A CLOSED-LOOP NON-ASYMPTOTIC CONVERGENCE ANALYSIS OF PPO WITH LEARNED CRITICS AND CLIPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite its widespread use, Proximal Policy Optimization with clipping (PPO-Clip) remains difficult to tune, and the interactions among critic learning, clipping, and rollout reuse remain incompletely understood. We develop a *non-asymptotic* analysis of PPO-Clip as a *closed-loop actor–critic* system. It captures actor–critic coupling, nonsmooth probability-ratio clipping, finite-batch reuse, and predictable early stopping under explicit coverage and critic regularity assumptions, using raw GAE and Monte Carlo critic targets. Our synchronous and asynchronous guarantees jointly characterize policy stationarity and the tracking accuracy of the learned critic, with explicit dependence on algorithmic parameters. A sufficient coupling condition gives optimization, critic tracking, clipping, and finite-batch errors a common amplification bound. The asynchronous result also requires a delay-dependent critic stepsize restriction; violating these conditions does not establish divergence. For finite layered MDPs with tabular critics, a uniform bound on the actual clipped-gradient class replaces complete-trajectory counting. A verified growing-horizon family has polynomial sample complexity, and a two-time-scale schedule gives $O(T^{-2/5})$ stationarity and critic-tracking bounds with explicit fresh-rollout accounting. These results together advance our understanding about PPO and provide theoretical guidance in tuning.

1 INTRODUCTION

Proximal Policy Optimization (PPO), particularly its clipped variant PPO-Clip, is widely used in reinforcement learning (RL) (Schulman et al., 2017). PPO-Clip typically uses an actor–critic framework: the *actor* updates the policy using sampled policy-gradient estimates, while the *critic* estimates values to compute advantage signals. Its practical success is commonly attributed to combining policy-gradient updates with stabilization mechanisms (Yu et al., 2022; Jin et al., 2023; Huang et al., 2024; Dwyer et al., 2025; Su et al., 2025). One motivating application is fine-tuning large language models (LLMs), including reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Stiennon et al., 2020; Kirk et al., 2024; Havrilla et al., 2024).

PPO-Clip uses a clipped surrogate objective (Markowitz & Staley, 2023; Huang et al., 2024; Dwyer et al., 2025), which limits incentives for large probability-ratio changes relative to a reference policy. This trust-region–inspired design avoids explicit second-order optimization. Beyond clipping, practical implementations use synchronous or asynchronous *actor–critic updates*, *minibatch reuse* across multiple epochs, *generalized advantage estimation*, and *early stopping* based on KL divergence (Mnih et al., 2016; Dossa et al., 2021; Hilton et al., 2022; Yu et al., 2022; Xiao et al., 2022; Sun et al., 2022; Sheng et al., 2025). These mechanisms improve stability and sample efficiency but couple actor and critic dynamics across multiple time scales. Deploying PPO-Clip on new tasks often requires substantial tuning of learning rates, clipping thresholds, and update schedules to avoid instability or performance collapse (Engstrom et al., 2020; Agarwal et al., 2021; Zheng et al., 2023; Adkins et al., 2024; Moalla et al., 2024). Our limited understanding of these interactions motivates a joint analysis of PPO's update mechanisms under explicit assumptions.

Existing Gap. Existing results include finite-time two-time-scale actor–critic analysis (Wu et al., 2020), stationary-point analysis retaining nonsmooth PPO clipping and multiple updates (Jin et al.,

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.

2023), and PPO-Clip methods with learned value estimation (Huang et al., 2024). *Establishing a joint guarantee for interacting PPO mechanisms remains a substantive analytical challenge.* The critic changes the actor's advantage estimates, the actor moves the critic's target, and both adapt to reused data; separate guarantees cannot simply be combined without controlling these dependencies. We address this challenge through a *non-asymptotic* analysis of interleaved raw-GAE updates, Monte Carlo critic targets, and finite-batch reuse. Under explicit coverage and critic regularity, the analysis jointly controls actor stationarity and critic tracking, including empirical and trajectory-policy discrepancies. Table 1 compares the analyzed processes and statistical premises.

Technical Challenges. Three technical challenges arise. *First*, PPO-Clip exhibits a bidirectionally coupled learning dynamic: approximation errors in the critic can bias bootstrapped policy updates, while policy drift simultaneously induces non-stationarity in the critic's learning objective. *Second*, actor and critic iterates adapt to the same reused batch. Their conditional gradients are empirical means, while the desired guarantees concern current-policy population quantities; multistep GAE also retains the behavior policy's continuation law. *Third*, sign-dependent clipping acts on evolving, trajectory-dependent advantage labels and makes the surrogate nonsmooth. We use smoothness of the return objective J and explicitly control the clipped update's bias.

Contributions and Results. We develop a non-asymptotic analysis of PPO-Clip as a *closed-loop actor–critic system*. The central contribution is to combine and adapt policy-gradient, concentration, and tracking arguments into a unified analysis of these dependent mechanisms. By deriving their coupled errors within one specified learning process, we make their combined effects on actor stationarity and critic tracking explicit. Our contributions are:

1. Unified finite-time analysis of interacting PPO-Clip mechanisms. Theorem 3.1 jointly characterizes (i) the stationarity of the policy iterates with respect to the underlying reinforcement learning objective, and (ii) the finite-time tracking accuracy of the critic relative to its moving, policy-dependent optimum. A central outcome is an explicit *coupling strength* that provides a sufficient condition for finite-time control. The bound quantifies their feedback through a common amplification factor for stochastic noise, critic tracking, reuse-induced drift, and clipping bias. The theory yields sufficient tuning conditions within the stated model (see Sec. 3). A uniform empirical bound for the actual clipped-gradient class replaces complete-path counting in a tabular specialization. A growing-horizon example verifies every premise along actual reused-batch updates, with polynomial horizon dependence (Corollary 3.2 and Proposition J.1).

2. Finite-time refinement of classical actor–critic theory. In the asymptotic limit, the analysis recovers classical actor–critic convergence under time-scale separation, growing batch sizes, and vanishing KL budgets (Proposition 3.4). The finite-time bounds quantify this approach under constant step sizes. In particular, the analysis reveals how time-scale separation enters through an explicit coupling parameter and how minibatch reuse, clipping, and behavior-policy mismatch contribute finite-time residual terms that a vanishing-error asymptotic conclusion does not quantify. Score-baseline cancellation sharpens the coupling coefficient. An explicit schedule yields $O(T^{-2/5})$ stationarity and tracking, and a rollout-budget bound quantifies the competing effects of batch size and reuse (Corollary 3.3 and Proposition F.1). These are sufficient upper bounds, not necessary instability thresholds.

3. Extension to asynchronous PPO-Clip. Appendix K extends the analysis to a single-gradient parameter-server model (Theorem K.1). The guarantee includes staleness penalties and requires both a delay-dependent critic stepsize restriction and a sufficient coupling condition, under prescribed batch assignment and nonanticipative sampling.

Scope and limitations. This is a theoretical study of PPO-Clip under explicit coverage, value realizability, and critic regularity assumptions, using raw recomputed GAE with Monte Carlo critic targets. The guarantees have conservative constants and do not cover unrestricted neural PPO; the tabular experiments provide qualitative illustrations. The analysis informs PPO practice by identifying how actor–critic timescales, KL budgets, clipping ranges, and batch reuse affect finite-time error bounds, motivating coordinated tuning and diagnostics of critic error and policy drift. Its conditions are sufficient and problem-dependent.

2 PRELIMINARIES AND PROBLEM FORMULATION

In this section, we present our analytical setting. Further background is provided in Appendix A.

2.1 FINITE-HORIZON EPISODIC SETTING

We consider a finite-horizon episodic MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with horizon H and a fixed initial-state distribution. States include the time index. Under a parameterized policy $\pi_\theta(\cdot \mid s)$, a trajectory is

$$\tau := (s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}, s_H).$$

The reward mechanism is fixed during training. Since true advantages are unavailable in closed form, PPO estimates them from finite data using a learned value-function baseline, leading to an actor–critic architecture.

As a motivating example, LLM fine-tuning represents each action as a token from a vocabulary $\mathcal{V}$. A prompt $x \in \mathcal{X}$ is drawn from a distribution $\mathbb{P}$, and the state records the prompt and generated prefix:

$$s_t := (x, a_{0:t-1}), \qquad s_0 := (x, \emptyset).$$

Here $\mathcal{A} := \mathcal{V}$, transitions append the chosen token, $s_{t+1} = s_t \circ a_t$, and a fixed reward model evaluates the generated trajectory.

2.2 ACTOR–CRITIC FRAMEWORK

PPO is implemented as an actor–critic method in which the policy (actor) and value function (critic) are parameterized separately and updated jointly. This separation supports value-based advantage estimation when rewards are delayed. The critic enables advantage estimation and can reduce variance, but its approximation error can bias bootstrapped actor directions. Conversely, as the actor evolves, the critic's regression target drifts. This bidirectional dependency gives rise to a *closed-loop coupling* that is central to our analysis.

Critic parameterization and learning objective. The critic is a parameterized value function $V_w(s)$ trained to approximate the policy-dependent value $V_\theta(s)$ by minimizing a policy-conditioned regression loss

$$\mathcal{L}_c(\theta, w) \;:=\; \mathbb{E}_{\tau\sim\pi_\theta}\left[\frac{1}{H}\sum\nolimits_{h=0}^{H-1}(V_w(s_h) - G_h(\tau))^2\right], \tag{2.1}$$

where $G_h(\tau) = \sum_{j=h}^{H-1}\gamma^{j-h}\dot{r}(s_j, a_j) + \gamma^{H-h}v_H(s_H)$ is the Monte Carlo return, with a prescribed terminal value v_H. The theorem analyzes these stored Monte Carlo targets. Bootstrapped targets require an additional target-error bound. The explicit dependence on θ is crucial: as the actor changes, both the state distribution and the target change, turning critic learning into a tracking problem. Let $w^\star(\theta) \in \arg\min_w L_c(\theta, w)$ denote a population minimizer. We quantify critic suboptimality via the *critic tracking error*

$$\Delta(\theta, w) := \|w - w^\star(\theta)\|^2, \tag{2.2}$$

which measures the distance to the moving solution $w^\star(\theta)$. In our analysis, $\Delta(\theta, w)$ controls the bias injected into the actor update through imperfect advantage estimation.

Advantage estimation via GAE. Since the true advantage is unknown, PPO constructs an estimator using the critic. Let $\widehat{\delta}_t$ be the temporal-difference (TD) residual. Generalized advantage estimation (GAE) defines

$$\widehat{\delta}_t := \dot{r}(s_t, a_t) + \gamma V_w(s_{t+1}) - V_w(s_t), \qquad \widehat{A}_t^{(\lambda)} := \sum\nolimits_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell\widehat{\delta}_{t+\ell}, \tag{2.3}$$

where $\lambda \in [0, 1]$ trades variance against bias. At the terminal state, the estimator uses the prescribed boundary value from the return convention, so $V_w(s_H) = V_\theta(s_H)$. This value may be zero when all rewards are assigned to transitions, or the terminal payoff when that payoff is represented as a boundary value. A nonmatching terminal bootstrap requires the additional term in Lemma C.1. Bootstrapping changes the bias–variance trade-off. Our bounds quantify the critic-induced bias through explicit dependence on λ and $\Delta(\theta, w)$.

Actor update (ideal form). Let $\mathcal{J}(\theta)$ denote the canonical RL objective optimized by the actor; the policy-gradient theorem yields

$$\mathcal{J}(\theta) := \frac{1}{H}\mathbb{E}_{\tau\sim\pi_\theta}[G_0(\tau)], \qquad \nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\frac{1}{H}\sum\nolimits_{h=0}^{H-1} \gamma^h \nabla_\theta \log \pi_\theta(a_h \mid s_h)\, A^{\pi_\theta}(s_h, a_h)\right],$$

where A^{π_θ} is the true advantage for G_h. The factor $1/H$ fixes the objective normalization; the actor uses the weights γ^h (equal to one when $\gamma = 1$). Replacing the true advantage with GAE produces the idealized update direction

$$\nabla_\theta \mathcal{J}(\theta) \;\approx\; \mathbb{E}_\tau\left[\frac{1}{H}\sum\nolimits_{h=0}^{H-1} \gamma^h \nabla_\theta \log \pi_\theta(a_h \mid s_h)\, \widehat{A}_h^{(\lambda)}\right]. \tag{2.4}$$

While equation 2.4 targets the gradient of the underlying RL objective, PPO training typically reuses a finite batch of rollouts for multiple epochs. Without additional control, reuse can produce policy drift and off-policy bias, motivating the modern PPO-Clip surrogate.

2.3 Modern PPO-Clip Surrogate

In PPO-Clip, data are collected under a behavior policy $\pi_{\theta_{\text{old}}}$ and reused for multiple actor updates. This improves sample efficiency but introduces distribution shift. PPO-Clip controls this shift using importance ratios, clipping, and KL-based early stopping.

Importance sampling ratio. For any state–action pair (s, a), define the likelihood ratio

$$r_{\theta,\theta_{\text{old}}}(s,a) := \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)}.$$

This ratio allows objectives evaluated under $\pi_{\theta_{\text{old}}}$ to approximate policy updates. However, large ratios can lead to unstable gradient estimates and brittle optimization, which motivates the use of clipping.

Clipped surrogate objective. PPO-Clip maximizes the surrogate objective

$$\begin{aligned}\mathcal{L}^{\text{clip}}(\theta;\theta_{\text{old}}) := \mathbb{E}_{\substack{\tau\sim\pi_{\theta_{\text{old}}},\\ h\sim\text{Unif}\{0,\dots,H-1\}}}\Big[\gamma^h \min\big(r_{\theta,\theta_{\text{old}}}(s,a)\,\widehat{A}(s,a),\\ \text{clip}(r_{\theta,\theta_{\text{old}}}(s,a), 1-\delta, 1+\delta)\,\widehat{A}(s,a)\big)\Big],\end{aligned} \tag{2.5}$$

where $(s,a) = (s_h, a_h)$, $\widehat{A}(s,a) = \widehat{A}_h^{(\lambda)}(\tau; w)$ remains a random trajectory label, and $\delta \in (0,1)$ is the clipping parameter. Clipping suppresses selected sample gradients according to the likelihood ratio and advantage sign. It is intended to limit excessive updates and introduces two analytical challenges. First, equation 2.5 is *non-smooth* at the boundaries $r_{\theta,\theta_{\text{old}}} = 1 \pm \delta$, requiring subgradient-based arguments. Second, clipping can induce systematic *bias*: active sample directions differ from their unclipped counterparts. The latter still has trajectory-policy and critic error under reused data. A central goal of our analysis is therefore to control the impact of clipping-induced distortion.

Actor update and KL trust region. At outer iteration n, the actor performs K inner updates

$$\theta_n^{(k+1)} = \theta_n^{(k)} + \eta\, g_n^{(k)}, \tag{2.6}$$

where $g_n^{(k)}$ denotes stochastic (sub)gradients of $\mathcal{L}^{\text{clip}}(\cdot;\theta_n)$. Because updates reuse data generated by π_{θ_n}, the analysis assumes a population KL budget $\kappa > 0$ at every inner iterate $\theta_n^{(k)}$:

$$\mathbb{E}_{s\sim\pi_{\theta_n}}\Big[\text{KL}\big(\pi_{\theta_n}(\cdot \mid s) \,\|\, \pi_{\theta_n^{(k)}}(\cdot \mid s)\big)\Big] \le \kappa. \tag{2.7}$$

This trust-region constraint plays a dual role in the analysis: (i) it limits the distribution shift induced by multi-epoch reuse, ensuring that the surrogate remains meaningful; and (ii) it provides quantitative control over clipping activity, since small KL divergence implies that likelihood ratios remain close to one for most state–action pairs.

2.4 Objective of the Analysis and Assumptions

We establish non-asymptotic convergence guarantees for PPO-Clip with a learned actor–critic architecture. To keep the analysis tractable and interpretable, we impose the following regularity assumptions.

Assumption 1 (Smoothness of objective)**.** *The objective $\mathcal{J}(\cdot)$ is bounded by $J_{\max}$ and L_J-smooth.*

Assumption 2 (Bounded conditional variance)**.** *Let $\mathcal{F}_t$ denote the algorithm filtration. The stochastic actor update direction g_t satisfies $\mathbb{E}\big[\|g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]\|^2 \mid \mathcal{F}_t\big] \le \sigma_a^2$ for some $\sigma_a^2 > 0$.*

Assumption 3 (Bounded advantages)**.** *There exists $A_{\max} > 0$ such that the raw GAE estimates, recomputed using the pre-update critic, satisfy $|\widehat{A}_t^{(\lambda)}| \le A_{\max}$ for every trajectory in the data support and every visited critic. No advantage normalization or clipping is applied in the analyzed update; such transformations require a separate bias term.*

Assumption 4 (Critic learnability and stability)**.** *For each fixed policy θ, the critic loss $\mathcal{L}_c(\theta, \cdot)$ is locally convex, satisfies PL and quadratic-growth conditions with constants $\mu_c, \mu_q > 0$, and has L_c-Lipschitz gradients in w. Quadratic growth is measured from the selected stationary minimizer on a convex region containing that minimizer and the critic iterate. Moreover, the selected map $w^\star(\theta)$ is measurable and $L_{w^\star}$-Lipschitz in θ.*

Assumption 5 (Value realization)**.** *In the visited region, the value parameterization is L_V-Lipschitz with respect to w and is realizable. For the critic-to-value step, the selected population regression minimizer is assumed to satisfy $V_{w^\star(\theta)} = V^{\pi_\theta}$ on the states used in the analysis.*

Assumption 6 (KL trust-region control)**.** *During each outer iteration, the inner actor updates satisfy the population KL budget with parameter $\kappa > 0$, and the score is bounded by G_π on the relevant parameter regions. The state law in the budget is the uniform average of the H behavior-policy state marginals.*

These premises hold on regions containing all iterates used in the proof; KL control alone does not establish critic geometry or realizability. Appendix M.1 explains their roles and limitations, including value identification and the separate certification needed for empirical KL stopping.

