $p_{\mathrm{alg}} = (x,U,S)$ the paired target is

$$Z_{m_Q}(p_{\mathrm{alg}}) = \left(\theta^*_{m_Q}(x,U,U),\ \frac{\theta^*_{m_Q}(x,U,U+cS) - \theta^*_{m_Q}(x,U,U)}{c}\right). \tag{284}$$

The linear solve (282) computes the exact value coefficients, not this nonlinear TD root. Monitoring the theorem's $\|z_t - Z_{m_Q}(p_t)\|_W$ therefore requires the nonlinear solves, with tolerances small enough after division by c.

Population evaluation away from the lower response. The local perturbed response solves $g_\alpha(x, \alpha_c^*) + c f_\alpha(x, \alpha_c^*) = 0$, with $S_c^* = (\alpha_c^* - \alpha_0^*)/c$. Away from the unperturbed response, preference expectations must retain the transition likelihoods; the independent-product reduction in Lemma C.73 is specific to the response policy. The exact population calculations provide evaluation quantities and do not replace the sampling estimators in Algorithm 1.

C.12 Conditional global lower responses on a prescribed domain

This subsection proves Corollary A.2 and specifies the scope of its additional assumption. Fix the finite actor architecture $\mathbf{m}$ and the same fixed regularization parameter τ as in Theorem 4.6. Write $a(x) = \alpha_{0,\mathbf{m}}^*(x)$ and

$$g_{\tau,\mathbf{m},\mathcal{D}}^*(x) := \min_{\beta \in \mathcal{D}} g_{\tau,\mathbf{m}}(x, \beta), \qquad \mathcal{A}_{\tau,\mathbf{m},\mathcal{D}}^*(x) := \arg\min_{\beta \in \mathcal{D}} g_{\tau,\mathbf{m}}(x, \beta).$$

The minimum exists by continuity and compactness. The actor map remains defined on $\mathbb{R}^r$; $\mathcal{D}$ specifies the comparison domain for global lower optimality.

Proof of Corollary A.2. Proposition B.1, with $c = 0$, gives $a(x) \in B_{R_{\mathrm{geom}}/4} \subset \mathcal{D}$ and $\nabla_\alpha g_{\tau,\mathbf{m}}(x, a(x)) = 0$ for every x. Assumption A.1 therefore yields

$$0 \le g_{\tau,\mathbf{m}}(x, a(x)) - g_{\tau,\mathbf{m},\mathcal{D}}^*(x) \le \frac{\|\nabla_\alpha g_{\tau,\mathbf{m}}(x, a(x))\|^2}{2\mu_{\mathrm{PL}}} = 0. \tag{285}$$

Thus $a(x) \in \mathcal{A}_{\tau,\mathbf{m},\mathcal{D}}^*(x)$. The selected global-response objective $\Phi_{\tau,\mathbf{m},\mathcal{D}}^{\mathrm{sel}}(x) = f_{\mathbf{m}}(x, a(x))$ equals $\Phi_{\tau,\mathbf{m}}(x)$ everywhere. Smoothness follows from the existing response construction, and their gradients agree. Theorem 4.6, applied to the same algorithm and output, therefore gives expected squared gradient norm $O(\epsilon)$ with $\widetilde{O}(\epsilon^{-2})$ segments and transitions, and its unchanged label and numerical KL-solve costs. □

Domain, constants, and algorithmic scope. The prescribed $\mathcal{D}$ is independent of x, perturbation c, critic scale m_Q, and accuracy ϵ. A meaningful extension compares against actors beyond the neighborhood already covered by derived strong convexity. The additional PL condition uses the true fixed-τ lower objective and its ambient Euclidean gradient, including at boundary points of $\mathcal{D}$. It is not a projected-gradient condition and need not certify arbitrary constrained stationary points. The proof evaluates it at the interior, zero-gradient response. Compactness and inclusion of that response suffice for the identification argument; convexity is a convenient domain convention.