Finite-batch protocol and coverage. The baseline finite-sample bridge uses a finite set Ω of complete trajectories (including initial states and rewards), with $S_{\mathrm{traj}} := |\Omega|$. Corollary 3.2 replaces this support premise by finite layered states and actions, a bounded tabular critic, and bounded returns; the reward distribution may then be continuous. Write P_θ for the complete trajectory law. At epoch n, conditionally on the past, collect B independent trajectories under the behavior policy $\bar{\theta}_n = \theta_n^{(0)}$, and denote their empirical law by $\widehat{P}_n$. Retain their Monte Carlo returns and reuse this batch for K joint updates. Before each update, compute raw GAE with the current critic $w_n^{(k)}$; sample trajectory–time pairs uniformly with replacement from the stored batch and $\{0, \dots, H-1\}$, or sample trajectories uniformly with replacement and average all H times. Both conventions have the same conditional mean. Use the empirical clipped actor gradient (with the advantage detached) and empirical Monte Carlo critic gradient, both evaluated at the pre-update parameters, then apply both updates. At a clipping tie, fix a measurable samplewise derivative rule in advance (for example, select the unclipped branch); it must not depend on the other samples in the minibatch. The two minibatches may coincide. Thus their noises may be correlated. The pre-update filtration contains the entire stored batch, all past draws, and the current parameters; conditional means are empirical means, not population gradients.

Assume $\pi_\theta(a \mid s) \ge p_{\min} > 0$ for every relevant action and policy, and put $R_{\max} := 1/p_{\min}$ (or any verified smaller ratio bound). This coverage condition is imposed separately; it does not follow from the average KL constraint. Assume $|G_h| \le G_{\max}$, $|V_w(s)| \le V_{\max}$, and $\|\nabla_w V_w(s)\| \le G_V$ on the visited region, with matched terminal values. Consequently every single-sample critic gradient has norm at most $Q_c := 2(V_{\max} + G_{\max})G_V$. Appendix C.4 gives the full sampling specification and derives variance bounds and uniform population discrepancies under adaptive reuse. Appendix M.1 discusses the finite-support and structured cases and their verified examples.

3 MAIN RESULTS

This section presents our main finite-time convergence guarantees for PPO-Clip. We analyze the specified *synchronous* actor–critic protocol; Appendix K extends the framework to bounded asynchronous staleness. All proofs are deferred to the appendix.

3.1 CONVERGENCE OF PPO-CLIP

We begin with our main closed-loop guarantee in the synchronous setting. The result jointly controls (i) the *stationarity* of the actor iterates with respect to the underlying RL objective $J(\theta)$ and (ii) the *tracking error* of the learned critic. Crucially, the bounds are stated in terms of explicit hyperparameters and intrinsic noise/regularity constants.

Theorem 3.1 (Closed-loop non-asymptotic convergence of PPO-Clip)**.** *Consider the finite-batch protocol in Section 2, with fixed initial actor θ_0, N batches of B trajectories, and K actual joint updates per batch; set $T = NK$. Suppose Assumptions 1–6 and the stated coverage and boundedness premises hold, with $0 < \eta \le 1/(4L_J)$ and $0 < \beta_c \le \min\{\mu_q/(2L_c^2), 1/\mu_q\}$. For matched terminal values, define*

$$\widetilde{C}_{\lambda,H} = \gamma(1-\lambda)\sum_{j=0}^{H-2}(\gamma\lambda)^j,\ a = 4G_\pi^2 L_V^2 \widetilde{C}_{\lambda,H}^2,\ \epsilon_B = \min\{1, S_{\mathrm{traj}}/(4B)\},\ E_c = 8Q_c^2\epsilon_B + 4Q_c^2 H\kappa.$$

An empty sum is zero. The state-only GAE error cancels under the on-policy score identity (Corollary C.2); thus $0 \le \widetilde{C}_{\lambda,H} \le \gamma$ and $a = 0$ when $\lambda = 1$ or $H = 1$. The empirical and clipping terms in F remain. The actor bias floor is

$$F := \underbrace{16R_{\max}^2 G_\pi^2 A_{\max}^2 \epsilon_B}_{\textit{finite-batch reuse}} + \underbrace{8G_\pi^2 A_{\max}^2 \kappa(1+\sqrt{H})^2}_{\textit{trajectory/policy drift}} + \underbrace{8G_\pi^2 A_{\max}^2 \kappa(1+1/\delta)^2}_{\textit{clipping}}. \tag{3.1}$$

Set

$$A_0 = \underbrace{\frac{4(J_{\max} - J(\theta_0))}{\eta T}}_{\textit{optimization}} + \underbrace{2L_J\eta\sigma_a^2}_{\textit{actor noise}},\quad d = \frac{40L_{w^\star}^2\eta^2}{\mu_q^2\beta_c^2},\quad U = \underbrace{\frac{4\mathbb{E}\Delta_0}{\mu_q\beta_c T}}_{\textit{critic initialization}} + \underbrace{\frac{5\beta_c\sigma_c^2}{\mu_q}}_{\textit{critic noise}} + \underbrace{\frac{40E_c}{\mu_q^2}}_{\textit{batch/target mismatch}}.$$

Define the explicit coupling coefficient $\rho := 10ad$ and suppose $\rho < 1$. Then the actor stationarity and critic tracking bounds are

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla J(\theta_t)\|^2 \le \frac{A_0 + 4F + 4aU + 4ad\sigma_a^2}{1-\rho},\quad \frac{1}{T}\sum_{t<T}\mathbb{E}\Delta_t \le \frac{U + d(\sigma_a^2 + 2A_0 + 10F)}{1-\rho}. \tag{3.2}$$

The flattened index is $t = nK + k$, and $\Delta_t = \|w_t - w^\star(\theta_t)\|^2$.

Corollary 3.2 (A batch bound from the clipped-gradient class)**.** *Retain the protocol and regularity premises of Theorem 3.1, but replace finite complete-trajectory support by finite layered state sets $\mathcal{S}_h$, finite actions, bounded rewards/returns, and a tabular critic $w \in [0,W]^{d_c}$ along every update, where $d_c = \sum_{h<H}|\mathcal{S}_h|$ and $M = \sum_{h<H}|\mathcal{S}_h||\mathcal{A}|$. Assume a positive raw-GAE envelope $A_{\max} > 0$ holds on this entire critic cube, and select the unclipped derivative at clipping ties. Put $C = C_{\lambda,H} \le 2$ and*

$$\ell_B = \log(6M) + d_c\log(2 + CWB/A_{\max}),\qquad e_c(B) = \frac{8(W^2 + G_{\max}^2)}{HB},$$
$$e_a(B) = \min\left\{4R_{\max}^2 G_\pi^2 A_{\max}^2,\ \frac{4R_{\max}^2 G_\pi^2 A_{\max}^2 M^2}{H^2 B}(\ell_B + 3)\right\}.$$

Both conclusions of the theorem hold after replacing the residuals by

$$F_{\mathrm{str}} = 4e_a(B) + 8G_\pi^2 A_{\max}^2\kappa[(1+\sqrt{H})^2 + (1+\delta^{-1})^2],\quad E_{c,\mathrm{str}} = 2e_c(B) + 4Q_c^2 H\kappa. \tag{3.3}$$

Here Q_c can be any uniform envelope of the whole-trajectory critic gradient; minibatch variances are bounded separately for the chosen sampling convention.

Lemma C.6 proves these uniform deviations for the actual clipped vector gradient, including adaptive reuse. Within each time–state–action cell the advantage multiplier is $\widehat{A}$, its positive part, or its negative part. A critic-parameter cover controls these three classes without assuming continuity of the clipped gradient in actor parameters. The actor bound is $O(M^2 d_c \log B/(H^2 B))$ times its squared envelope; it contains no complete-path cardinality. This is a structured tabular result: a prefix-tree state representation can itself have exponentially many states. Proposition J.1 verifies all premises along actual updates on a nonconstant-objective family with 2^H action paths but $2H-1$ critic coordinates, and gives a full bound polynomial in H.

Corollary 3.3 (Update and fresh-rollout complexity)**.** *Fix the problem constants, clipping range, minibatch size, and K. Consider runs with $T = NK$, stepsizes constant within each run, $\eta = c_\eta T^{-3/5}$ and $\beta_c = c_\beta T^{-2/5}$, where the positive constants ensure the stepsize restrictions and $\rho \le \rho_0 < 1$. If $\kappa_T = O(T^{-2/5})$, then both average squared stationarity and average critic parameter error are $O(T^{-2/5})$, using either $B_T = \lceil S_{\mathrm{traj}} T^{2/5} \rceil$ in the finite-support theorem or $B_T = \lceil c_B T^{2/5} \log(eT) \rceil$, $c_B > 0$, in Corollary 3.2. The fresh-trajectory count is $Q = NB = TB/K$; thus the respective sufficient counts for error at most ε are $O(\varepsilon^{-7/2})$ and $\widetilde{O}(\varepsilon^{-7/2})$, with problem-dependent constants. Environment transitions number HQ.*

Appendix F proves the rate, supplies an implementable KL certificate, and derives a fixed-rollout-budget tradeoff between B and K. The rate balances optimization, critic noise, and target movement; it is a sufficient rate for this analysis, not a minimax claim or a comparison with global optimality results for different PPO variants. The explicitly verified horizon family has polynomial, but conservative, sample complexity.

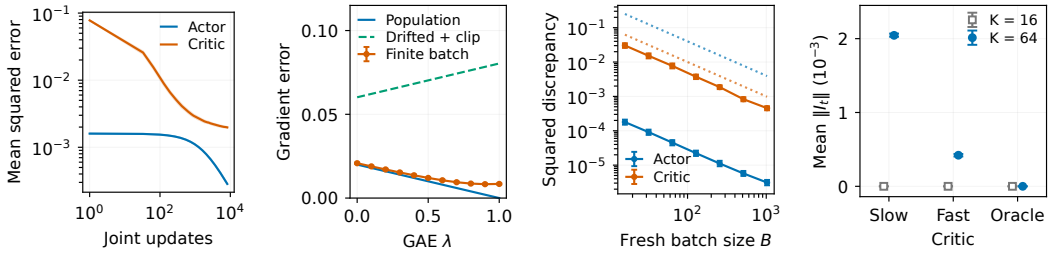
Reading the guarantee: decomposition and amplification. The bounds separate optimization and stochastic noise from finite-batch reuse, trajectory drift, clipping, and critic-target mismatch. A defining structural feature is the prefactor $(1-\rho)^{-1}$. We interpret $(1-\rho)^{-1}$ as a *coupling amplification factor*: it quantifies how actor–critic feedback amplifies *all* finite-time error sources (optimization error, noise, drift, clipping bias, and critic tracking error). The condition $\rho < 1$ suffices to close the coupled inequalities; numerical informativeness also depends on the numerator's constants and error floors (Appendix M). For nonzero coupling, it motivates time-scale separation but does not establish a universal 5–10× learning-rate ratio or a necessary instability threshold.

Practical implications and tuning guidance. The optimization contribution decays as $1/(\eta NK)$ when the remaining constants and error floors are controlled. The effect of increasing K (more mini-batch reuse) depends on the resulting policy drift. Proposition F.1 makes this trade-off explicit under a certificate $\kappa \le DK^2\eta^2$ and a fixed fresh-rollout budget. At fixed κ, the theorem does not imply that larger K increases drift. The coupling factor ρ scales as $(\eta/\beta_c)^2$ up to problem constants and $\widetilde{C}^2_{\lambda,H}$.

Once residual terms dominate the bound, further reducing its optimization term gives little improvement. The bound motivates coordinating three controls: *(i)* tightening the trust region (smaller κ, e.g., more aggressive KL-based early stopping), *(ii)* balancing reduced critic target lag against increased noise within its admissible stepsize range, and *(iii)* improving critic quality (reducing tracking error or variance).

Beyond these aggregate trade-offs, the bound surfaces several concrete PPO-Clip–specific insights. First, the KL budget κ controls two distinct failure modes simultaneously: within-epoch distribution shift ($\propto \kappa$) and clipping distortion ($\propto \kappa/\delta^2$), giving distinct reasons to control the behavior-policy mismatch. Certified predictable stopping is treated in Appendix H; an uncalibrated empirical check alone does not establish the population premise. Second, a smaller clipping range δ is not automatically more conservative: for fixed κ, the distortion term scales as κ/δ^2, so this upper bound becomes less informative for overly small clipping ranges; it does not establish a monotone law for the actual bias. Finally, bootstrapping enters the coupling through $\widetilde{C}^2_{\lambda,H}$, with $\widetilde{C}_{\lambda,H} \le \gamma \le 1$. Matched terminal values give $a = \rho = 0$ when $\lambda = 1$ or $H = 1$. Holding other constants fixed, smaller λ can tighten the sufficient stepsize-ratio bound. Other constants, including critic coverage and return bounds, can still worsen with H.

Connection to classical two-time-scale theory. Classical actor–critic analyses typically rely on asymptotic time-scale separation, often expressed as $\eta_t/\beta_{c,t} \to 0$. Theorem 3.1 provides a finite-time counterpart tailored to PPO-Clip, in which time-scale separation appears explicitly through the dimensionless coupling coefficient $\rho \;\propto\; (\eta/\beta_c)^2$ (up to problem constants and $\widetilde{C}^2_{\lambda,H}$). Thus, $\rho \ll 1$ means weak amplification in the bound; critic tracking also depends on U and d. In particular, $\lambda = 1$ or $H = 1$ gives $\rho = 0$ independently of critic speed. As $\rho \to 1$, the bound increasingly amplifies noise and bias terms. In the asymptotic limit of diminishing step sizes and vanishing trust-region budgets, the bound recovers the classical two-time-scale picture. More formally, we have the following result.



(a) Joint errors (b) GAE and residuals (c) Batch discrepancy (d) Clipping interaction

Figure 1: Two-step diagnostics: (a) joint average squared errors, (b) GAE bias and finite-batch RMS error, (c) adaptive batch discrepancy with expectation bounds (dotted). Eight-step diagnostic: (d) mean critic–clipping interaction with clipping enabled. Bands/bars show 95% bootstrap intervals.

Proposition 3.4 (Two-time-scale stationarity in a single-index t and horizon $T = NK$)**.** *Fix H, γ, λ, $\delta \in (0,1)$, and the initial actor θ_0. For each $T \in \mathbb{N}$, consider a PPO-Clip run with $N = N(T)$ outer iterations and $K = K(T)$ actual inner actor–critic updates per iteration, where $T = N(T)K(T)$ is deterministic. Assume the model and regularity premises of Theorem 3.1, including the finite-batch sampling and Monte Carlo target protocol, hold with constants uniform in T, fixed finite S_{traj}, batch size $B_T \to \infty$, and $\sup_T \mathbb{E}[\Delta_{0,0}] < \infty$. Define the flattened time index $t := nK + k, (n,k) \in \{0,\dots,N-1\} \times \{0,\dots,K-1\}$. Let $\theta_t := \theta_n^{(k)}$ and $\Delta_t := \Delta_{n,k}$ denote the corresponding actor iterate and critic tracking error, respectively. Assume the run uses positive (possibly T-dependent) hyperparameters $(\eta_T, \beta_{c,T}, \kappa_T)$ satisfying*

$$\eta_T \to 0, \quad \beta_{c,T} \to 0, \quad \kappa_T \to 0, \quad \eta_T\, T \to \infty, \quad \beta_{c,T}\, N(T) \to \infty, \quad \frac{\eta_T}{\beta_{c,T}} \to 0, \quad \frac{\eta_T}{\beta_{c,T}^2\, T} \to 0.$$

If $\widehat{t}$ is sampled uniformly from $\{0,\dots,T-1\}$ independently of the run, then

$$\mathbb{E}\big[\|\nabla J(\theta_{\widehat{t}})\|^2\big] \to 0, \qquad \mathbb{E}\big[\Delta_{\widehat{t}}\big] \to 0,$$

recovering the classical two-time-scale actor–critic stationarity and critic tracking conclusions.

Takeaway. Proposition 3.4 shows that the finite-time closed-loop bound is consistent with the classical two-time-scale actor–critic limit. What our analysis adds beyond classical two-time-scale theory is an explicit *finite-time* characterization of how this limit is approached: time-scale separation enters quantitatively through $(\eta/\beta_c)^2$ (via ρ and $(1-\rho)^{-1}$), while PPO-Clip introduces additional bias floors controlled by the KL budget κ and the clipping threshold δ through κ/δ^2. These terms describe how clipping, trust-region enforcement, and critic tracking enter the finite-time bound; the displayed diminishing schedules are sufficient to recover the asymptotic two-time-scale limit.

Asynchronous extension. The closed-loop analysis extends to a single-gradient parameter-server model with bounded actor and critic delays and a nonanticipative sampling schedule. The resulting stationarity and critic-tracking guarantees include additional staleness penalties and require a delay-dependent critic stepsize restriction alongside the sufficient coupling condition $\rho_{\mathrm{async}} < 1$. The complete model, assumptions, Theorem K.1, and its proof appear in Appendix K.

3.2 Empirical Study

Our experiments provide controlled illustrations of the main mechanisms appearing in the theory. Figures 1a–1c use the two-step example from Appendix I. Figure 1a shows that actor stationarity error and critic tracking error decrease together under joint updates, illustrating the two quantities controlled simultaneously by Theorem 3.1. Figure 1b isolates the actor-bias decomposition: the critic-induced population GAE bias vanishes at $\lambda = 1$, as predicted by the score–baseline cancellation, while finite-batch and behavior-policy/clipping residuals remain. Figure 1c then examines the statistical effect of reusing the same rollout batch: despite the actor and critic adapting to that batch, their empirical-to-population discrepancies decrease as the fresh batch size B grows, consistent with our uniform finite-batch bounds.

Figure 1d isolates the critic–clipping interaction during training on an eight-step chain. At $K = 64$, critic error changes both clipping decisions and distortion: mean $\|I_t\|$ is 2.04×10^{-3} for the slow

Table 1: Closest theoretical comparisons under each work's assumptions. Our row describes the structured tabular specialization.

Work	Actor update	Critic treatment	Statistical control	Conclusion
Jin et al. (2023)	Clipped surrogate gradients	Estimated advantages; no critic recursion	Sampling/advantage bias retained as ϕ_n	Stationarity up to bias
Huang et al. (2024)	Entropic mirror descent; neural policy regression	Neural action-value TD	Evaluation/improvement errors controlled by width and inner iterations	Global optimality for analyzed scheme
Wu et al. (2020)	Online score update with TD error	Simultaneous linear TD(0)	Markovian data; uniform mixing; critic tracking	Stationarity up to approximation error
Doering et al. (2026)	Symmetric clipped-gradient proxy	Bounded advantage bias assumed	Finite-buffer reuse with random reshuffling	Epoch-start stationarity with error floors
This work	Interleaved clipping; recomputed raw GAE	Tabular MC regression on stored returns	Uniform adaptive actor/critic batch bounds; population KL control	Joint stationarity and critic parameter tracking

critic and 4.21×10^{-4} for the fast critic. This illustrates the closed-loop mechanism in our theory: critic tracking affects clipped actor updates, while actor updates move the critic's value target. At $K = 16$, clipping is inactive; the exact-value oracle has $I_t = 0$ by construction.