No projection onto $\mathcal{D}$ or new global lower solver is introduced. All initializations, local containment and failure-output rules, sampling blocks, parameter choices, and finite KL solves remain those of Theorem 4.6. In particular, the argument imposes no PL condition on the perturbed objective $h_c = g_{\tau,\mathbf{m}} + c f_{\mathbf{m}}$ and requires no critic realizability or training beyond the neighborhoods in C1–C2. Those conditions continue to justify the estimator locally. The extra assumption changes the interpretation of the selected response, not the algorithm or its distribution of outputs.

For a fixed architecture, μ_{PL} may depend on $\mathbf{m}$, $\mathcal{D}$, and τ. It enters no step-size or sample-size choice in this corollary, so no new inverse-PL factor is introduced into the inherited interaction bound. Positivity separately at each x would already suffice in (285); uniformity in x is the stated stronger condition. If widths are instead selected as $\mathbf{m} = \mathbf{m}(\epsilon)$, impose the condition for every architecture used on an accuracy-independent domain and retain the target notation $\Phi_{\tau,\mathbf{m}(\epsilon),\mathcal{D}}^{\mathrm{sel}}$. The argument does not identify targets at different widths, and does not need a width-uniform PL constant. The pointwise certificate

$$g_{\tau,\mathbf{m}}(x, \alpha) - g_{\tau,\mathbf{m},\mathcal{D}}^*(x) \le (2\mu_{\mathrm{PL}})^{-1} \|\nabla_\alpha g_{\tau,\mathbf{m}}(x, \alpha)\|^2, \qquad \alpha \in \mathcal{D},$$

would require a separate residual estimate to give a finite-sample lower-suboptimality bound for an approximate actor output.

Selection independence and optimization target. PL alone does not generally guarantee uniqueness of a lower minimizer. The corollary retains the specified smooth selection $a(x)$. If, in addition,

$\mathcal{A}^*_{\tau,\mathbf{m},\mathcal{D}}(x)$ is a singleton for every x, that selection is the unique global lower response. A weaker sufficient condition for a selection-independent upper objective is upper-loss invariance:

$$f_{\mathbf{m}}(x,\alpha) = f_{\mathbf{m}}(x,\beta) \quad \text{for all } \alpha,\beta \in \mathcal{A}^*_{\tau,\mathbf{m},\mathcal{D}}(x), \quad \text{for every } x. \tag{286}$$

Then every choice of global lower solution induces the same scalar objective, which is smooth because it equals $\Phi_{\tau,\mathbf{m}}$, even if some selections themselves are nonsmooth. Without either condition or another selection argument, the corollary does not identify its objective with the optimistic or pessimistic upper objective obtained by minimizing or maximizing $f_{\mathbf{m}}$ over the lower solution set.

The conclusion is upper stationarity for a selected global *regularized* lower response within the declared actor class and domain. Matching another bilevel target additionally requires matching the policy class, domain, regularization, and selection convention. It gives no unregularized-response comparison or guarantee from arbitrary actor initialization. In particular, τ stays at the fixed value required for the original geometry; the analysis does not permit $\tau \downarrow 0$. PL only for the unregularized lower objective would not justify (285).

Why a full-space PL assumption is unavailable here. The bounded activations and fixed readout in (80) and (92) imply, at every branch parameter, $|G_{\mathbf{m},\vartheta}(s,a)| \le \sqrt{m_L} M_{L,0}$. Consequently the antithetic energy (85) obeys

$$|F_{\mathbf{m},\alpha}(s,a)| \le B_{F,\mathbf{m}} := \sqrt{2m_L} M_{L,0} < \infty \qquad (\alpha \in \mathbb{R}^r). \tag{287}$$

For the exponential tilt (87), (88) gives $|A_{\mathbf{m},s}(\alpha)| \le B_{F,\mathbf{m}}$, and hence $|\log(\pi_{\mathbf{m},\alpha}/\pi_{\mathrm{ref}})| \le 2B_{F,\mathbf{m}}$. The expected log ratio at each state is a nonnegative KL divergence. The reward bound and normalized discounting in (2) therefore yield

$$-B_r \le g_{\tau,\mathbf{m}}(x,\alpha) \le B_r + 2\tau B_{F,\mathbf{m}} \qquad \text{for all } x,\alpha. \tag{288}$$