These studies isolate and validate three ingredients of the theory: joint actor–critic tracking, finite-batch error under adaptive reuse, and the interaction between critic error and clipping. They are mechanism checks rather than validations of the theorem's rates or constants: the eight-step setting does not satisfy a separately verified theorem certificate, and neither bound tightness nor the predicted convergence rate is tested here. Full protocols and additional results are given in Appendix L.

4 Related Work

An extended review appears in Appendix B.

Policy gradients and actor–critic methods. Classical theory establishes stochastic policy-gradient convergence under regularity and sampling assumptions (Sutton et al., 1999; Williams, 1992). Learned critics can reduce gradient variance but can introduce bias when approximate (Schulman et al., 2016; 2017; Wen et al., 2021; Kumar et al., 2023). Asymptotic and non-asymptotic actor–critic analyses use stochastic approximation and non-convex optimization (Liu et al., 2025; Zhong & Zhang, 2023; Holzleitner et al., 2021), including finite-time critic-tracking guarantees (Wu et al., 2020). We analyze interleaved critic fitting with clipping, raw GAE, and finite-batch reuse under explicit assumptions.

Theory of PPO and clipping. Existing PPO analyses vary in policy, value-estimation, and sampling models (Huang et al., 2024; Liu et al., 2023; Huang et al., 2021; Jin et al., 2023; Liu et al., 2025; Chen & Zhao, 2023). Jin et al. (2023) retain nonsmooth clipping and multiple updates; Huang et al. (2024) include learned value estimation. These features alone do not distinguish our analysis. Table 1 compares the updates and statistical premises. Our contribution derives advantage and sampling discrepancies for interleaved raw-GAE/Monte Carlo updates and closes the resulting actor–critic recursions. For finite layered MDPs with tabular critics, Corollary 3.2 controls adaptive reuse through trajectory-function classes rather than complete-path counting; Proposition J.1 verifies a growing-horizon family with polynomial constants. These conditional stationarity and tracking guarantees cannot be directly ranked against rates under different objectives, sampling laws, or geometries.

5 Concluding Discussion

Our non-asymptotic, closed-loop analysis of PPO-Clip jointly characterizes policy stationarity and learned-critic tracking in synchronous and asynchronous settings. Under an explicit sampling protocol, it captures actor–critic coupling, nonsmooth clipping, old-policy trajectories, and finite-sample effects from multi-epoch batch reuse.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing portions of the manuscript and preparing verification code. The authors remain responsible for all claims and proofs.

REPRODUCIBILITY STATEMENT

Code is available at
`https://anonymous.4open.science/status/PPO-Analysis-D4CB`.
The appendices provide the complete set of assumptions, proofs, and protocols. Details of the controlled study are provided in the appendices.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Jacob Adkins, Michael Bowling, and Adam White. A method for evaluating hyperparameter sensitivity in reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 124820–124842, 2024.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36:7017–7049, 2023.

Leif Doering, Daniel Schmidt, Moritz Melcher, Sebastian Kassing, Benedikt Wille, Tilman Aach, and Simon Weissmann. An approximate ascent approach to prove convergence of ppo. *arXiv preprint arXiv:2602.03386*, 2026.

Rousslan Fernand Julien Dossa, Shengyi Huang, Santiago Ontañón, and Takashi Matsubara. An empirical investigation of early stopping optimizations in proximal policy optimization. *IEEE access*, 9:117981–117992, 2021.

Madeleine Dwyer, Adam Sobey, and Adriane Chapman. It's not you, it's clipping: A soft trust-region via probability smoothing for llm rl. *arXiv preprint arXiv:2509.21282*, 2025.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729*, 2020.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025a.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025b.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pp. 3610–3619. PMLR, 2021.

Yunxiao Guo, Han Long, Xiaojun Duan, Kaiyuan Feng, Maochu Li, and Xiaying Ma. Cim-ppo: proximal policy optimization with liu-correntropy induced metric. *arXiv preprint arXiv:2110.10522*, 2021.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning, 2024. URL https://arxiv.org/abs/2403.04642.

Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization. *Advances in Neural Information Processing Systems*, 35:17086–17098, 2022.

Markus Holzleitner, Lukas Gruber, José Arjona-Medina, Johannes Brandstetter, and Sepp Hochreiter. Convergence proof for actor-critic methods applied to ppo and rudder. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XLVIII: Special Issue In Memory of Univ. Prof. Dr. Roland Wagner*, pp. 105–130. Springer, 2021.

Nai-Chieh Huang, Ping-Chun Hsieh, Kuo-Hao Ho, Hsuan-Yu Yao, Kai-Chun Hu, Liang-Chun Ouyang, I Wu, et al. Neural ppo-clip attains global optimality: A hinge loss perspective. *arXiv preprint arXiv:2110.13799*, 2021.

Nai-Chieh Huang, Ping-Chun Hsieh, Kuo-Hao Ho, and I-Chen Wu. Ppo-clip attains global optimality: Towards deeper understandings of clipping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12600–12607, 2024.

Ruinan Jin, Shuai Li, and Baoxiang Wang. On stationary point convergence of ppo-clip. In *The Twelfth International Conference on Learning Representations*, 2023.

Sajad Khodadadian, Thinh T Doan, Justin Romberg, and Siva Theja Maguluri. Finite-sample analysis of two-time-scale natural actor–critic algorithm. *IEEE Transactions on Automatic Control*, 68(6): 3273–3284, 2022.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024.

Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 112(7): 2433–2467, 2023.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy, 2023. URL https://arxiv.org/abs/1906.10306.

Yin Liu, Qiming Dai, Junyu Zhang, and Zaiwen Wen. Non-asymptotic global convergence of ppo-clip. *arXiv preprint arXiv:2512.16565*, 2025.

Jared Markowitz and Edward W Staley. Clipped-objective policy gradients for pessimistic policy optimization. *arXiv preprint arXiv:2311.05846*, 2023.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking. *arXiv preprint arXiv:2501.19358*, 2025.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016. URL https://arxiv.org/abs/1602.01783.

Skander Moalla, Andrea Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No representation, no trust: connecting representation, collapse, and trust issues in ppo. *Advances in Neural Information Processing Systems*, 37:69652–69699, 2024.

Arun Nair, Praveen Srinivasan, Sam Blackwell, Cagdas Alcicek, Rory Fearon, Alessandro De Maria, Vedavyas Panneershelvam, Mustafa Suleyman, Charles Beattie, Stig Petersen, Shane Legg, Volodymyr Mnih, Koray Kavukcuoglu, and David Silver. Massively parallel methods for deep reinforcement learning, 2015. URL https://arxiv.org/abs/1507.04296.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

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

John Schulman, Philipp Moritz, Sergey Levine, Michael I Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *arXiv preprint arXiv:2509.20712*, 2025.

Mingfei Sun, Vitaly Kurin, Guoqing Liu, Sam Devlin, Tao Qin, Katja Hofmann, and Shimon Whiteson. You may not need ratio clipping in ppo. *arXiv preprint arXiv:2202.00079*, 2022.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Uncertainty in artificial intelligence*, pp. 113–122. PMLR, 2020.

Junfeng Wen, Saurabh Kumar, Ramki Gummadi, and Dale Schuurmans. Characterizing the gap between actor-critic and policy gradient. In *International conference on machine learning*, pp. 11101–11111. PMLR, 2021.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Zifan Wu, Chao Yu, Deheng Ye, Junge Zhang, Hankz Hankui Zhuo, et al. Coordinated proximal policy optimization. *Advances in Neural Information Processing Systems*, 34:26437–26448, 2021.

Yuchen Xiao, Weihao Tan, and Christopher Amato. Asynchronous actor-critic for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:4385–4400, 2022.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pp. 11204–11213. PMLR, 2020.

Rui Zheng, Wei Shen, Yuan Hua, Wenbin Lai, Shihan Dou, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Haoran Huang, Tao Gui, et al. Improving generalization of alignment with human preferences through group invariant learning. *arXiv preprint arXiv:2310.11971*, 2023.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36: 73666–73690, 2023.

Lu Zou, Wendi Ren, Weizhong Zhang, Liang Ding, and Shuang Li. Sampling complexity of td and ppo in rkhs. *arXiv preprint arXiv:2509.24991*, 2025.

APPENDIX CONTENTS AND PROOF GUIDE

The appendices provide background, complete proofs, verified examples, and experimental details. The contents and result references below are linked to their locations in the manuscript.

	Appendix	Page
A	Further Background	17
B	Extended Related Work	20
C	Useful Lemmas	22
D	Open-Loop Convergence Results	37
E	Closed-Loop Convergence Results	40
F	Rates and Rollout-Budget Consequences	41
G	The Classical Two-Time-Scale Limit	43
H	Predictable Early Stopping and Population KL Control	44
I	An Explicit Finite-Batch Actor–Critic Example	45
J	A Growing-Horizon Example with Polynomial Resources	48
K	Asynchronous PPO-Clip	53
L	Experimental Details	61
M	Discussion	65
N	Extensions and Alternative Critic Error Measures	68

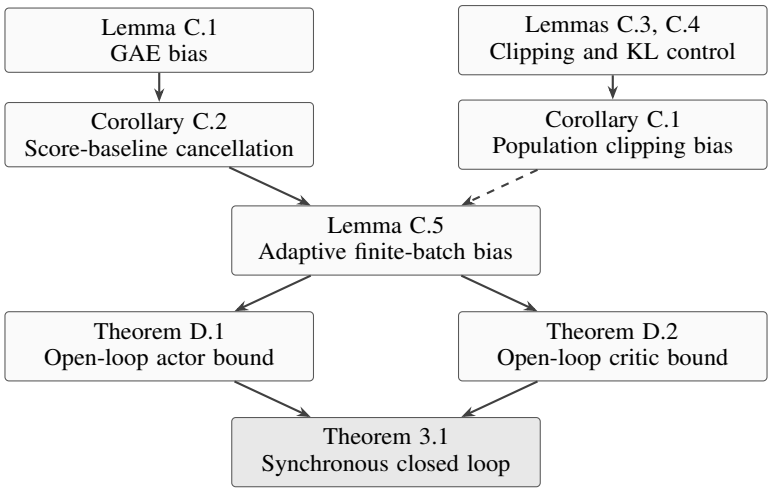
Locations of the main proofs. The results stated in the main text are proved at the following locations. The asynchronous theorem is stated and proved in Appendix K.

Result	Conclusion	Proof location
Theorem 3.1	Synchronous stationarity and critic tracking	Appendix E, p. 40
Corollary 3.2	Structured finite-batch bound	Appendix F, p. 41
Corollary 3.3	Update and fresh-rollout complexity	Appendix F, p. 41
Proposition 3.4	Classical two-time-scale limit	Appendix G, p. 43
Theorem K.1	Asynchronous stationarity and tracking	Section K.4, p. 58

Reading the dependency graphs. An arrow points from a proof ingredient to the result that uses it. Solid arrows denote applications of results; dashed arrows denote reuse or adaptation of their proof arguments. The graphs show the principal dependencies and omit routine inequalities and some transitive edges. Each result retains its own stated assumptions. Lemma C.1 and Corollary C.1 have the same numerical identifier but are distinct results, so every node includes its result type.

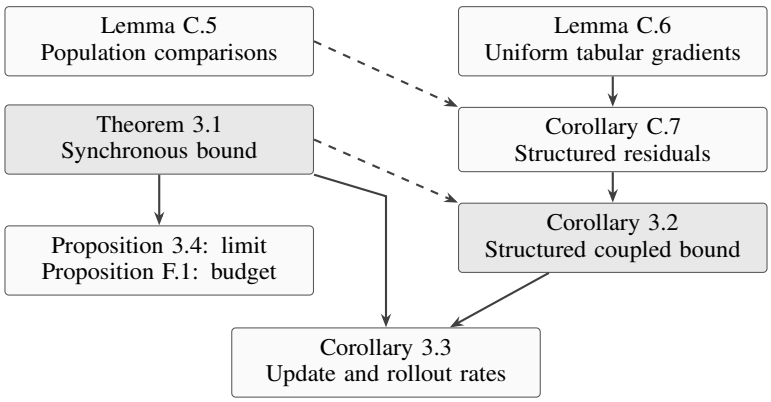
SYNCHRONOUS RESULT DEPENDENCIES

From GAE and clipping bounds to the coupled guarantee. The finite-batch bound supplies the actor bias and the critic's empirical-to-population error. The two open-loop estimates are then combined to prove the synchronous result.



The clipping argument is reproduced within Lemma C.5. Theorem D.2 proves a general biased tracking bound; the arrow into it identifies the finite-batch specialization of that bias.

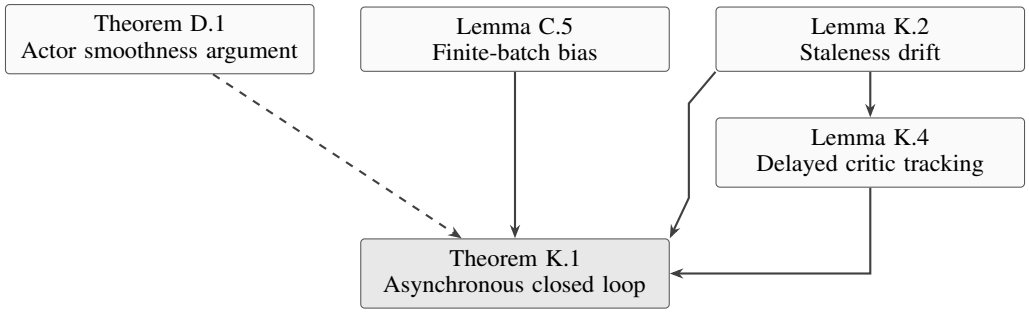
Structured statistics, rates, and the classical limit. The structured concentration lemma replaces complete-trajectory counting. Its residuals enter the same synchronous proof, after which the rate and budget conclusions follow.



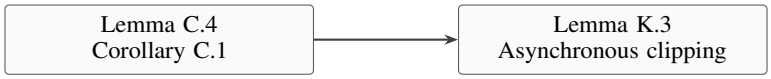
The two incoming arrows to Corollary 3.3 represent its finite-support and structured alternatives. Proposition F.1 also admits the structured residual coefficients described in Appendix F.

DEPENDENCIES FOR EXTENSIONS AND EXAMPLES

Asynchronous convergence. The asynchronous theorem combines finite-batch bias at stale parameters with a separately proved delayed-critic bound. It reuses the actor smoothness argument, rather than applying the synchronous convergence theorem directly.

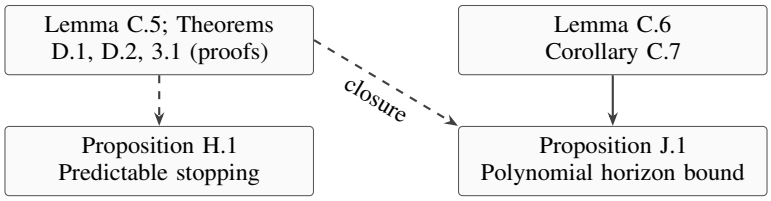


Theorem K.1 additionally requires its nonanticipative schedule, mixed-parameter regularity, delay-dependent critic stepsize, and coupling condition. The population clipping specialization is recorded separately:



The final asynchronous proof uses this clipping control through the finite-batch argument; Lemma K.3 is not a separate prerequisite for that proof.

Predictable stopping and verified examples. The stopping result adapts the one-step inequalities to a predictable gate. The growing-horizon example combines structured statistics with explicit bounds on every remaining problem constant.



Appendix I gives the two-step example, using the synchronous closure with instance-specific constants and the schedules of Corollary 3.3. Appendix J extends that construction, but derives its horizon-dependent resources explicitly. The loss-gap discussion in Section N.2 is a conditional route requiring additional drift and calibration estimates; it is not a proved replacement theorem. Section N.3 instead derives the approximation-error extension using the parameter-distance argument.

A FURTHER BACKGROUND

This appendix reviews the reinforcement-learning (RL) background for PPO-style fine-tuning of large language models (LLMs), with emphasis on two features of the analyzed training loop: *(i) multi-epoch mini-batch reuse* (multiple gradient steps on a fixed rollout batch), and *(ii) a jointly trained critic* whose approximation error and tracking lag feed back into the actor update. The *closed-loop* actor–critic coupling arises because the actor update relies on critic-based advantage estimates, while the critic is trained to regress to policy-dependent targets that drift as the actor evolves.

A.1 RL PARADIGMS AND WHERE PPO-CLIP SITS

At a high level, RL algorithms can be categorized along two orthogonal axes: *(i) on-policy vs. off-policy learning* (whether updates are computed using data from the current policy), and *(ii) value-based vs. policy-based optimization* (whether one learns a value function and derives a policy, or directly optimizes a parameterized policy). PPO belongs to the family of *on-policy policy-gradient methods* augmented with a learned value baseline (actor–critic). However, modern PPO implementations used for LLM fine-tuning depart from "pure" on-policy updates in an important way: they *reuse* a fixed on-policy rollout batch for multiple epochs. This reuse improves sample efficiency and amortizes the high cost of LLM rollouts, but it creates a *controlled distribution shift* within an outer iteration because the policy is updated while the data batch is held fixed. This training structure consists of the following outer and inner loops:

- **Outer loop (on-policy data collection).** At iteration n, a fresh batch $\mathcal{D}_n$ is collected under $\pi_{\theta_{\text{old}}}$.
- **Inner loop (multi-epoch reuse).** The batch $\mathcal{D}_n$ is reused for K inner updates, during which the actor parameters drift away from θ_{old}.

PPO-Clip controls this drift by combining *importance ratios* (to correct for mismatch), *clipping* (to prevent unstable corrections), and a *KL trust region* (to keep the mismatch small on average). Our analysis is tailored to this "weakly off-policy" regime: the updates are computed using the behavior policy $\pi_{\theta_{\text{old}}}$ but remain meaningful because the KL budget limits how far π_θ can move during reuse.