This bound needs fixed finite widths, but no uniform positive lower bound on reference action probabilities across states. For fixed x, the objective is C^2 on $\mathbb{R}^r$: bounded inputs, bounded activation derivatives, and finite layer weights give uniform actor and score derivative bounds on each compact parameter set. Differentiated length-t trajectory likelihoods have polynomial-in-t bounds, which are summable with γ^t, justifying twice differentiating the discounted expectation locally.

Lemma C.78 (Bounded objectives obstruct full-space uniform PL)**.** *Let $v : \mathbb{R}^r \to \mathbb{R}$ be C^2 and bounded above, with $v_* := \inf_{z\in\mathbb{R}^r} v(z) > -\infty$. If there is $\mu > 0$ such that $\frac{1}{2}\|\nabla v(z)\|^2 \ge \mu[v(z) - v_*]$ for every $z \in \mathbb{R}^r$, then v is constant.*

Proof. If v is nonconstant, choose z_0 with $v(z_0) > v_*$ and consider the maximal solution of $\dot z = \nabla v(z)$ from z_0. Local existence and uniqueness follow from C^2 regularity. With $M = \sup v < \infty$,

$$\int_0^t \|\nabla v(z(u))\|^2\, du = v(z(t)) - v(z_0) \le M - v(z_0), \qquad \|z(t) - z(s)\| \le \sqrt{t-s}\sqrt{M - v(z_0)}$$

for $0 \le s < t$ in the solution interval. If the maximal forward time were finite, the second estimate would give a finite limit for $z(t)$ there; local existence at that limit would extend the solution. Thus it exists for all $t \ge 0$. But $e(t) = v(z(t)) - v_* > 0$ satisfies $e'(t) = \|\nabla v(z(t))\|^2 \ge 2\mu e(t)$, so $e(t) \ge e(0)e^{2\mu t}$. This contradicts $e(t) \le M - v_*$. □

Apply Lemma C.78 to $v(\alpha) = g_{\tau,\mathbf{m}}(x,\alpha)$ at any fixed x. Equation (288) supplies both bounds, while Proposition B.1's positive local Hessian makes v nonconstant for $r \ge 1$. A positive full-space uniform Euclidean PL constant is therefore impossible for this construction. Replacing $\mathcal{D}$ by $\mathbb{R}^r$ in Assumption A.1 would make the conditional extension vacuous in these nontrivial instances.

Compatibility on compact domains. The obstruction does not apply when ascent can leave the comparison domain. For illustration, $v(a) = \tanh^2(a)$ satisfies on any $[-A, A]$, $0 < A < \infty$,

$$\tfrac{1}{2}|v'(a)|^2 = 2\,\mathrm{sech}^4(a)v(a) \ge 2\,\mathrm{sech}^4(A)[v(a) - \min_{b\in[-A,A]} v(b)]. \tag{289}$$

It is smooth, globally bounded, and nonconstant, with positive curvature near its minimum and a positive PL constant on each fixed compact interval. This illustrates compatibility with bounded nonlinearities; it does not verify Assumption A.1 for the RL examples.

A separate condition for upper optimality. Set $V = \Phi^{\mathrm{sel}}_{\tau,\mathbf{m},\mathcal{D}}$. If a separately specified upper domain $\mathcal{X}$ contains x_{out} almost surely and, for $V_* := \inf_{x\in\mathcal{X}} V(x) > -\infty$, one additionally has

$$\tfrac{1}{2}\|\nabla V(x)\|^2 \geq \mu_{\mathrm{up}}[V(x) - V_*] \quad (x \in \mathcal{X}), \qquad \mu_{\mathrm{up}} > 0,$$

then taking expectations gives $\mathbb{E}[V(x_{\mathrm{out}}) - V_*] = O(\epsilon/\mu_{\mathrm{up}})$. Neither the upper PL condition nor output membership in such a domain is asserted here. They do not follow from lower PL. In fact, Corollary B.2 gives a bounded C^2 upper objective, so Lemma C.78 also precludes positive full-space upper PL when that objective is nonconstant.

D Continuous-state experimental details

This section specifies the sampled continuous-state experiment and its population diagnostics. The structural construction and its C1–C2 verification appear in Appendix C.11, especially Lemmas C.69 and C.70. We use the two-hidden-layer actor in Remark C.68. The code was ran on a 12700k processor with an RTX 5090 and 32GB RAM.

D.1 Instance, parameterization, and reference calculations

Let $I_- = [-1, 0)$, $I_+ = [0, 1]$, and $\mu_i = \mathrm{Unif}(I_i)$. The reset distribution is uniform on $[-1, 1]$, and the transition law is

$$P(ds' \mid s, a) = \left(\tfrac{1}{2} + \tfrac{as}{4}\right)\mu_+(ds') + \left(\tfrac{1}{2} - \tfrac{as}{4}\right)\mu_-(ds'), \qquad a \in \{-1, +1\}.$$

We use $\gamma = 1/2$, reward $r_x(s, a) = \tanh(x)as$, and reference-KL coefficient $\tau = 2$ with a uniform reference policy. The lower objective is the normalized discounted cost with $q_{x,\alpha} = -r_x + \tau\log(\pi_\alpha/\pi_{\mathrm{ref}})$. Two independent length-two state–action trajectories define $D = \sum_{t=0}^{1}(a_t^1 s_t^1 - a_t^2 s_t^2)$; the fixed feedback kernel is $Y \mid D \sim \mathrm{Bernoulli}(\sigma(D/2))$. The preference loss is

$$L(x; D, Y) = \log(1 + \exp(\tanh(x)D)) - Y\tanh(x)D.$$

The upper objective $\Phi(x) = f(x, \alpha_0^*(x))$ averages this loss over trajectories generated by the lower response. There is no upper projection or added quadratic upper regularizer.

The tied actor has two hidden layers of width eight and two trainable chart coordinates. Its logits are $aH_4(\alpha_{i(s)})$, where

$$H_4(v) = \tanh((1 + t)\tanh t) + \tanh((1 - t)\tanh t), \qquad t = v/(2\sqrt{8}).$$

Both hidden layers vary with these coordinates. The critic is $Q_\theta(s, a) = \sqrt{m_Q}\tanh(\varphi(s, a)^\top\theta/\sqrt{m_Q})$. The five fixed features comprise four region–action indicators and $a\sqrt{12}(s - i(s)/2)$, with region sign $i(s) \in \{-1, +1\}$. Thus m_Q varies the critic's nonlinearity while its parameter dimension remains five. Training samples continuous states without a state grid; the two-region structure permits separate deterministic reference calculations.

The learned variants use old-coordinate updates, independent random streams for the update blocks, no critic warmup, and no population targets. Exact responses, gradients, and TD roots are evaluated after training or in explicitly identified population studies.

D.1.1 Connection to the theoretical sample complexity

The ordinate in Figure 1(a) matches the stationarity criterion of Theorem 4.6. For a successful run and an independent index J uniform on $\{0, \ldots, T - 1\}$,

$$\mathbb{E}_J\left[|\Phi'(x_J)|^2 \mid x_0, \ldots, x_{T-1}\right] = \frac{1}{T}\sum_{t=0}^{T-1}|\Phi'(x_t)|^2 = \overline{G}_T.$$

Thus averaging the displayed run-level statistic across training seeds estimates the theorem's uniform-output squared-gradient criterion; the prescribed fallback accounts for failed runs as described above. The theoretical relation

$$\mathbb{E}\|\nabla\Phi(x_{\mathrm{out}})\|^2 = O(\varepsilon), \qquad N = \widetilde{\mathcal{O}}(\varepsilon^{-2})$$

has leading polynomial dependence $N^{-1/2}$ when expressed as squared stationarity versus interaction budget, up to logarithmic factors. This is the reference slope shown in the main-text plot.