A.2 RL FOR LLM FINE-TUNING AND THE TOKEN-LEVEL MDP VIEW

We model RL fine-tuning of an LLM as a finite-horizon episodic MDP at the token level. At the beginning of each episode, a prompt $x \in \mathcal{X}$ is sampled from $\mathbb{P}$. Conditioned on x, the LLM generates a length-H sequence $(a_0, \dots, a_{H-1}) \in \mathcal{V}^H$. The state encodes the prompt and prefix,

$$s_t := (x, a_{0:t-1}), \qquad s_0 := (x, \emptyset),$$

and actions correspond to next-token choices, $\mathcal{A} := \mathcal{V}$. Transitions are deterministic, $s_{t+1} = s_t \circ a_t$. A trajectory is

$$\tau := (s_0, a_0, s_1, a_1, \dots, s_{H-1}, a_{H-1}, s_H).$$

In RLHF-style settings, per-token rewards may be sparse, and a terminal reward $\bar{r}(\tau)$ (from a learned reward model) often dominates. Given per-step rewards $\dot{r}(s_t, a_t)$ and an optional terminal reward $\bar{r}(\tau)$, the return from time t is

$$G_t(\tau) := \sum_{k=t}^{H-1} \gamma^{k-t} \dot{r}(s_k, a_k) + \gamma^{H-t} \bar{r}(\tau), \tag{A.1}$$

where $\gamma \in (0, 1]$ (often $\gamma = 1$ in finite-horizon alignment). The corresponding value functions under π_θ are

$$V^{\pi_\theta}(s) := \mathbb{E}[G_t(\tau) \mid s_t = s], \qquad Q^{\pi_\theta}(s, a) := \mathbb{E}[G_t(\tau) \mid s_t = s, a_t = a],$$

and the advantage is $A^{\pi_\theta}(s, a) := Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$. Advantages provide the variance-reduced learning signal for policy gradients, but they are not available in closed form for LLMs; they must be estimated from finite rollouts, motivating an actor–critic implementation.

A.3 Actor–critic coupling as a closed-loop learning system

Critic learning as policy-conditioned regression. PPO trains a critic $V_w(s)$ to approximate the policy-dependent value $V^{\pi_\theta}(s)$. At the population level, this can be written as a policy-conditioned regression problem

$$\mathcal{L}_c(\theta, w) := \mathbb{E}_{\tau \sim P_\theta}\left[\frac{1}{H}\sum_{h=0}^{H-1}(V_w(s_h) - G_h(\tau))^2\right], \tag{A.2}$$

where the state weighting is the uniform average over times and G_h is the Monte Carlo target in equation A.1. The analyzed update reuses stored Monte Carlo targets; bootstrapped critic targets are outside the stated theorem. Let $w^\star(\theta) \in \arg\min_w \mathcal{L}_c(\theta, w)$ denote a population minimizer. We measure critic suboptimality by the tracking error

$$\Delta(\theta, w) := \|w - w^\star(\theta)\|^2. \tag{A.3}$$

This dependence on θ is the root of the coupling: as the actor changes, both the critic's sampling distribution and its regression target drift.

Interleaved critic training. In the interleaved training pattern, the critic is updated *inside the same multi-epoch inner loop* as the actor. Each inner iteration k draws a minibatch from the fixed rollout batch $\mathcal{D}_n$ and performs:

1. compute both the Monte Carlo critic gradient and the clipped actor gradient at the pre-update parameter pair, recomputing raw GAE with that critic; and
2. apply the critic descent and actor ascent steps, allowing their minibatches to coincide,

with fixed update counts in the main theorem. Appendix H handles a predictable stopping gate with a population-KL certificate. This joint inner-loop design captures the *interleaving* between critic fitting and actor drift found in large-scale PPO codebases: the critic tracks a moving target while shaping the actor's update direction through advantages.

Advantage estimation and the bias–variance trade-off. Given a critic V_w, PPO forms temporal-difference residuals

$$\widehat{\delta}_t := \dot{r}(s_t, a_t) + \gamma V_w(s_{t+1}) - V_w(s_t),$$

and computes generalized advantage estimation (GAE)

$$\widehat{A}_t^{(\lambda)} := \sum_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell \widehat{\delta}_{t+\ell}, \tag{A.4}$$

where $\lambda \in [0, 1]$ trades variance against bias. We use the matched terminal value $V_w(s_H) = \bar{r}(\tau)$, recompute the raw estimates before each update, and detach them when differentiating the actor surrogate. Bootstrapping changes the bias–variance trade-off. The bounds quantify the critic-induced population bias through $\widetilde{C}_{\lambda,H}$ and $\Delta(\theta, w)$; they do not establish a monotone variance law.

A.4 From policy gradients to PPO-Clip and KL trust regions

Policy gradients and importance ratios. Let $J(\theta) := H^{-1}\mathbb{E}_{\tau\sim\pi_\theta}[G_0(\tau)]$ be the normalized RL objective used throughout the analysis. The policy-gradient theorem yields

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim P_\theta}\left[\frac{1}{H}\sum_{h=0}^{H-1}\gamma^h \nabla_\theta \log \pi_\theta(a_h \mid s_h)\, A^{\pi_\theta}(s_h, a_h)\right].$$

In multi-epoch reuse, updates are computed using data sampled from $\pi_{\theta_{\text{old}}}$ with importance ratios

$$r_{\theta,\theta_{\text{old}}}(s, a) := \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)}.$$

Large ratios can yield high-variance updates; the clipped surrogate limits incentives for some large ratio changes.

Clipping as stability control (and as a source of bias). The PPO-Clip surrogate suppresses the sample gradient when the ratio exceeds $1 + \delta$ for a positive advantage or falls below $1 - \delta$ for a negative advantage. The opposite-sign branches remain active. Clipping is intended to limit excessive surrogate updates but can introduce distortion relative to the unclipped importance-weighted direction. A single action ratio does not correct the complete trajectory law; Lemma C.5 bounds that remaining discrepancy. In our analysis, the impact of clipping is quantified through explicit "clip-activation" terms and their dependence on the KL budget and clipping threshold.

KL trust region and early stopping. Modern implementations often combine clipping with an empirical KL constraint (early stopping) to limit distribution shift during batch reuse. If the empirical KL on a minibatch exceeds κ, the inner loop breaks. Such an empirical statistic alone does not imply a population budget. The main theorem assumes that budget, and Appendix H certifies it for a specified predictable stopping rule. Analytically, the population budget plays two roles:

- it bounds the policy drift induced by multi-epoch reuse, keeping the surrogate meaningful; and
- it controls the prevalence of clipping by keeping likelihood ratios near one for most samples.

These mechanisms are central to obtaining interpretable finite-time bounds that reflect practical PPO behavior in LLM training.

A.5 The role of actor–critic dependency

A defining feature of PPO-Clip in practical LLM fine-tuning is that the actor and critic are *trained jointly* and *interleaved*: each inner iteration performs (at least) one critic regression step and one actor policy update on the same reused rollout batch. This produces an intrinsically *coupled* learning system rather than two independent optimizers. The following observations explain how this dependency affects the update directions, critic targets, and interaction with clipping.

(i) The actor update is computed through the critic. In PPO, the update direction is not the true policy gradient $\nabla J(\theta)$ but an estimate that depends on critic-based advantages, e.g., $\widehat{A}^{(\lambda)}$ in equation A.4. Consequently, the actor update admits a natural decomposition

$$g(\theta, w) \;=\; \nabla J(\theta) \;+\; b(\theta, w) \;+\; \zeta,$$

where $b(\theta, w)$ is the conditional-mean discrepancy from the true policy gradient and ζ is centered sampling noise. Exact values remove the population critic-induced component; empirical-batch, trajectory-policy, and clipping discrepancies can remain. Treating the critic as exact also removes its dynamic tracking error from the analysis.

(ii) The critic target drifts with the actor (tracking, not static regression). Critic learning is not a stationary supervised-learning problem. The regression target is *policy-dependent*: both the visitation distribution and the conditional return law in equation A.2 change as θ changes. Thus the critic must *track* a moving minimizer $w^\star(\theta)$ rather than converge to a fixed optimum. Neglecting this drift—for example, by treating $w^\star$ as fixed or assuming the critic "converges" between actor updates—misses an essential mechanism in PPO: actor updates induce parameter drift, which creates critic tracking error, which can feed back into subsequent actor updates. This two-way dependency is particularly acute under multi-epoch reuse: within a single outer iteration, θ moves while the batch and empirical regression targets remain fixed. The current-policy population minimizer moves, creating a mismatch between the empirical gradient and the gradient of the moving population loss.

(iii) Interleaving matters: critic and actor have comparable update frequencies. Many theoretical treatments simplify actor–critic by assuming a strict two-time-scale separation (e.g., the critic is updated "to convergence" for each actor step, or the actor is effectively fixed during critic fitting). While analytically convenient, such a separation is typically *not* representative of modern PPO practice, where both actor and critic are updated a small number of times per outer iteration and often with comparable compute budgets. Interleaved inner updates capture this training pattern; equal update frequencies do not imply equal stepsizes. If one assumes the critic is fully converged at each step, the analysis may predict negligible critic-induced bias and may incorrectly attribute observed instability primarily to clipping or stochasticity, whereas in practice critic lag and advantage mis-estimation can dominate early training and can remain non-negligible throughout.

(iv) PPO-specific mechanisms interact with critic error. PPO-Clip introduces additional nonlinearities—clipping and KL-based early stopping—whose effects depend on the advantage estimates. For instance, the sign and magnitude of $\widehat{A}(s,a)$ determine which branch of the clipped objective is active and how much gradient distortion occurs when clipping activates. If the critic underestimates or overestimates values, the resulting advantage errors can change: (i) the empirical clip-activation frequency, (ii) the effective step size implied by the surrogate, and (iii) the likelihood of triggering the KL early-stopping condition. Therefore, simplifying away critic error can also lead to an unfaithful characterization of clipping-induced bias and trust-region behavior. In other words, clipping/KL control do not operate on an abstract "true" gradient signal; they operate on a critic-dependent, data-dependent surrogate whose geometry is shaped by w.

(v) Scope of an open-loop surrogate. A common modeling shortcut is to analyze PPO as (stochastic) gradient ascent on a fixed surrogate objective $\mathcal{L}^{\mathrm{clip}}(\theta;\theta_{\mathrm{old}})$ with bounded noise. This viewpoint is useful for isolating clipping effects, but it becomes incomplete when used as a proxy for PPO dynamics in practice because:

- critic error can bias the population actor direction, although its contribution cancels when $\lambda = 1$ or $H = 1$ for a fixed critic under on-policy sampling and matched terminal values; empirical-batch, policy-drift, and clipping errors can remain;
- the critic update rule influences the distribution of advantages, and hence the surrogate geometry itself;
- the critic error evolves endogenously with actor drift (a feedback loop), rather than acting as exogenous noise.

As a result, theoretical predictions derived under open-loop assumptions (e.g., monotone improvement, stable steps under a fixed KL budget, or negligible bias floors) may fail to match observed behavior in LLM PPO runs, where the critic is learned online, targets are bootstrapped, and updates are interleaved with multi-epoch reuse.

Takeaway. The coupled updates describe feedback pathways through which critic tracking and actor drift affect PPO training. Our analysis treats PPO as a *closed-loop learning system* in which these errors co-evolve, with clipping and KL control modulating this coupling.

A.6 Objective of the analysis

The analysis establishes *non-asymptotic* convergence guarantees for PPO-Clip with a learned critic under a restricted training regime that retains interleaved fitting, clipping, and multi-epoch reuse. The actor update uses critic-dependent advantages, and the critic follows a policy-dependent regression objective whose minimizer moves with the actor. The bounds quantify this *closed-loop coupling* through the contribution of critic error to actor stationarity and the effect of actor drift on critic tracking, alongside error terms from clipping and multi-epoch reuse.

B Extended Related Work

Our work is most closely related to research on (i) policy-gradient and actor–critic theory, (ii) trust-region and proximal policy optimization, (iii) theoretical analyses of PPO and clipping, (iv) two-timescale stochastic approximation, and (v) RLHF and large language model (LLM) alignment.

Policy gradients and actor–critic methods. Classical policy-gradient theory establishes convergence of stochastic gradient ascent on the expected return under suitable regularity and sampling assumptions (Sutton et al., 1999; Williams, 1992). Actor–critic methods extend this framework by introducing a learned value function to reduce variance, which can introduce bias when approximate (Schulman et al., 2016; 2017; Wen et al., 2021; Kumar et al., 2023). A large body of work studies asymptotic and non-asymptotic convergence of actor–critic algorithms using tools from stochastic approximation and non-convex optimization, often under i.i.d. trajectory abstractions or mixing assumptions for Markovian sampling (Liu et al., 2025; Zhong & Zhang, 2023; Holzleitner et al., 2021). In particular, Wu et al. (2020) analyze simultaneous actor and linear TD(0) updates

for an average-reward objective under Markovian sampling. Their Theorems 4.5 and 4.7 already couple actor stationarity to critic tracking; Corollary 4.9 retains a function-approximation residual. These results establish finite-time actor–critic coupling under a different protocol. Our analysis also accounts for the clipped surrogate, recomputed raw GAE, and adaptive reuse of stored rollout batches.

Trust-region and proximal policy optimization. Trust-region methods such as TRPO were introduced to stabilize policy optimization by explicitly constraining policy updates using KL divergence, yielding monotonic improvement guarantees under idealized assumptions (Schulman et al., 2015; Achiam et al., 2017; Garg et al., 2021). PPO was proposed as a first-order alternative that approximates trust-region behavior using surrogate objectives, most notably probability-ratio clipping and, in some variants, explicit KL penalties (Schulman et al., 2017; Wang et al., 2020; Guo et al., 2021; Wu et al., 2021). While PPO is now the dominant algorithm in large-scale RL and RLHF, its theoretical understanding remains less complete than that of earlier trust-region methods. Our analysis treats PPO as optimizing a clipped surrogate while targeting a smooth return objective under a population KL constraint, and makes explicit how KL budgets simultaneously control within-epoch distribution shift and clipping activation.

Theory of PPO and clipping. Several works study PPO and PPO-Clip under different policy, critic, and sampling models (Huang et al., 2024; Liu et al., 2023; Huang et al., 2021; Jin et al., 2023; Liu et al., 2025; Chen & Zhao, 2023). In their ICLR 2024 paper, Jin et al. (2023) analyze repeated clipped-surrogate gradient updates and retain an advantage/sampling discrepancy ϕ_n (their Eq. (3)) in the stationarity conclusions of Theorems 3.1–3.3. They do not derive this discrepancy from an interleaved critic-fitting recursion. Our distinction is to bound the actual conditional actor and critic discrepancies under stored-batch reuse and then eliminate the tracking term through the coupled inequalities. Huang et al. (2024) learn neural action values by TD, construct a target policy by entropic mirror descent, and fit the neural actor to that target. Their Algorithm 1, Theorem 2, and Corollary 1 control policy evaluation and improvement errors through network widths and inner iterations. This establishes global optimality for a different update and sampling model, including uniformly sampled actions for policy improvement; it is not the joint raw-GAE/Monte Carlo gradient process analyzed here.

The function-space approach of Zou et al. (2025) controls kernel-TD evaluation using RKHS complexity and combines it with exponentiated natural-gradient policy improvement (their Algorithm 1 and Theorem 12). Its statistical complexity bounds concern a different policy-evaluation and improvement procedure. Our structured specialization instead controls the actual clipped-gradient trajectory class and tabular critic-gradient class uniformly over adaptively selected iterates (Corollary 3.2). Proposition J.1 verifies a growing-horizon family in which these bounds and the remaining regularity constants are polynomial despite exponentially many complete trajectories. The different update and sampling models prevent a direct comparison of rates.

Two-timescale stochastic approximation and tracking. The coupled actor–critic recursion fits naturally within the theory of two-timescale stochastic approximation, where one component tracks a moving target induced by another (Holzleitner et al., 2021; Wu et al., 2020; Zhang et al., 2020; Khodadadian et al., 2022). Classical results establish convergence under timescale separation (Holzleitner et al., 2021; Zhang et al., 2020), while more recent work provides finite-time tracking bounds that decompose error into transient, noise-floor, and drift-induced terms (Wu et al., 2020; Khodadadian et al., 2022). Our critic analysis builds on this perspective by treating the critic as a stochastic optimizer tracking the policy-dependent minimizer of a regression objective, yielding an explicit finite-time bound in which actor stepsizes directly control critic drift. This characterization is central to closing the loop in our PPO analysis.

RLHF and LLM alignment. Reinforcement learning from human feedback has become a standard paradigm for aligning large language models, typically combining learned reward models, KL regularization against a reference policy, and PPO-style optimization with minibatch reuse and multiple epochs (Ouyang et al., 2022; Schulman et al., 2016; Stiennon et al., 2020; Hilton et al., 2022; Yu et al., 2022; Sun et al., 2022). While extensive empirical work studies stability, reward hacking, and hyperparameter heuristics in RLHF, comparatively little theory addresses the finite-time behavior of PPO in this regime (Dossa et al., 2021; Miao et al., 2025; Yu et al., 2025; Fu et al., 2025a; Chen et al., 2024). RLHF motivates the present model, but the proved objective is normalized expected

return with a fixed reward model and a population KL constraint. The explicit coverage, raw-GAE, Monte Carlo-target, and critic regularity premises delimit its applicability.

Relation to our analysis. Building on these lines of work, we derive and close actor and critic bias bounds for the specified interleaved, reused-batch PPO process. The structured specialization replaces trajectory-count dependence by bounds on the relevant function classes while retaining the coverage, realizability, critic geometry, and population-KL premises. The resulting guarantees concern stationarity and critic tracking; they do not assert global optimality or cover arbitrary neural PPO or LLM fine-tuning implementations.

C Useful Lemmas

The lemmas in this appendix quantify how critic error, clipping, and batch reuse affect the actor update. They provide two components of the convergence proof.

- **Critic-induced bias control.** These results bound the effect of value-estimation error on the policy update through GAE, with explicit dependence on critic tracking error and the bootstrapping parameter.
- **Clipping- and trust-region–induced distortion control.** These results localize the non-smoothness of the PPO-Clip surrogate to clipping events and bound the resulting deviation from the policy gradient under KL control.

The bounds depend explicitly on critic tracking error, the GAE parameter λ, the clipping threshold δ, and the KL budget κ. They apply at intermediate iterates under batch reuse, without requiring either component to have converged.

C.1 Bounds for critic-based advantage estimation

We begin with lemmas controlling the accuracy of critic-based advantage estimates. In PPO, the actor update is driven by $\widehat{A}^{(\lambda)}$, not the true advantage A^{π_θ}. Discrepancies between V_w and V^{π_θ} can induce actor bias through GAE, although some value errors cancel in the population policy gradient.

The following lemma bounds the conditional mean error of GAE in terms of value-estimation error, with a coefficient depending on λ, the horizon, and the terminal convention. This is a bias bound; the variance of GAE is outside its scope.