The experimental schedule also follows the theorem's powers of the iteration budget. Writing $\varepsilon_T \asymp T^{-1/2}$ gives

$$\eta_x, \eta_y, \beta, c^2, m_Q^{-2} \asymp T^{-1/2}, \qquad \frac{\Psi_0}{\eta_x T} = O(T^{-1/2}),$$

for a uniformly bounded initial potential Ψ_0. These are the terms in the stationarity bound (23), with the remaining truncation and certificate terms controlled at the corresponding accuracy in the theorem. With fixed batches and geometric rollout lengths of bounded mean, charged interactions grow proportionally to T in expectation, apart from the initial metric cost. Accordingly, the fitted slopes near -0.505 and nearly constant $\sqrt{N}\,\overline{G}_T$ provide finite-budget support for the predicted interaction scaling. An eightfold increase in budget predicts a reduction by approximately $\sqrt{8} = 2.828$ in this metric; the observed reductions are 2.856 and 2.859.

D.1.2 Role of the two critic initializations

The two curves use the same actor and reward initialization, matched training seeds, interaction budgets, and update schedules. They differ in the initial critic coordinates. The zero-initialization curve starts from $u_0 = v_0 = 0$, which represents the exact values at $x_0 = U_0 = S_0 = 0$. The displaced-initialization curve starts from $\|u_0\| = 2$ and $\|v_0\| = 4$, so its initial critics are inaccurate. Both settings subsequently train the critics through sampled TD updates while the reward and actor parameters evolve.

Theorem 4.6 permits any deterministic critic initialization in $\mathcal{K}_Q$. Initial critic error contributes to Ψ_0 in (23); with a displacement fixed across budgets, its contribution $\Psi_0/(\eta_x T)$ retains the $T^{-1/2}$ dependence. The theorem therefore allows initialization-dependent constants with the same interaction exponent. Comparing the two curves checks whether the predicted decay is also observed when training begins with inaccurate value estimates. Their similar slopes support this aspect of the guarantee, while the critic-error measurements in Section D.2 directly document recovery from the displaced initialization.

D.2 Critic initialization and recovery

We evaluate persistent critic learning from zero and displaced initializations using the full update in both cases. The fixed-budget study uses $T = 2000$, batch size 256 in each sampled block, geometric cap $H = 32$, $c = 0.05$, $m_Q = 256$, TD step $\beta = 0.2$, and actor and upper steps $\eta_y = \eta_x = 0.02$. The initial metric uses 4722 reset states; critic radii are $R_u = 4$ and $R_v = 8$. The two initializations use matched seeds 100–109, with ten runs per initialization. All 20 runs complete.

At $x_0 = U_0 = S_0 = 0$, zero critics represent the exact initial values. The displaced initialization has $\|u_0\| = 2$ and $\|v_0\| = 4$, in fixed directions shared across seeds. Both settings train the critics from the first iteration, without critic warmup or population targets in the updates.

For evaluation, population TD roots at the current reward and actor parameters supply (u^*, v^*). The weighted parameter tracking error is $100\|u - u^*\|^2 + \|v - v^*\|^2$, evaluated at 26 saved reference checkpoints per run. Table 4 reports initial and final seed-averaged errors for the displaced initialization. Value errors are measured against the true regularized values. The base metric is the squared sup-norm error; the paired metric is the squared sup-norm of the endpoint-error difference after division by c.

Critic error metric	Initial mean	Final mean
Weighted parameter tracking	416	0.1904
Squared base value error	5.8796	1.141×10^{-3}
Squared normalized paired value error	23.7719	4.437×10^{-4}

Table 4: Critic recovery from displaced initialization. Entries are means over ten training seeds; population reference calculations are used only for evaluation.

The reductions in both base and normalized paired value errors show recovery from initially inaccurate critics on this continuous-state instance while the reward and actor parameters evolve.

E END-TO-END CONTINUOUS-STATE BENCHMARK DETAILS

This section describes Walker Walk and Highway-fast. They both have continuous observations; Highway-fast has a finite action set. The separate continuous-state theory-aligned study has its own protocol. These practical SAC experiments evaluate end-to-end learning performance.

E.1 COHORT, UNCERTAINTY, AND INTERACTION ACCOUNTING

Both figures use 5 training seeds. Each seed contributes the mean of ten held-out evaluation episodes at each checkpoint. Lines average these five means; bands show their sample standard deviation with denominator $5 - 1$, not a confidence interval or a pooled episode SD. Each preference run uses 100,000 main interactions and 500 labels. Walker Walk additionally uses 9,216 auxiliary response interactions per run. Highway-fast auxiliary interactions are listed below. Known-reward SAC uses 100,000 main interactions and no preference labels or auxiliary responses. The horizontal axis counts total training interactions: main plus cumulative auxiliary response transitions. For each common evaluation checkpoint, the plotted cost is averaged across the five seeds.

Table 5: Highway-fast auxiliary interactions, in addition to 100,000 main interactions per run.

Seed	Secant	Gaur et al.	Known-reward SAC
1	1682	1800	0
3	1712	1607	0
4	1800	1734	0
5	1612	1591	0
6	1704	1578	0

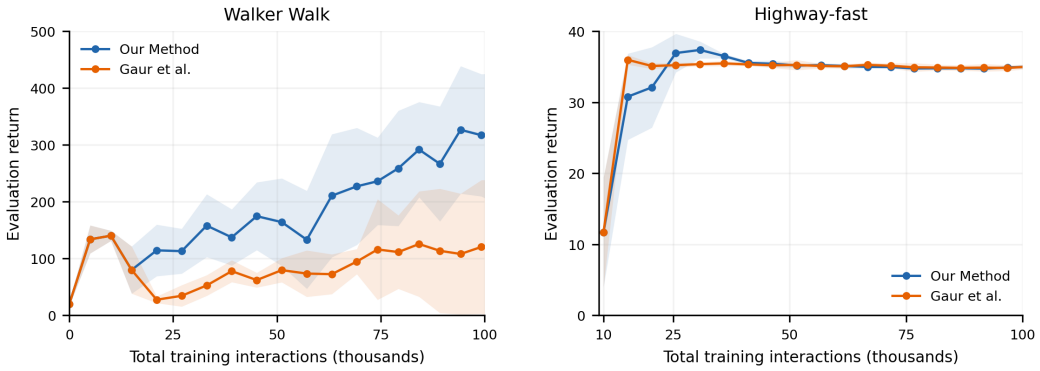

Figure 3: End-to-end deterministic evaluation returns: mean and one sample SD across five matched training seeds, each contributing a ten-episode mean. The horizontal axis includes cumulative auxiliary response interactions; checkpoint costs are averaged across seeds. Walker Walk starts at its first measured checkpoint, and Highway-fast is shown after exploration warm-up.

E.2 END-TO-END IMPLEMENTATION

Environments. Walker Walk uses DeepMind Control, a 24-dimensional observation, a bounded six-dimensional continuous action, and 1000-step episodes with per-step task reward in $[0, 1]$. Highway-fast uses `highway-fast-v0`, with 25 continuous observation features: presence, position (x, y), and velocity (v_x, v_y) for five vehicles. Its five actions are lane left, idle, lane right, faster, and slower. The configuration uses three lanes, 20 traffic vehicles, vehicle density 1, policy frequency 1 Hz, simulation frequency 5 Hz, and at most 50 policy steps per episode. Collisions terminate episodes;

`offroad_terminal` is false. Task rewards are normalized. Returns are undiscounted sums of true environment rewards.