Lemma C.1 (Finite-horizon GAE bias bound (critic-induced bias))**.** *Fix a policy π_θ, a finite horizon $H \geq 1$, and $\gamma, \lambda \in [0, 1]$. The state includes the timestep, and rewards are integrable. Let V_θ denote the true value function under π_θ for the reward and terminal-value convention used by the algorithm, and let V_w be an arbitrary fixed critic. The value function satisfies the finite-horizon Bellman equation with that terminal convention. Define $e(s) := V_w(s) - V_\theta(s)$ and assume $\|e\|_\infty < \infty$, including terminal states. The policy and critic are fixed before sampling the on-policy trajectory; the critic is held fixed within it. In particular, $e(s_H)$ is not assumed to vanish.*

For a trajectory generated under π_θ, define the TD residual based on V_w:

$$\widehat{\delta}_t^{(w)} := r_t + \gamma V_w(s_{t+1}) - V_w(s_t),$$

and the finite-horizon GAE estimator

$$\widehat{A}_t^{(\lambda,w)} \;:=\; \sum_{\ell=0}^{H-1-t} (\gamma\lambda)^\ell \widehat{\delta}_{t+\ell}^{(w)}.$$

Let

$$\bar{A}_t^{(\lambda,w)}(s,a) \;:=\; \mathbb{E}\Big[\widehat{A}_t^{(\lambda,w)} \;\Big|\; s_t = s,\, a_t = a\Big]$$

denote its conditional expectation. Then for all supported (s, a),

$$\big|\bar{A}_t^{(\lambda,w)}(s,a) - A_\theta(s,a)\big| \;\leq\; C^{\mathrm{term}}_{\lambda,H,t}\,\|e\|_\infty \;\leq\; C^{\mathrm{term}}_{\lambda,H}\,\|e\|_\infty, \tag{C.1}$$

where one may take

$$C^{\mathrm{term}}_{\lambda,H,t} := 1 + (1-\lambda)\sum_{\ell=1}^{H-1-t}(\gamma\lambda)^{\ell-1}\gamma + \gamma(\gamma\lambda)^{H-1-t}, \qquad C^{\mathrm{term}}_{\lambda,H} := \sup_{t\in\{0,\dots,H-1\}} C^{\mathrm{term}}_{\lambda,H,t}.$$

In particular, the uniform bound

$$C^{\mathrm{term}}_{\lambda,H} = 1 + \gamma \tag{C.2}$$

holds for all $\gamma, \lambda \in [0,1]$, with empty sums equal to zero and zeroth powers equal to one. If the critic and true value agree at the prescribed terminal boundary, $V_w(s_H) = V_\theta(s_H)$, the terminal term can be omitted. In this case $|\bar{A}_t^{(\lambda,w)}(s,a) - A_\theta(s,a)| \le C_{\lambda,H}\|e\|_\infty$, where

$$C_{\lambda,H} := 1 + \gamma(1-\lambda)\sum_{j=0}^{H-2}(\gamma\lambda)^j.$$

This matching condition permits a zero terminal value or a known terminal payoff.

Proof. Fix t and condition on $(s_t, a_t) = (s, a)$. We compare the conditional mean of GAE computed with V_w with the true advantage $A_\theta(s,a)$.

Step 1: Express the V_w-TD residual as a perturbation of the V_θ-TD residual. Define the TD residual based on the true value V_θ:

$$\delta_t^{(\theta)} := r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t).$$

Using $V_w = V_\theta + e$, we can write

$$\begin{aligned}\widehat{\delta}_t^{(w)} &= r_t + \gamma(V_\theta(s_{t+1}) + e(s_{t+1})) - (V_\theta(s_t) + e(s_t)) \\ &= \delta_t^{(\theta)} + \gamma e(s_{t+1}) - e(s_t).\end{aligned} \tag{C.3}$$

Step 2: Decompose $\widehat{A}_t^{(\lambda,w)}$ into the true GAE term plus an error term. By definition,

$$\begin{aligned}\widehat{A}_t^{(\lambda,w)} &= \sum_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell \widehat{\delta}_{t+\ell}^{(w)} \\ &= \sum_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell \delta_{t+\ell}^{(\theta)} \;+\; \sum_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell\big(\gamma e(s_{t+\ell+1}) - e(s_{t+\ell})\big).\end{aligned} \tag{C.4}$$

The first sum is precisely the finite-horizon GAE computed with the true value function:

$$\widehat{A}_t^{(\lambda,\theta)} := \sum_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell \delta_{t+\ell}^{(\theta)}.$$

Step 3: The true-value GAE has the correct conditional mean. For finite-horizon on-policy GAE/TD(λ),

$$\mathbb{E}\Big[\widehat{A}_t^{(\lambda,\theta)} \;\Big|\; s_t = s,\, a_t = a\Big] \;=\; A_\theta(s,a). \tag{C.5}$$

To see this, the Bellman identity gives $\mathbb{E}[\delta_t^{(\theta)} \mid s_t = s, a_t = a] = A_\theta(s,a)$. For $j > t$, let $\mathcal{H}_j$ contain the trajectory history through s_j, before drawing a_j. On-policy sampling and the Bellman identity imply $\mathbb{E}[\delta_j^{(\theta)} \mid \mathcal{H}_j] = 0$. Since (s_t, a_t) is measurable with respect to $\mathcal{H}_j$, the tower property gives $\mathbb{E}[\delta_j^{(\theta)} \mid s_t, a_t] = 0$. Summing these conditional means with the finite GAE weights proves equation C.5.

Thus, taking conditional expectation in equation C.4 and using equation C.5,

$$\bar{A}_t^{(\lambda,w)}(s,a) - A_\theta(s,a) \;=\; \mathbb{E}\left[\sum_{\ell=0}^{H-1-t}(\gamma\lambda)^\ell\big(\gamma e(s_{t+\ell+1}) - e(s_{t+\ell})\big) \;\middle|\; s_t = s,\, a_t = a\right]. \tag{C.6}$$

Step 4: Telescope the error terms with λ. Let $L := H - 1 - t$ for brevity. Rewrite the right-hand side of equation C.6 as

$$\begin{aligned} S &:= \sum_{\ell=0}^{L} (\gamma\lambda)^{\ell} \big(\gamma e(s_{t+\ell+1}) - e(s_{t+\ell})\big) \\ &= -e(s_t) + \sum_{\ell=1}^{L} \gamma^{\ell}\lambda^{\ell-1}(1-\lambda)\, e(s_{t+\ell}) \;+\; (\gamma\lambda)^{L}\gamma\, e(s_{t+L+1}). \end{aligned} \tag{C.7}$$

To verify equation C.7, collect coefficients of each $e(s_{t+j})$:

- For $j = 0$, only the $-\ell = 0$ term contributes: coefficient -1.
- For $1 \le j \le L$, $e(s_{t+j})$ appears twice: from $+\gamma(\gamma\lambda)^{j-1}e(s_{t+j})$ and from $-(\gamma\lambda)^{j}e(s_{t+j})$. The net coefficient is
$$\gamma(\gamma\lambda)^{j-1} - (\gamma\lambda)^{j} = \gamma^{j}\lambda^{j-1}(1-\lambda).$$
- The last state $e(s_{t+L+1})$ appears only from the final $+\gamma(\gamma\lambda)^{L}e(s_{t+L+1})$ term.

Now take conditional expectation of equation C.7 given $(s_t, a_t) = (s, a)$ and apply the triangle inequality:

$$\begin{aligned} \big|\bar{A}_t^{(\lambda,w)}(s,a) - A_\theta(s,a)\big| &= \big|\mathbb{E}[S \mid s_t = s, a_t = a]\big| \\ &\le \mathbb{E}\big[|e(s_t)| \mid s_t = s, a_t = a\big] \\ &\quad + (1-\lambda)\sum_{\ell=1}^{L} \gamma^{\ell}\lambda^{\ell-1}\mathbb{E}\big[|e(s_{t+\ell})| \mid s_t = s, a_t = a\big] \\ &\quad + (\gamma\lambda)^{L}\gamma\,\mathbb{E}\big[|e(s_{t+L+1})| \mid s_t = s, a_t = a\big]. \end{aligned} \tag{C.8}$$

Step 5: Sup-norm bound and the constant $C^{\mathrm{term}}_{\lambda,H,t}$. Since $|e(s)| \le \|e\|_\infty$ for all states s, equation C.8 yields

$$\big|\bar{A}_t^{(\lambda,w)}(s,a) - A_\theta(s,a)\big| \le \Big(1 + (1-\lambda)\sum_{\ell=1}^{L}\gamma^{\ell}\lambda^{\ell-1} + (\gamma\lambda)^{L}\gamma\Big)\|e\|_\infty.$$

This is exactly equation C.1 with

$$C^{\mathrm{term}}_{\lambda,H,t} = 1 + (1-\lambda)\sum_{\ell=1}^{H-1-t}(\gamma\lambda)^{\ell-1}\gamma + \gamma(\gamma\lambda)^{H-1-t}.$$

Finally, write $n = H - t$ and $q = \gamma\lambda$. Since $q \le \lambda$,

$$\begin{aligned} (1-\lambda)\sum_{j=0}^{n-2} q^{j} + q^{n-1} &\le (1-\lambda)\sum_{j=0}^{n-2}\lambda^{j} + \lambda^{n-1} \\ &= 1. \end{aligned}$$

Thus $C^{\mathrm{term}}_{\lambda,H,t} \le 1 + \gamma$, with equality at $t = H - 1$. This proves equation C.2. If $e(s_H) = 0$, deleting the terminal term in equation C.8 gives the stated sharper coefficient. The argument is on-policy and does not control old-policy sampling, a critic fitted adaptively to the same trajectory, or further normalization or clipping of GAE. □

Corollary C.2 (Baseline cancellation in the population actor bias)**.** *Under Lemma C.1, suppose the terminal values match and the policy has a differentiable finite-action distribution with bounded score $z_h = \nabla_\theta \log \pi_\theta(a_h \mid s_h)$, $\|z_h\| \le G_\pi$. For the normalized objective $J = \mathbb{E}G_0/H$, define*

$$\widetilde{C}_{\lambda,H} := C_{\lambda,H} - 1 = \gamma(1-\lambda)\sum_{j=0}^{H-2}(\gamma\lambda)^{j} \le \gamma.$$

Then

$$\left\| \frac{1}{H} \mathbb{E}_{P_\theta} \sum_{h=0}^{H-1} \gamma^h z_h \widehat{A}_h^{(\lambda,w)} - \nabla J(\theta) \right\| \le G_\pi \widetilde{C}_{\lambda,H} \|V_w - V_\theta\|_\infty. \tag{C.9}$$

In particular, this population critic-induced bias is zero if $\lambda = 1$ or $H = 1$. This conclusion concerns the on-policy expectation with a fixed critic; it does not remove the empirical, clipping, or old-policy discrepancies.

Proof. For each fixed state, $\mathbb{E}[z_h \mid s_h] = \sum_a \nabla_\theta \pi_\theta(a \mid s_h) = 0$. Consequently the leading term $-e(s_h)$ in equation C.7 cancels after multiplication by z_h and expectation. The matched terminal term is zero. The remaining nonnegative coefficients sum to at most $\gamma(1-\lambda)\sum_{j=0}^{H-2}(\gamma\lambda)^j$ at every time. Use $\|z_h\| \le G_\pi$, $\gamma^h \le 1$, and the policy-gradient identity. Finally, $(1-\lambda)\sum_{j=0}^{H-2}(\gamma\lambda)^j \le 1$, including the endpoints by the finite-sum convention. □

C.2 Handling PPO-Clip non-smoothness via event localization

A major analytical difficulty in PPO arises from the non-smooth nature of the clipped surrogate objective. The objective is not differentiable at likelihood ratios $r_{\theta,\theta_{\text{old}}} = 1 \pm \delta$, which precludes a direct application of standard smooth optimization arguments.

The key observation underlying our analysis is that this non-smoothness is *localized*: it only affects samples at or beyond the relevant sign-dependent clipping boundary. We enclose the sign-dependent clip-active region and its boundary in the ratio event

$$E_t := \{\, |r_{\theta_t,\theta_{\text{old}}}(s,a) - 1| \ge \delta \,\}.$$

On the complement of this event, the sample derivative coincides with the unclipped importance-weighted surrogate derivative. This does not by itself make the reused-batch update unbiased for the policy gradient.

The next lemma makes this localization explicit by showing that the clipped gradient agrees with the unclipped gradient whenever E_t does not occur. This allows us to decompose PPO updates into a smooth component and a clipping-induced distortion term that is weighted by $\mathbb{P}(E_t)$.

Lemma C.3 (PPO-Clip gradient localization)**.** *Let $L^{\text{clip}}(\theta;\bar\theta_t)$ be the PPO-Clip objective computed with a fixed behavior policy $\pi_{\bar\theta_t}$. Policies are probabilities or densities with respect to a common action measure, with a fixed common support and positive values there. On a parameter ball U of radius $R > 0$ centered at θ, assume the current log-policy is continuously differentiable and $\|\nabla_\vartheta \log \pi_\vartheta(a \mid s)\| \le G_\pi$ for almost every relevant (s,a), uniformly over $\vartheta \in U$. Let P be a fixed joint law of $(s,a,\widehat{A})$ with state–action marginal given by $s \sim \nu$, $a \sim \pi_{\bar\theta_t}(\cdot \mid s)$, and $|\widehat{A}| \le A_{\max}$. The advantage label and its law are held fixed when differentiating, while its sign remains inside the expectations. Fix $\delta \in (0,1)$ and write $r_\theta = \pi_\theta/\pi_{\bar\theta_t}$. Define the active boundary $B_\theta := \{\widehat{A} > 0, r_\theta = 1+\delta\} \cup \{\widehat{A} < 0, r_\theta = 1-\delta\}$. For any θ with $P(B_\theta) = 0$, the population objective is differentiable and*

$$\nabla_\theta L^{\text{clip}}(\theta;\bar\theta_t) = \mathbb{E}_P\Big[\nabla_\theta r_{\theta,\bar\theta_t}(s,a)\, \widehat{A}\, \mathbf{1}\{\neg E(\theta;s,a,\widehat{A})\}\Big]. \tag{C.10}$$

On B_θ, a sample Clarke derivative is selected by a measurable coefficient $\alpha \in [0,1]$ multiplying $\nabla_\theta r_\theta \widehat{A}$. Off B_θ, take $\alpha = 0$ on the strict clip-active event E and $\alpha = 1$ otherwise. The expected selected sample direction is then $\mathbb{E}_P[\alpha \nabla_\theta r_\theta \widehat{A}]$; the indicator formula need not hold when $P(B_\theta) > 0$.

Proof. Fix an update index t and suppress the dependence on t when it is clear. Let the behavior policy be $\pi_{\bar\theta_t}$ and define the likelihood ratio

$$r_\theta(s,a) \;:=\; r_{\theta,\bar\theta_t}(s,a) \;=\; \frac{\pi_\theta(a \mid s)}{\pi_{\bar\theta_t}(a \mid s)}.$$

Condition on a sample $(s,a,\widehat{A})$ and hold its advantage label fixed for differentiation. The label remains random under the joint expectation P. The PPO-Clip objective can be written as

$$\begin{aligned} L^{\text{clip}}(\theta;\bar\theta_t) \;&=\; \mathbb{E}_P\big[\ell^{\text{clip}}(\theta;s,a)\big], \\ \ell^{\text{clip}}(\theta;s,a) \;&:=\; \min\Big\{ r_\theta(s,a)\widehat{A},\; \text{clip}(r_\theta(s,a), 1-\delta, 1+\delta)\widehat{A} \Big\}. \end{aligned} \tag{C.11}$$

Step 1: Express the surrogate as a scalar piecewise function. For fixed (s, a), $\ell^{\text{clip}}(\theta; s, a)$ depends on θ only through the scalar $r_\theta(s, a)$. Define the scalar function (parameterized by $\widehat{A}$)

$$\psi_{\widehat{A}}(u) \;:=\; \min\Big\{u\,\widehat{A},\; \text{clip}(u, 1-\delta, 1+\delta)\,\widehat{A}\Big\}, \qquad u \geq 0.$$

Then for each sample (s, a),

$$\ell^{\text{clip}}(\theta; s, a) \;=\; \psi_{\widehat{A}}\big(r_\theta(s, a)\big). \tag{C.12}$$

Step 2: Compute $\psi_{\widehat{A}}$ explicitly and identify the clip-active region. Recall

$$\text{clip}(u, 1-\delta, 1+\delta) = \begin{cases} 1-\delta, & u < 1-\delta, \\ u, & 1-\delta \leq u \leq 1+\delta, \\ 1+\delta, & u > 1+\delta. \end{cases}$$

For $\widehat{A} = 0$, the sample function and all its derivatives are zero.

Case 1: $\widehat{A} > 0$. Since $\widehat{A} > 0$, the map $u \mapsto u\widehat{A}$ is increasing in u and

$$\psi_{\widehat{A}}(u) = \widehat{A} \cdot \min\Big\{u,\; \text{clip}(u, 1-\delta, 1+\delta)\Big\} = \widehat{A} \cdot \min\{u,\, 1+\delta\}.$$

Indeed, if $u > 1+\delta$, then $\text{clip}(u, \cdot) = 1+\delta < u$ so the minimum is the clipped value; otherwise the minimum is u. Thus,

$$\psi_{\widehat{A}}(u) = \begin{cases} u\widehat{A}, & u < 1+\delta, \\ (1+\delta)\widehat{A}, & u > 1+\delta, \end{cases} \quad \text{and it is non-differentiable at } u = 1+\delta. \tag{C.13}$$

Hence, for $\widehat{A} > 0$, clipping is active exactly on $\{u > 1+\delta\}$.

Case 2: $\widehat{A} < 0$. Since $\widehat{A} < 0$, multiplying by $\widehat{A}$ reverses inequalities, and

$$\psi_{\widehat{A}}(u) = \widehat{A} \cdot \max\Big\{u,\; \text{clip}(u, 1-\delta, 1+\delta)\Big\} = \widehat{A} \cdot \max\{u,\, 1-\delta\}.$$

Indeed, if $u < 1-\delta$, then $\text{clip}(u, \cdot) = 1-\delta > u$ so the maximum is the clipped value; otherwise the maximum is u. Thus,

$$\psi_{\widehat{A}}(u) = \begin{cases} (1-\delta)\widehat{A}, & u < 1-\delta, \\ u\widehat{A}, & u > 1-\delta, \end{cases} \quad \text{and it is non-differentiable at } u = 1-\delta. \tag{C.14}$$

Hence, for $\widehat{A} < 0$, clipping is active exactly on $\{u < 1-\delta\}$.