Actor and critics. Walker Walk uses a tanh-squashed diagonal-Gaussian actor with two 1024-unit ReLU layers, outputting action means and log standard deviations; the latter are clamped to $[-5, 2]$. Its two Q networks each map concatenated state/action through two 1024-unit ReLU layers to a scalar. Highway-fast uses a categorical actor with two 1024-unit ReLU layers and five logits. Each of its two Q networks maps the state through two 1024-unit ReLU layers to five action values. Discrete SAC computes the action expectation exactly. Highway actions are represented by five-dimensional one-hot vectors in the reward model.

Reward model and labels. Both preference methods use an ensemble of three reward networks, each with three 256-unit LeakyReLU layers and a scalar tanh output. Inputs concatenate state and action. Reward Adam learning rate is 3×10^{-4}; preference batch size and reward updates per query block are both 50. Segment length is 50 for Walker Walk and 10 for Highway-fast. Synthetic labels use segment task rewards, $\beta = -1$, $\gamma_{\text{teacher}} = 1$, zero mistake and skip probabilities, and equality-threshold coefficient 0.1. This coefficient scales the teacher's equality margin; it is not an independent probability of assigning an equal label. Acquisition uses disagreement sampling from a ten-times-larger candidate batch.

Table 6: Common SAC and collection settings. Feedback settings apply to the two preference methods. Evaluation schedules are specified below.

Setting	Value
Main environment interactions / replay capacity	100,000 / 100,000
Random / intrinsic warm-up interactions	1000 / 9000
Discount / SAC minibatch size	0.99 / 1024
Actor and critic learning rates / Adam betas	5×10^{-4} / $(0.9, 0.999)$
Temperature learning rate / initial temperature	10^{-4} / 0.1 (learned)
Actor / critic-target update frequencies	1 / 2
Target Polyak coefficient	0.005
Feedback interval / total label budget	5000 / 500
Critic reset updates after warm-up	100
Held-out evaluation episodes per checkpoint	10
Training device	CUDA

Response and upper-update settings. Secant uses perturbation $c = 0.1$ and chain coefficient 1. Each response branch receives 100 shadow-SAC updates and supplies two auxiliary rollout episodes, capped at 256 steps on Walker Walk and 50 on Highway-fast. The 10k feedback block performs preference bootstrapping; nine subsequent blocks apply the upper correction, beginning at 15k. Across those blocks there are 1800 shadow-SAC updates (two branches per block).

The Gaur et al. (2025) comparator uses the upper objective $G + \widehat{V}_{\text{base}}/\sigma$, with $\sigma = 0.05$ on Walker Walk and $\sigma = 0.08$ on Highway-fast. This is the upper level loss used for the empirical results in Sectio F of the work. The rest of the pipeline, including both response branches, the perturbation $c = 0.1$, their update counts, and auxiliary rollout collection, is retained. Thus c configures the retained response construction; it is not substituted for σ in the Gaur upper loss. The value contribution uses the base response's normalized, truncated discounted learned-reward return and is averaged across reward-ensemble members. The positive sign is retained, including for negative learned returns. σ was chosen via grid search over $[0.01, 0.2]$.

Evaluation and display. Evaluation uses 10 held-out environment seeds: the Gaussian mean with squashing on Walker Walk and the categorical argmax on Highway-fast. Walker Walk uses saved actors at 25k, 50k, 75k, and 100k. The Walker curve begins at its first measured checkpoint, 25k main interactions; no artificial origin is included. Highway-fast evaluations are recorded at the first episode boundary after each 5k milestone (less than 50 steps after the nominal milestone); the exact step is retained in the data. The figure includes 10k–100k main-interaction checkpoints, displayed at their total training costs and omitting the earlier exploration period. The preference methods' shared

warm-up produces identical early policies within each seed; their upper updates first differ at 15k. Lines join available evaluations without smoothing, jitter, or synthetic intermediate measurements.

Reproduction material. The accompanying code regenerates figures and tables from the archived episode returns and verifies the recorded source hashes. Exact source versions and the support snapshot are supplied for reconstructing each historical training run. The code README distinguishes rebuilding reported figures from rerunning GPU training with the recorded environment.