Combining the two cases gives the clip-active event

$$E(\theta; s, a, \widehat{A}) := \Big\{\widehat{A} > 0,\; r_\theta(s, a) > 1+\delta\Big\} \;\cup\; \Big\{\widehat{A} < 0,\; r_\theta(s, a) < 1-\delta\Big\}. \tag{C.15}$$

On $E(\theta; s, a, \widehat{A})$, $\psi_{\widehat{A}}(r_\theta)$ equals a *constant* (either $(1+\delta)\widehat{A}$ or $(1-\delta)\widehat{A}$), while on its complement it equals $r_\theta\widehat{A}$.

Step 3: Pointwise gradient for differentiable points. Assume θ is such that $\ell^{\text{clip}}(\theta; s, a)$ is differentiable for P-almost every sample $(s, a, \widehat{A})$. From equation C.13–equation C.14, whenever $\psi_{\widehat{A}}$ is differentiable at $u = r_\theta(s, a)$ we have

$$\psi'_{\widehat{A}}(u) = \begin{cases} \widehat{A}, & \neg E(\theta; s, a, \widehat{A}), \\ 0, & E(\theta; s, a, \widehat{A}), \end{cases} \qquad \text{(ignoring the boundary points } u \in \{1 \pm \delta\}\text{)}.$$

By the chain rule and equation C.12,

$$\nabla_\theta \ell^{\text{clip}}(\theta; s, a) = \psi'_{\widehat{A}}\big(r_\theta(s, a)\big)\, \nabla_\theta r_\theta(s, a) = \nabla_\theta r_\theta(s, a)\, \widehat{A}\, \mathbf{1}\{\neg E(\theta; s, a, \widehat{A})\}. \tag{C.16}$$

Step 4: Interchange gradient and expectation. The joint law P does not depend on θ. Integrating the uniformly bounded score along the segment from θ to $\vartheta \in U$ gives

$$r_\vartheta(s,a) \le e^{G_\pi R} r_\theta(s,a), \qquad \|\nabla_\vartheta r_\vartheta(s,a)\| \le G_\pi e^{G_\pi R} r_\theta(s,a).$$

Since $u \mapsto \psi_{\widehat{A}}(u)$ is $|\widehat{A}|$-Lipschitz, sample difference quotients and selected derivative norms are bounded by $M(s,a) := A_{\max} G_\pi e^{G_\pi R} r_\theta(s,a)$. Common support and normalization imply $\mathbb{E}_P[r_\theta] = 1$, so $\mathbb{E}_P[M] = A_{\max} G_\pi e^{G_\pi R} < \infty$. This provides an integrable local envelope without finite state/action sets or a pointwise likelihood-ratio cap. Hence, by dominated convergence (or standard differentiation-under-the-integral rules),

$$\nabla_\theta L^{\mathrm{clip}}(\theta;\bar\theta_t) = \mathbb{E}_P\big[\nabla_\theta \ell^{\mathrm{clip}}(\theta;s,a)\big].$$

Substituting the pointwise identity equation C.16 yields

$$\nabla_\theta L^{\mathrm{clip}}(\theta;\bar\theta_t) = \mathbb{E}_P\Big[\nabla_\theta r_{\theta,\bar\theta_t}(s,a)\,\widehat{A}\,\mathbf{1}\{\neg E(\theta;s,a,\widehat{A})\}\Big],$$

which is exactly equation C.10.

Step 5: Non-differentiable boundary and Clarke subgradients. The only non-differentiable points of $\psi_{\widehat{A}}$ occur at $u = 1+\delta$ when $\widehat{A} > 0$ and at $u = 1-\delta$ when $\widehat{A} < 0$. At those points the (scalar) Clarke subdifferential is

$$\partial^\circ \psi_{\widehat{A}}(u) = \begin{cases} \mathrm{conv}\{0,\widehat{A}\}, & \widehat{A} > 0,\ u = 1+\delta, \\ \mathrm{conv}\{0,\widehat{A}\}, & \widehat{A} < 0,\ u = 1-\delta, \\ \{\widehat{A}\}, & \neg E(\theta;s,a,\widehat{A}) \text{ and not on boundary}, \\ \{0\}, & E(\theta;s,a,\widehat{A}) \text{ and not on boundary}. \end{cases}$$

By the chain rule for Clarke subgradients of compositions with smooth maps,

$$\partial^\circ_\theta \ell^{\mathrm{clip}}(\theta;s,a) = \partial^\circ \psi_{\widehat{A}}\big(r_\theta(s,a)\big) \cdot \nabla_\theta r_\theta(s,a),$$

so any measurable selection $\widetilde{g}(\theta;s,a) \in \partial^\circ_\theta \ell^{\mathrm{clip}}(\theta;s,a)$ can be written as

$$\widetilde{g}(\theta;s,a) = \alpha(\theta;s,a,\widehat{A})\,\nabla_\theta r_\theta(s,a)\,\widehat{A},$$

$$\text{with } \alpha(\theta;s,a,\widehat{A}) = \begin{cases} 1, & \neg E(\theta;s,a,\widehat{A}) \text{ (away from boundary)}, \\ 0, & E(\theta;s,a,\widehat{A}) \text{ (away from boundary)}, \\ \in [0,1], & \text{on the boundary}. \end{cases}$$

At a zero advantage label one may take $\alpha = 1$ without changing the derivative. If $P(B_\theta) = 0$, the choice of boundary selector does not affect the expectation, and equation C.10 follows from Step 4. When the boundary has positive mass, the joint expectation includes the chosen selector:

$$g_{\mathrm{sel}}(\theta) = \mathbb{E}_P[\alpha(\theta;s,a,\widehat{A})\nabla_\theta r_\theta(s,a)\,\widehat{A}].$$

The clipping difference is supported on $E \cup B_\theta \subseteq \{|r_\theta - 1| \ge \delta\}$. No equality between this expression and an arbitrary population Clarke subgradient is assumed.

This completes the proof. □

C.3 Bounding clipping-induced bias

Clipping changes active sample directions and can introduce population bias. The following results control this distortion through the clipping event and the advantage and score envelopes. They contribute a KL-dependent term to the main theorem; an active clipping event need not produce nonzero aggregate bias.

Lemma C.4 (KL control bounds clip activation)**.** *For any $\delta \in (0,1)$ and any state s, let $r(a) = \pi_{\theta_t}(a \mid s)/\pi_{\bar\theta_t}(a \mid s)$. Both policies have probability mass functions or densities with respect to a common action measure, share a fixed support, and are positive there. The state law may be arbitrary. For densities, the sums below are interpreted as integrals. Then*

$$\Pr_{a\sim\pi_{\bar\theta_t}(\cdot|s)}\big(|r(a)-1| \ge \delta\big) \;\le\; \frac{\sqrt{2\,\mathrm{KL}(\pi_{\bar\theta_t}(\cdot \mid s)\,\|\,\pi_{\theta_t}(\cdot \mid s))}}{\delta}. \tag{C.17}$$

Consequently, for any fixed state law ν satisfying the population condition $\mathbb{E}_{s\sim\nu}[\mathrm{KL}(\pi_{\bar{\theta}_t}(\cdot \mid s)\|\pi_{\theta_t}(\cdot \mid s))] \le \kappa$, the boundary-inclusive clip-active event E_t defined below satisfies

$$\Pr(E_t) \le \frac{\sqrt{2\kappa}}{\delta}. \tag{C.18}$$

Proof. Fix an arbitrary time index t and a state s. For brevity, define the conditional action distributions

$$p(a) := \pi_{\bar{\theta}_t}(a \mid s), \qquad q(a) := \pi_{\theta_t}(a \mid s), \qquad r(a) := \frac{q(a)}{p(a)}.$$

Actions with $p(a) = q(a) = 0$ may be omitted; a current-policy action outside the behavior support is excluded.

Step 1: A tail bound via Markov's inequality and total variation. Let

$$A_\delta := \{a : \ |r(a) - 1| \ge \delta\}.$$

By Markov's inequality applied to the nonnegative random variable $|r(a) - 1|$ under $a \sim p$,

$$\Pr_{a\sim p}(A_\delta) \ = \ \Pr_{a\sim p}\big(|r(a) - 1| \ge \delta\big) \ \le \ \frac{1}{\delta}\,\mathbb{E}_{a\sim p}\big[|r(a) - 1|\big]. \tag{C.19}$$

We now compute the expectation explicitly:

$$\mathbb{E}_{a\sim p}\big[|r(a) - 1|\big] = \sum_a p(a)\left|\frac{q(a)}{p(a)} - 1\right| = \sum_a \big|q(a) - p(a)\big| \ = \ 2\,\mathrm{TV}(p, q). \tag{C.20}$$

Substituting equation C.20 into equation C.19 yields

$$\Pr_{a\sim p}\big(|r(a) - 1| \ge \delta\big) \ \le \ \frac{2\,\mathrm{TV}(p, q)}{\delta}. \tag{C.21}$$

Step 2: Convert the total-variation bound to a KL bound. By Pinsker's inequality,

$$\mathrm{TV}(p, q) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(p\|q)}. \tag{C.22}$$

Combining equation C.21 and equation C.22 gives

$$\Pr_{a\sim p}\big(|r(a) - 1| \ge \delta\big) \ \le \ \frac{2}{\delta}\sqrt{\frac{1}{2}D_{\mathrm{KL}}(p\|q)} \ = \ \frac{\sqrt{2\,D_{\mathrm{KL}}(p\|q)}}{\delta},$$

which is exactly equation C.17 after substituting back $p(\cdot) = \pi_{\bar{\theta}_t}(\cdot \mid s)$ and $q(\cdot) = \pi_{\theta_t}(\cdot \mid s)$.

Step 3: Average over the state distribution. Let $d_{\bar{\theta}_t} := \nu$ denote the specified state distribution under which the population KL budget is enforced, and write

$$K(s) := D_{\mathrm{KL}}\big(\pi_{\bar{\theta}_t}(\cdot \mid s) \,\|\, \pi_{\theta_t}(\cdot \mid s)\big).$$

Taking expectation of equation C.17 over $s \sim d_{\bar{\theta}_t}$ yields

$$\begin{aligned}\Pr_{s\sim d_{\bar{\theta}_t},\, a\sim\pi_{\bar{\theta}_t}(\cdot|s)}\big(|r(a) - 1| \ge \delta\big) &= \mathbb{E}_{s\sim d_{\bar{\theta}_t}}\Big[\Pr_{a\sim\pi_{\bar{\theta}_t}(\cdot|s)}\big(|r(a) - 1| \ge \delta\big)\Big] \\ &\le \frac{\sqrt{2}}{\delta}\,\mathbb{E}_{s\sim d_{\bar{\theta}_t}}\big[\sqrt{K(s)}\big].\end{aligned} \tag{C.23}$$

Since $\sqrt{\cdot}$ is concave on $\mathbb{R}_+$, Jensen's inequality gives

$$\mathbb{E}_{s\sim d_{\bar{\theta}_t}}\big[\sqrt{K(s)}\big] \ \le \ \sqrt{\mathbb{E}_{s\sim d_{\bar{\theta}_t}}[K(s)]}. \tag{C.24}$$

Using the stated population KL budget,

$$\mathbb{E}_{s\sim d_{\bar{\theta}_t}}[K(s)] = \mathbb{E}_{s\sim d_{\bar{\theta}_t}} D_{\mathrm{KL}}\big(\pi_{\bar{\theta}_t}(\cdot \mid s) \,\|\, \pi_{\theta_t}(\cdot \mid s)\big) \ \le \ \kappa,$$

and combining with equation C.23–equation C.24 yields

$$\Pr_{s\sim d_{\bar{\theta}_t},\, a\sim\pi_{\bar{\theta}_t}(\cdot|s)}\big(|r(a) - 1| \ge \delta\big) \ \le \ \frac{\sqrt{2\kappa}}{\delta}. \tag{C.25}$$

Step 4: Relating clip activation to likelihood-ratio deviation. Include the active boundaries in the clip-active event (for a sample $(s, a, \widehat{A})$):

$$E_t \;:=\; \Big\{\widehat{A} > 0,\; r(a) \geq 1+\delta\Big\} \;\cup\; \Big\{\widehat{A} < 0,\; r(a) \leq 1-\delta\Big\}.$$

In either case, E_t implies $|r(a) - 1| \geq \delta$. Hence, regardless of the distribution of $\widehat{A}$,

$$\Pr(E_t) \;\leq\; \Pr\big(|r(a) - 1| \geq \delta\big).$$

Combining this with equation C.25 proves equation C.18:

$$\Pr(E_t) \leq \frac{\sqrt{2\kappa}}{\delta}.$$

□

Corollary C.1 (Population clipping distortion)**.** *Under the preceding lemma, let P be a fixed joint law of $(s, a, \widehat{A})$ with state–action marginal $\nu(s)\pi_{\bar{\theta}_t}(a \mid s)$. Suppose $|\widehat{A}| \leq A_{\max}$ and $\|\nabla \log \pi_{\theta_t}(a \mid s)\| \leq G_\pi$. For a sample selector α as in Lemma C.3, the clipping difference from the unclipped direction under the same law is*

$$c_t := \mathbb{E}_P[(1-\alpha) r \nabla \log \pi_{\theta_t}(a \mid s)\widehat{A}].$$

Then, without a pointwise likelihood-ratio cap,

$$\|c_t\|^2 \leq 2G_\pi^2 A_{\max}^2 \kappa (1+\delta^{-1})^2 \leq 8G_\pi^2 A_{\max}^2 \kappa/\delta^2. \tag{C.26}$$

Proof. Equation C.20, Pinsker's inequality, and Jensen's inequality give $\mathbb{E}_P|r-1| \leq \sqrt{2\kappa}$. Since $0 \leq 1-\alpha \leq \mathbf{1}_{E_t}$, we have

$$\begin{aligned}\|c_t\| &\leq G_\pi A_{\max} \mathbb{E}_P[r \mathbf{1}_{E_t}] \\ &\leq G_\pi A_{\max}\big(P(E_t) + \mathbb{E}_P|r-1|\big) \\ &\leq G_\pi A_{\max}\sqrt{2\kappa}(1+\delta^{-1}).\end{aligned}$$

Squaring and using $\delta < 1$ proves the claim. The random advantage and its sign remain inside the joint expectation throughout. □

The corollary compares two population surrogate directions. Relating the unclipped direction to the current-policy return gradient, or relating empirical KL and adaptive batch averages to population quantities, requires additional arguments.

Role in the convergence proof. Smoothness of the return objective J relates each actor update to progress in the objective. The localization and KL lemmas bound clipping distortion, while the GAE lemma controls the current-policy advantage error. The finite-batch bridge below adds the empirical-law and complete-trajectory discrepancies and combines all four terms. These bounds then enter the coupled analysis of actor stationarity and critic tracking.

C.4 Finite-batch bias under adaptive rollout reuse

The following lemma bounds the discrepancy between empirical actor and critic directions and their current-policy population counterparts. Resampling a stored batch leaves a statistical error because its empirical gradients can differ from the population gradients.

Lemma C.5 (Adaptive finite-batch actor and critic bias)**.** *Let P_θ be the law of a length-H trajectory, including its prompt and rewards, on a common finite support Ω of cardinality S_{traj}. The environment, reward, and prompt kernels are policy independent, and policies have common positive action support. At outer iteration n, conditional on the history $\mathcal{G}_n$ before collection, draw B independent trajectories under the $\mathcal{G}_n$-measurable behavior parameter $\bar{\theta} = \theta_n$. Write $\widehat{P}_n$ for their empirical law and*

$$\epsilon_B := \min\left\{1, \frac{S_{\mathrm{traj}}}{4B}\right\}.$$

For each deterministic inner index $t = nK + k$, the pre-update history $\mathcal{F}_t$ contains this batch and (θ_t, w_t), which may depend on all preceding uses of the batch. Uniform minibatch sampling with replacement is the only additional randomness in the two update directions.

Define the return, normalized objective, and population critic loss by

$$G_h(\tau) := \sum_{j=h}^{H-1} \gamma^{j-h} r_j + \gamma^{H-h} v_{\mathrm{term}}(s_H),$$

$$J(\theta) := \frac{1}{H}\mathbb{E}_{P_\theta} G_0, \qquad \mathcal{L}_c(\theta, w) := \mathbb{E}_{P_\theta} \frac{1}{H} \sum_{h=0}^{H-1} (V_w(s_h) - G_h(\tau))^2. \tag{C.27}$$

Here $\gamma \in (0, 1]$, the state includes time, and the critic matches the prescribed terminal boundary. At every update, raw GAE is recomputed on the stored paths using the current pre-update critic, held fixed along each trajectory:

$$\widehat{A}_h(w, \tau) = \sum_{\ell=0}^{H-1-h} (\gamma\lambda)^\ell [r_{h+\ell} + \gamma V_w(s_{h+\ell+1}) - V_w(s_{h+\ell})], \qquad \lambda \in [0, 1].$$

The actor holds these labels fixed when differentiating the surrogate and weights the token-h direction by γ^h. The critic uses the stored Monte Carlo targets G_h, which do not depend on w. Both directions use pre-update parameters; the actor and critic minibatches may be shared. Fix a measurable samplewise clipping-tie derivative rule in advance, independently of the other minibatch samples.

Assume uniformly over supported trajectories and visited parameters that

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\| \le G_\pi, \qquad |\widehat{A}_h(w, \tau)| \le A_{\max}, \qquad \frac{\pi_\theta(a \mid s)}{\pi_{\bar\theta}(a \mid s)} \le R_{\max},$$

$$|V_w(s)| \le V_{\max}, \qquad |G_h(\tau)| \le G_{\max}, \qquad \|\nabla_w V_w(s)\| \le G_V.$$

The ratio envelope is a separate coverage premise: a uniform behavior-action probability floor $p_{\min} > 0$ permits $R_{\max} = 1/p_{\min}$. For the selected critic minimizer assume

$$V_{w^\star(\theta)} = V^{\pi_\theta}, \qquad \|V_w - V_{w^\star(\theta)}\|_\infty \le L_V \|w - w^\star(\theta)\|.$$

At each such inner iterate, suppose the population KL budget holds almost surely:

$$\frac{1}{H} \sum_{h=0}^{H-1} \mathbb{E}_{s_h \sim P_{\bar\theta}} \mathrm{KL}\big(\pi_{\bar\theta}(\cdot \mid s_h) \,\|\, \pi_{\theta_t}(\cdot \mid s_h)\big) \le \kappa. \tag{C.28}$$

Let $\Delta_t = \|w_t - w^\star(\theta_t)\|^2$, and define

$$a := 4G_\pi^2 L_V^2 \widetilde{C}_{\lambda,H}^2, \qquad Q_c := 2(V_{\max} + G_{\max})G_V,$$

$$F_{B,\kappa,\delta} := 16R_{\max}^2 G_\pi^2 A_{\max}^2 \epsilon_B + 8G_\pi^2 A_{\max}^2 \kappa\big[(1 + \sqrt{H})^2 + (1 + \delta^{-1})^2\big],$$

$$E_c := 8Q_c^2 \epsilon_B + 4Q_c^2 H\kappa, \qquad \widetilde{C}_{\lambda,H} := \gamma(1 - \lambda) \sum_{j=0}^{H-2} (\gamma\lambda)^j. \tag{C.29}$$

The sum is empty for $H = 1$. For the actor and critic conditional biases

$$b_t := \mathbb{E}[g_t \mid \mathcal{F}_t] - \nabla J(\theta_t), \qquad u_t := \mathbb{E}[h_t \mid \mathcal{F}_t] - \nabla_w \mathcal{L}_c(\theta_t, w_t),$$

one has

$$\mathbb{E}\|b_t\|^2 \le a\,\mathbb{E}[\Delta_t] + F_{B,\kappa,\delta}, \tag{C.30}$$

$$\mathbb{E}\|u_t\|^2 \le E_c. \tag{C.31}$$

If each direction averages m_a or m_c independent uniform-with-replacement samples, respectively, valid conditional variance bounds are

$$\mathbb{E}[\|g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]\|^2 \mid \mathcal{F}_t] \le \frac{R_{\max}^2 G_\pi^2 A_{\max}^2}{m_a},$$

$$\mathbb{E}[\|h_t - \mathbb{E}[h_t \mid \mathcal{F}_t]\|^2 \mid \mathcal{F}_t] \le \frac{Q_c^2}{m_c}. \tag{C.32}$$

A sample may be a whole trajectory with its token-averaged contribution, or a uniformly chosen trajectory–token pair; both give the same conditional means.

Proof. We first bound the discrepancy between the empirical and population trajectory laws, then apply this bound to the actor and critic integrands.

Step 1: a uniform empirical-law comparison. Condition on $\mathcal{G}_n$, and let p_i and $\widehat{p}_i$ be the population and empirical probabilities of the S_{traj} paths. With $D_n := \mathrm{TV}(\widehat{P}_n, P_{\bar{\theta}})$, Cauchy–Schwarz and the variance of each empirical frequency imply

$$\mathbb{E}[D_n^2 \mid \mathcal{G}_n] \le \frac{S_{\mathrm{traj}}}{4} \sum_i \mathbb{E}[(\widehat{p}_i - p_i)^2 \mid \mathcal{G}_n] = \frac{S_{\mathrm{traj}}}{4B} \left(1 - \sum_i p_i^2\right).$$

Since $D_n \le 1$, this proves $\mathbb{E}[D_n^2] \le \epsilon_B$. For every vector-valued integrand f satisfying $\sup_\tau \|f(\tau)\| \le L$,

$$\|(\widehat{P}_n - P_{\bar{\theta}})f\| \le 2LD_n \quad \text{pathwise.} \tag{C.33}$$

This inequality holds simultaneously for all bounded integrands. In particular, f may depend on the realized batch through the current actor and critic. We use the expectation bound on D_n^2 only after this pathwise comparison; conditioning on the entire batch does not make its empirical error vanish.

Step 2: compare whole trajectory laws. The shared prompt and environment kernels cancel in the likelihood ratio. The KL chain rule and equation C.28 give

$$\mathrm{KL}(P_{\bar{\theta}} \| P_{\theta_t}) = \sum_h \mathbb{E}_{P_{\bar{\theta}}} \mathrm{KL}(\pi_{\bar{\theta}}(\cdot \mid s_h) \| \pi_{\theta_t}(\cdot \mid s_h)) \le H\kappa, \qquad \mathrm{TV}(P_{\bar{\theta}}, P_{\theta_t})^2 \le H\kappa/2. \tag{C.34}$$

These are pointwise comparisons at the realized parameters, so their validity is unaffected by how those parameters were selected from the batch.

Step 3: identify the empirical actor direction. Suppress t and write $z_h := \nabla_\theta \log \pi_\theta(a_h \mid s_h)$ and $r_h := \pi_\theta(a_h \mid s_h)/\pi_{\bar{\theta}}(a_h \mid s_h)$. Let α_h be the selected sample derivative multiplier from Lemma C.3; in particular, $0 \le 1 - \alpha_h \le \mathbf{1}_{\{|r_h - 1| \ge \delta\}}$. Define

$$\Phi^{\mathrm{clip}}(\tau) := \frac{1}{H} \sum_h \gamma^h \alpha_h r_h z_h \widehat{A}_h(w, \tau),$$
$$\Phi(\tau) := \frac{1}{H} \sum_h \gamma^h r_h z_h \widehat{A}_h(w, \tau), \qquad \Psi(\tau) := \frac{1}{H} \sum_h \gamma^h z_h \widehat{A}_h(w, \tau).$$

The random advantage and its sign stay inside α_h. Uniform sampling from the stored batch gives exactly $\mathbb{E}[g_t \mid \mathcal{F}_t] = \widehat{P}_n \Phi^{\mathrm{clip}}$. Introduce four differences:

$$b_t = \underbrace{(\widehat{P}_n - P_{\bar{\theta}})\Phi^{\mathrm{clip}}}_{b^{\mathrm{batch}}} + \underbrace{P_{\bar{\theta}}(\Phi^{\mathrm{clip}} - \Phi)}_{b^{\mathrm{clip}}}$$
$$+ \underbrace{(P_{\bar{\theta}}\Phi - P_\theta \Psi)}_{b^{\mathrm{traj}}} + \underbrace{(P_\theta \Psi - \nabla J(\theta))}_{b^{\mathrm{adv}}}. \tag{C.35}$$

Since $\|\Phi^{\mathrm{clip}}\| \le R_{\max} G_\pi A_{\max}$, Step 1 gives

$$\|b^{\mathrm{batch}}\|^2 \le 4R_{\max}^2 G_\pi^2 A_{\max}^2 D_n^2. \tag{C.36}$$

Step 4: population clipping distortion. For any fixed state and timestep, put $p = \pi_{\bar{\theta}}(\cdot \mid s)$ and $q = \pi_\theta(\cdot \mid s)$. The boundary-inclusive clip-active event is contained in $\{|r - 1| \ge \delta\}$. Thus, even when the label is random,

$$\mathbb{E}_p[r(1 - \alpha)] \le \mathbb{E}_p[r \mathbf{1}_{\{|r-1| \ge \delta\}}] \le \Pr_p(|r - 1| \ge \delta) + \mathbb{E}_p|r - 1| \le \sqrt{2\mathrm{KL}(p\|q)}(1 + \delta^{-1}),$$

where the last inequality follows from Markov and Pinsker as in Lemma C.4. Averaging over states and timesteps, applying Jensen, and using $\gamma^h \le 1$ yields

$$\|b^{\mathrm{clip}}\|^2 \le 2G_\pi^2 A_{\max}^2 \kappa (1 + \delta^{-1})^2. \tag{C.37}$$

No likelihood-ratio envelope is needed for this population comparison.

Step 5: trajectory and action-ratio drift. Decompose

$$b^{\mathrm{traj}} = P_{\bar{\theta}}(\Phi - \Psi) + (P_{\bar{\theta}} - P_\theta)\Psi.$$

The first term is bounded by

$$\|P_{\bar{\theta}}(\Phi - \Psi)\| \le \frac{G_\pi A_{\max}}{H} \sum_h \mathbb{E}_{P_{\bar{\theta}}}|r_h - 1| \le G_\pi A_{\max}\sqrt{2\kappa}.$$

Here $\mathbb{E}_{a\sim p}|q/p - 1| = 2\mathrm{TV}(p, q)$ and Jensen give the final inequality. Since $\|\Psi\| \le G_\pi A_{\max}$, Step 2 bounds the second term by $G_\pi A_{\max}\sqrt{2H\kappa}$. Therefore

$$\|b^{\mathrm{traj}}\|^2 \le 2G_\pi^2 A_{\max}^2 \kappa(1 + \sqrt{H})^2. \tag{C.38}$$

The full-trajectory comparison changes both the state visitation and the future continuation law used in GAE. It does not identify action-only importance weighting with on-policy trajectories.

Step 6: current-policy GAE error and the combined actor bound. Under P_θ, the realized current critic is fixed before this population expectation is evaluated. Corollary C.2 therefore applies. The policy-gradient identity with the stated return normalization is

$$\nabla J(\theta) = \frac{1}{H}\mathbb{E}_{P_\theta} \sum_h \gamma^h z_h A^{\pi_\theta}(s_h, a_h).$$

The score identity cancels the state-only baseline error before applying the norm bound. With value realization, this gives

$$\|b^{\mathrm{adv}}\|^2 \le G_\pi^2 L_V^2 \widetilde{C}_{\lambda,H}^2 \Delta_t. \tag{C.39}$$

Apply $\|\sum_{i=1}^4 v_i\|^2 \le 4\sum_{i=1}^4 \|v_i\|^2$ to equation C.35, use equation C.36–equation C.39, and take total expectation. The resulting constants are exactly equation C.29, proving equation C.30.

Step 7: the stored-target critic bias. The trajectory gradient integrand is

$$\Xi_w(\tau) := \frac{1}{H} \sum_h 2(V_w(s_h) - G_h(\tau))\nabla_w V_w(s_h), \qquad \|\Xi_w(\tau)\| \le Q_c.$$

Differentiation of the finite expectation in equation C.27 gives $P_\theta \Xi_w = \nabla_w \mathcal{L}_c(\theta, w)$, while uniform minibatch sampling gives $\mathbb{E}[h_t \mid \mathcal{F}_t] = \widehat{P}_n \Xi_{w_t}$. Consequently

$$u_t = (\widehat{P}_n - P_{\bar{\theta}})\Xi_{w_t} + (P_{\bar{\theta}} - P_{\theta_t})\Xi_{w_t}.$$

Steps 1–2 and the two-term squared triangle inequality imply

$$\|u_t\|^2 \le 8Q_c^2 D_n^2 + 8Q_c^2 \mathrm{TV}(P_{\bar{\theta}}, P_{\theta_t})^2 \le 8Q_c^2 D_n^2 + 4Q_c^2 H\kappa.$$

Taking total expectation proves equation C.31. This argument permits adaptive fitting of w_t to the same stored trajectories because the comparison in equation C.33 is uniform over bounded integrands.

Step 8: minibatch variances. Conditional on $\mathcal{F}_t$, a single actor contribution has norm at most $R_{\max} G_\pi A_{\max}$, and a single critic contribution has norm at most Q_c. Independence of the draws within each with-replacement minibatch and the variance identity give equation C.32. These are bounds on centered conditional variances; total second moments also contain the squared conditional means. The same bounds hold when the actor and critic share samples. □

Scope of the comparison. The finite-support constant may be large, and increasing the number of inner updates does not eliminate the S_{traj}/B statistical floor. The lemma covers raw GAE recomputed with the current pre-update critic and stored Monte Carlo critic targets. Further advantage normalization, critic-dependent bootstrapped targets, or an empirical KL acceptance rule require additional arguments. The expectation bounds above hold at deterministic outer–inner indices; selecting among batches based on their realized data requires control of that selection as well.

C.5 A STRUCTURED BRIDGE WITHOUT TRAJECTORY COUNTING

The finite-support comparison above controls every bounded function of a complete trajectory. For a tabular critic, the actual clipped actor gradient has a smaller class of trajectory-dependent scalar factors. Controlling that class gives the following alternative, including when rewards have continuous distributions.

Lemma C.6 (Uniform empirical gradients for a bounded tabular critic)**.** *Consider a finite layered Markov decision process with disjoint nonterminal state sets $\mathcal{S}_0, \dots, \mathcal{S}_{H-1}$ and a finite action set $\mathcal{A}$. Write*

$$d_V := \sum_{h=0}^{H-1} |\mathcal{S}_h|, \qquad M := \sum_{h=0}^{H-1} |\mathcal{S}_h||\mathcal{A}|.$$

The critic is tabular: $V_w(s) = w_s$ at nonterminal states, with $w \in \mathcal{W} := [0, W]^{d_V}$, and it has a fixed prescribed terminal boundary. Conditionally on the history $\mathcal{G}_n$ before collection, collect $B \geq 1$ independent complete trajectories under a $\mathcal{G}_n$-measurable behavior policy $\pi_{\bar{\theta}}$. Let $P = P_{\bar{\theta}}$ and $\widehat{P} = \widehat{P}_n$. The state, reward, and initial-state kernels are policy independent; the reward distribution need not have finite support. Policies have common positive action support and are continuously differentiable in a finite-dimensional parameter. All trajectory functions and parameter selections are measurable.

Use the stored targets G_h, raw GAE $\widehat{A}_h(w, \tau)$, and γ^h/H actor normalization from Lemma C.5. Suppose $|G_h| \leq G_{\max}$ and $|\widehat{A}_h(w, \tau)| \leq A_{\max}$ for every $w \in \mathcal{W}$ and supported trajectory, where $A_{\max} > 0$. All admissible actor parameters satisfy the score and likelihood-ratio envelopes G_π and $R_{\max}$ of that lemma. Fix $\delta \in (0, 1)$ and select the unclipped derivative at both clipping boundaries. *Define*

$$C := C_{\lambda,H} = 1 + \gamma(1-\lambda) \sum_{j=0}^{H-2} (\gamma\lambda)^j \leq 2, \qquad \ell_B := \log(6M) + d_V \log\left(2 + \frac{CWB}{A_{\max}}\right),$$

where the sum is empty for $H = 1$. For the actual clipped trajectory gradient $\Phi^{\mathrm{clip}}_{\theta,w}$ from the preceding proof and the stored-target critic gradient

$$\Xi_w(\tau) := \frac{2}{H} \sum_{h=0}^{H-1} (w_{s_h} - G_h(\tau)) e_{s_h},$$

where e_s is the tabular coordinate vector, there exist nonnegative, random variables $D_{a,n}, D_{c,n}$ measurable with respect to $\mathcal{G}_n$ and the collected batch such that, simultaneously for every admissible actor parameter θ and $w \in \mathcal{W}$,

$$\|(\widehat{P} - P)\Phi^{\mathrm{clip}}_{\theta,w}\| \leq D_{a,n}, \qquad \|(\widehat{P} - P)\Xi_w\| \leq D_{c,n}.$$

Their conditional second moments obey

$$\mathbb{E}[D^2_{a,n} \mid \mathcal{G}_n] \leq e_a(B), \tag{C.40}$$

$$\mathbb{E}[D^2_{c,n} \mid \mathcal{G}_n] \leq e_c(B), \tag{C.41}$$

where the following explicit constants are valid:

$$e_a(B) := \min\left\{4R^2_{\max}G^2_\pi A^2_{\max}, \frac{4R^2_{\max}G^2_\pi A^2_{\max}M^2}{H^2B}(\ell_B + 3)\right\},$$

$$e_c(B) := \frac{8(W^2 + G^2_{\max})}{HB}. \tag{C.42}$$

In particular these bounds hold when (θ, w) is selected adaptively after arbitrarily many preceding uses of the batch, at each deterministic outer–inner index. No covering number for the actor parameters is required.

Proof. All probabilities in this proof are conditional on $\mathcal{G}_n$, so the behavior law is fixed before the independent trajectories are collected.

Step 1: the critic dependence of raw GAE. For two critics $w, v \in \mathcal{W}$, the reward and fixed terminal terms cancel. Writing $q = \gamma\lambda$, direct collection of the coefficients gives

$$\widehat{A}_h(w,\tau) - \widehat{A}_h(v,\tau) = -(w_{s_h} - v_{s_h}) + \gamma(1-\lambda) \sum_{j=1}^{H-1-h} q^{j-1}(w_{s_{h+j}} - v_{s_{h+j}}). \tag{C.43}$$

Thus raw GAE is affine in w and $|\widehat{A}_h(w,\tau) - \widehat{A}_h(v,\tau)| \le C\|w - v\|_\infty$ uniformly over trajectories and timesteps. The maps

$$T_0(x) = x, \qquad T_+(x) = \max\{x, 0\}, \qquad T_-(x) = \min\{x, 0\}$$

are all 1-Lipschitz and have magnitude at most $A_{\max}$ on $[-A_{\max}, A_{\max}]$.

Step 2: the exact clipped gradient in each cell. For a cell $j = (h, s, a)$, let $I_j(\tau) = \mathbf{1}_{\{s_h = s, a_h = a\}}$, $r_j = \pi_\theta(a \mid s)/\pi_{\bar\theta}(a \mid s)$, and $z_j = \nabla_\theta \log \pi_\theta(a \mid s)$. With the stated boundary rule, differentiating the PPO surrogate while holding the advantage label fixed gives the scalar multiplier

$$\alpha_j \widehat{A}_h(w,\tau) = \begin{cases} T_+(\widehat{A}_h(w,\tau)), & r_j < 1-\delta, \\ T_0(\widehat{A}_h(w,\tau)), & 1-\delta \le r_j \le 1+\delta, \\ T_-(\widehat{A}_h(w,\tau)), & r_j > 1+\delta. \end{cases} \tag{C.44}$$

Here T_- is the signed negative part, so no additional minus sign is needed. For fixed $(\theta, \bar\theta)$, the choice $u_j \in \{0, +, -\}$ depends only on the cell. In particular, although the label sign depends on the rest of the trajectory and on w, all that dependence stays inside T_{u_j}. Consequently

$$\Phi^{\mathrm{clip}}_{\theta,w}(\tau) = \frac{1}{H} \sum_{j=(h,s,a)} \gamma^h r_j z_j I_j(\tau) T_{u_j}(\widehat{A}_h(w,\tau)).$$

Let $f_{j,u,w}(\tau) := I_j(\tau) T_u(\widehat{A}_h(w,\tau))$. The vector coefficient $\gamma^h r_j z_j$ is constant on its cell, with norm at most $R_{\max} G_\pi$. Therefore, pathwise and simultaneously for all admissible actor and critic parameters,

$$\|(\widehat{P} - P)\Phi^{\mathrm{clip}}_{\theta,w}\| \le \frac{R_{\max} G_\pi M}{H} \max_{j,u} \sup_{w \in \mathcal{W}} |(\widehat{P} - P) f_{j,u,w}|. \tag{C.45}$$

The bound requires no continuity of the clipped gradient in θ. Because it holds uniformly, it also covers parameters selected using the collected data.

Step 3: a finite critic grid and its expected squared maximum. Set $\varepsilon = A_{\max}/(CB)$. A product grid in $[0, W]^{d_V}$ with mesh at most ε is an ℓ_∞ cover with cardinality

$$N \le (2 + W/\varepsilon)^{d_V} = \left(2 + \frac{CWB}{A_{\max}}\right)^{d_V}.$$

Indeed one may use the multiples of ε in each coordinate and add the endpoint W; if $W = 0$, the grid is the singleton cube. For each of the at most $3MN$ grid functions, $f \in [-A_{\max}, A_{\max}]$. Hoeffding's inequality for the B independent trajectories and a union bound give, for their maximal absolute empirical deviation Z,

$$\Pr(Z \ge x \mid \mathcal{G}_n) \le \min\left\{1, 6MN \exp\left(-\frac{Bx^2}{2A^2_{\max}}\right)\right\}.$$

Integrating the tail of Z^2, split at $2A^2_{\max}\log(6MN)/B$, yields

$$\mathbb{E}[Z^2 \mid \mathcal{G}_n] \le \frac{2A^2_{\max}}{B}\big(\log(6MN) + 1\big) \le \frac{2A^2_{\max}}{B}(\ell_B + 1). \tag{C.46}$$

Step 1 implies that replacing any w by a grid point within ε changes $f_{j,u,w}$ by at most $C\varepsilon$ on each trajectory. Hence it changes its empirical-minus-population mean by at most $2C\varepsilon = 2A_{\max}/B$. It follows that

$$\begin{aligned} \mathbb{E}\left[\left(\max_{j,u} \sup_{w \in \mathcal{W}} |(\widehat{P} - P) f_{j,u,w}|\right)^2 \,\middle|\, \mathcal{G}_n\right] &\le 2\mathbb{E}[Z^2 \mid \mathcal{G}_n] + \frac{8A^2_{\max}}{B^2} \\ &\le \frac{4A^2_{\max}}{B}(\ell_B + 3), \end{aligned}$$

where $B \geq 1$ is used in the last inequality. Together with equation C.45 this proves the second entry in the minimum defining e_a. The first entry follows from $\|\Phi^{\mathrm{clip}}_{\theta,w}\| \leq R_{\max} G_\pi A_{\max}$ and the triangle inequality. Explicitly, a batch-measurable upper bound is

$$D_{a,n} := \min\left\{2R_{\max}G_\pi A_{\max}, \frac{R_{\max}G_\pi M}{H}\left(Z + \frac{2A_{\max}}{B}\right)\right\}.$$

Thus no measurability assumption on an uncountable actor supremum is needed.

Step 4: the sparse tabular critic gradient. Index nonterminal states by $s \in \bigcup_h \mathcal{S}_h$, and let $h(s)$ be the unique layer of s. Define

$$I_s(\tau) = \mathbf{1}_{\{s_{h(s)}=s\}}, \quad X_s = (\widehat{P} - P)I_s, \quad Y_s = (\widehat{P} - P)(I_s G_{h(s)}).$$

Every path visits exactly one state per layer, so $\sum_s P I_s = H$. The sth coordinate of $(\widehat{P} - P)\Xi_w$ is $2(w_s X_s - Y_s)/H$. Thus

$$\sup_{w \in \mathcal{W}} \|(\widehat{P} - P)\Xi_w\|^2 \leq \frac{8}{H^2}\sum_s (W^2 X_s^2 + Y_s^2).$$

Define $D_{c,n}$ as the square root of this last right-hand side. The trajectories are independent, and hence

$$\sum_s \mathbb{E}[X_s^2 \mid \mathcal{G}_n] = \frac{1}{B}\sum_s \mathrm{Var}_P(I_s) \leq \frac{H}{B},$$
$$\sum_s \mathbb{E}[Y_s^2 \mid \mathcal{G}_n] = \frac{1}{B}\sum_s \mathrm{Var}_P(I_s G_{h(s)}) \leq \frac{1}{B}\sum_s P(I_s G_{h(s)}^2) \leq \frac{H G_{\max}^2}{B}.$$

These inequalities prove equation C.41. Dependence between states or rewards along a trajectory is allowed; only independence between the B complete trajectories was used. □

Corollary C.7 (Structured replacement in the coupled bounds)**.** *Use the structured setting and clipping-boundary rule of Lemma C.6, and retain the population-KL, realizability, and other population comparison assumptions of Lemma C.5. A finite trajectory support is not required. Then the actor and critic bias conclusions of that lemma hold with the same coefficient a, and with*

$$F^{\mathrm{str}}_{B,\kappa,\delta} := 4e_a(B) + 8G_\pi^2 A_{\max}^2 \kappa\big[(1+\sqrt{H})^2 + (1+\delta^{-1})^2\big],$$
$$E^{\mathrm{str}}_c := 2e_c(B) + 4Q_c^2 H\kappa, \tag{C.47}$$

where Q_c may be any uniform bound on $\|\Xi_w(\tau)\|$. For example, layered tabular structure permits

$$Q_c = \frac{2(W + G_{\max})}{\sqrt{H}}; \qquad Q_c = \frac{2\max\{W, G_{\max}\}}{\sqrt{H}} \quad \text{if } 0 \leq G_h \leq G_{\max}.$$

The synchronous coupled theorem therefore applies with $F = F^{\mathrm{str}}$ and $E_c = E^{\mathrm{str}}_c$, whenever its remaining smoothness, critic contraction, and stepsize premises hold. For complete-trajectory minibatches the same Q_c also gives the critic variance bound Q_c^2/m_c; for token minibatches that variance instead uses a bound on the individual token gradient.

Proof. In equation C.35, replace only the expected squared empirical actor term by $e_a(B)$ from equation C.40. The other three population comparisons are unchanged and use bounded integrands and the KL chain rule, not the cardinality of the trajectory support. Their four-term squared triangle inequality contributes $4e_a(B)$. Continuous rewards cause no differentiation problem: conditional on each finite state–action sequence, their law is policy independent. The expectations defining J and $\mathcal{L}_c$ are therefore finite sums of differentiable policy path probabilities times bounded reward integrals. Their derivatives are the same policy-gradient and critic-gradient expressions used above. For the critic, the same decomposition gives

$$u_t = (\widehat{P} - P_{\bar{\theta}})\Xi_{w_t} + (P_{\bar{\theta}} - P_{\theta_t})\Xi_{w_t}.$$

The expected square of the first term is at most $e_c(B)$, and the square of the second is at most $4Q_c^2\mathrm{TV}(P_{\bar{\theta}}, P_{\theta_t})^2 \le 2Q_c^2 H\kappa$. The two-term squared triangle inequality proves the stated E_c^{str}. Each path visits H distinct tabular coordinates, so

$$\|\Xi_w(\tau)\|^2 = \frac{4}{H^2}\sum_{h=0}^{H-1}(w_{s_h} - G_h)^2.$$

Bounding each residual by $W + G_{\max}$, or by $\max\{W, G_{\max}\}$ for nonnegative targets, proves the displayed choices of Q_c and the corresponding complete-trajectory minibatch variance. This proof applies at deterministic outer–inner indices, as does the batch-collection conditioning above. □

For example, if $M, d_V = O(H)$, $W, G_{\max}, A_{\max} = O(H)$, and the score and ratio envelopes are bounded independently of H, then the uncapped actor bound is $e_a(B) = O(H^3 \log(B+H)/B)$ and $e_c(B) = O(H/B)$. These bounds depend on the dimensions of the specified gradient classes rather than the number of complete action sequences, with all complexity terms explicit.

D Open-Loop Convergence Results

This appendix establishes *open-loop* finite-time guarantees for the actor updates in PPO-Clip. An explicit *bias decomposition* separates four effects that enter the stationarity analysis: (i) critic-induced advantage error (controlled by critic tracking), (ii) within-epoch trajectory drift (controlled by the population KL budget), (iii) clipping-induced distortion, and (iv) the finite-batch error that persists under adaptive reuse. These bounds are "open-loop" because they treat critic tracking error as an input. The closed-loop analysis combines the actor and critic inequalities to control this error.

D.1 Open-Loop Convergence Results for the Actor

We first present a stationarity bound for the underlying smooth RL objective $J(\theta)$ that makes the N–K update structure explicit. The theorem is stated at the level of inner iterates $\{\theta_n^{(k)}\}$ and decomposes the deviation between the expected PPO update direction and the true policy gradient into critic, drift, clipping, and finite-batch terms.

Theorem D.1 (Non-asymptotic stationarity of the actor (open-loop, explicit N–K structure))**.** *Consider PPO-Clip with a fixed initial actor θ_0, deterministic outer iterations $n = 0, \dots, N-1$, and K actual inner actor updates per outer iteration, with $\theta_{n+1} = \theta_n^{(K)}$ and $\theta_n^{(0)} = \theta_n$. Let the inner iterates satisfy*

$$\theta_n^{(k+1)} = \theta_n^{(k)} + \eta\, g_n^{(k)}, \qquad k = 0, \dots, K-1,$$

with stepsize $0 < \eta \le 1/(4L_J)$. Assume Assumptions 1 and 2 and the finite-batch sampling, coverage, raw-GAE, and population-KL premises of Lemma C.5. The objective is $J(\theta) = H^{-1}\mathbb{E}_{P_\theta} G_0$, and the actor uses the γ^h-weighted, token-averaged sample direction defined in that lemma. Define the (conditional) actor bias at inner step (n, k) by

$$b_{n,k} := \mathbb{E}[g_n^{(k)} \mid \mathcal{F}_{n,k}] - \nabla J(\theta_n^{(k)}).$$

Then

$$\frac{1}{NK}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla J(\theta_n^{(k)})\|^2\big] \le \underbrace{\frac{4\big(J_{\max} - J(\theta_0)\big)}{\eta\, NK}}_{\text{optimization}} + \underbrace{2L_J\,\eta\,\sigma_a^2}_{\text{stochastic noise}} + \underbrace{\frac{4}{NK}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\|b_{n,k}\|^2}_{\text{systematic bias}}. \tag{D.1}$$

Moreover, for each inner step (n, k), the bias satisfies

$$\mathbb{E}\|b_{n,k}\|^2 \le a\,\mathbb{E}[\Delta_{n,k}] + F_{B,\kappa,\delta}, \tag{D.2}$$

where $\Delta_{n,k} := \|w_n^{(k)} - w^\star(\theta_n^{(k)})\|^2$ and $a, F_{B,\kappa,\delta}$ are the explicit constants in equation C.29. Advantages are recomputed from the stored trajectories using the current pre-update critic $w_n^{(k)}$, as required by that lemma.

Interpretation. Theorem D.1 averages stationarity over all *inner* iterates $\{\theta_n^{(k)}\}$, i.e., over the total number of actor gradient steps NK. The bound has the standard nonconvex SGD form: an $O(1/(\eta NK))$ optimization term, an $O(\eta)$ noise term, and an additive bias term. The PPO-specific effects appear in equation D.2: critic-induced advantage error enters through the expected tracking error and the GAE coefficient $\widetilde{C}_{\lambda,H}$; trajectory drift and clipping are controlled by κ and δ; and adaptive reuse contributes an explicit finite-batch term through $\epsilon_B = \min\{1, S_{\mathrm{traj}}/(4B)\}$. The bound is *open-loop* because critic tracking remains to be controlled by its own update recursion. The finite-batch term is not removed by taking more inner updates.

Proof of Theorem D.1. We prove equation D.1 first (smoothness ⇒ stationarity), and then equation D.2 (PPO-specific bias decomposition).

Part I: from smoothness to the open-loop stationarity bound. Let $T := NK$ and flatten the double index (n, k) into a single time index

$$t := nK + k \in \{0, 1, \dots, T-1\}.$$

Define the single-index iterates and quantities

$$\theta_t := \theta_n^{(k)}, \qquad g_t := g_n^{(k)}, \qquad b_t := b_{n,k}, \qquad \mathcal{F}_t := \mathcal{F}_{n,k}.$$

Then the actor recursion becomes

$$\theta_{t+1} = \theta_t + \eta g_t, \qquad t = 0, \dots, T-1,$$

and the averaged stationarity measure satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla J(\theta_t)\|^2\big] = \frac{1}{NK}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1} \mathbb{E}\big[\|\nabla J(\theta_n^{(k)})\|^2\big].$$

By L_J-smoothness of J (Assumption 1), for any x, y,

$$J(y) \;\geq\; J(x) + \langle \nabla J(x), y - x\rangle - \frac{L_J}{2}\|y - x\|^2. \tag{D.3}$$

Apply equation D.3 with $x = \theta_t$ and $y = \theta_{t+1} = \theta_t + \eta g_t$:

$$J(\theta_{t+1}) \;\geq\; J(\theta_t) + \eta\langle \nabla J(\theta_t), g_t\rangle - \frac{L_J\eta^2}{2}\|g_t\|^2. \tag{D.4}$$

Taking conditional expectation given $\mathcal{F}_t$ and using that θ_t is $\mathcal{F}_t$-measurable yields

$$\mathbb{E}[J(\theta_{t+1}) \mid \mathcal{F}_t] \;\geq\; J(\theta_t) + \eta\,\langle \nabla J(\theta_t), \mathbb{E}[g_t \mid \mathcal{F}_t]\rangle - \frac{L_J\eta^2}{2}\,\mathbb{E}\big[\|g_t\|^2 \mid \mathcal{F}_t\big]. \tag{D.5}$$

By definition of the (conditional) bias,

$$b_t := \mathbb{E}[g_t \mid \mathcal{F}_t] - \nabla J(\theta_t) \qquad \Longleftrightarrow \qquad \mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla J(\theta_t) + b_t,$$

so

$$\langle \nabla J(\theta_t), \mathbb{E}[g_t \mid \mathcal{F}_t]\rangle = \|\nabla J(\theta_t)\|^2 + \langle \nabla J(\theta_t), b_t\rangle.$$

Using $2\langle u, v\rangle \geq -\|u\|^2 - \|v\|^2$ with $u = \nabla J(\theta_t)$ and $v = b_t$, we get

$$\langle \nabla J(\theta_t), b_t\rangle \geq -\frac{1}{2}\|\nabla J(\theta_t)\|^2 - \frac{1}{2}\|b_t\|^2.$$

The conditional variance assumption gives

$$\mathbb{E}\big[\|g_t\|^2 \mid \mathcal{F}_t\big] = \mathbb{E}\big[\|g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]\|^2 \mid \mathcal{F}_t\big] + \|\nabla J(\theta_t) + b_t\|^2 \leq \sigma_a^2 + 2\|\nabla J(\theta_t)\|^2 + 2\|b_t\|^2.$$

Substituting this bound into equation D.5 yields

$$\mathbb{E}[J(\theta_{t+1}) \mid \mathcal{F}_t] \;\geq\; J(\theta_t) + \left(\frac{\eta}{2} - L_J\eta^2\right)\|\nabla J(\theta_t)\|^2 - \left(\frac{\eta}{2} + L_J\eta^2\right)\|b_t\|^2 - \frac{L_J\eta^2}{2}\sigma_a^2. \tag{D.6}$$

Since $\eta \leq 1/(4L_J)$, we have $\frac{\eta}{2} - L_J\eta^2 \geq \frac{\eta}{4}$ and $\frac{\eta}{2} + L_J\eta^2 \leq \eta$. Rearranging equation D.6 gives

$$\frac{\eta}{4}\|\nabla J(\theta_t)\|^2 \;\leq\; \mathbb{E}[J(\theta_{t+1}) \mid \mathcal{F}_t] - J(\theta_t) + \eta\|b_t\|^2 + \frac{L_J\eta^2}{2}\sigma_a^2. \tag{D.7}$$

Taking total expectation in equation D.7 and summing over $t = 0, \dots, T-1$ yields the telescoping bound

$$\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\theta_t)\|^2 \;\leq\; \mathbb{E}[J(\theta_T)] - J(\theta_0) + \eta\sum_{t=0}^{T-1}\mathbb{E}\|b_t\|^2 + \frac{L_J\eta^2}{2}T\sigma_a^2. \tag{D.8}$$

By the bounded-above assumption (Assumption 1), $\mathbb{E}[J(\theta_T)] \leq J_{\max}$. Plugging this into equation D.8 gives

$$\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\theta_t)\|^2 \leq \big(J_{\max} - J(\theta_0)\big) + \eta\sum_{t=0}^{T-1}\mathbb{E}\|b_t\|^2 + \frac{L_J\eta^2}{2}\,T\sigma_a^2.$$

Divide both sides by $\eta T/4$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\theta_t)\|^2 \leq \frac{4(J_{\max} - J(\theta_0))}{\eta T} + 2L_J\eta\sigma_a^2 + \frac{4}{T}\sum_{t=0}^{T-1}\mathbb{E}\|b_t\|^2.$$

Finally, substitute $T = NK$ and restore the double index $t \leftrightarrow (n, k)$ to obtain equation D.1.

Part II: decomposition and control of the actor bias. Fix an inner step (n, k) and condition on its pre-update history $\mathcal{F}_{n,k}$, which contains the stored rollout batch and the current actor and critic. Lemma C.5 identifies the conditional minibatch mean with the empirical clipped direction, not with a population gradient. It decomposes its bias into four terms: empirical-to-population error, population clipping distortion, whole-trajectory policy drift, and current-policy GAE error. The empirical total-variation bound holds uniformly over bounded integrands, so it continues to hold when both parameters have been fitted adaptively to the same batch. The trajectory comparison controls old-policy continuation errors in GAE in addition to state-visitation drift. The random advantage label and its sign remain inside the clipping selector. Applying equation C.30 at this deterministic index gives

$$\mathbb{E}\|b_{n,k}\|^2 \le 4G_\pi^2 L_V^2 \widetilde{C}_{\lambda,H}^2 \, \mathbb{E}[\Delta_{n,k}] + F_{B,\kappa,\delta},$$

which gives equation D.2. This completes the proof. □

D.2 Open-Loop Convergence for the Critic

Theorem D.2 (Non-asymptotic tracking of the critic (open-loop, explicit N–K structure))**.** *Consider $T = NK$ actual joint actor–critic updates, with $\theta_{t+1} = \theta_t + \eta g_t$ and $w_{t+1} = w_t - \beta_c h_t$. For $f_t(w) = \mathcal{L}_c(\theta_t, w)$, assume the convexity, L_c-smoothness and selected-minimizer quadratic-growth premises of Assumption 4, with modulus $\mu_q > 0$. Let $w_t^\star = w^\star(\theta_t)$ be measurable and $L_{w^\star}$-Lipschitz along the actor iterates. The pre-update history $\mathcal{F}_t$ includes the stored batch and current parameters. Define*

$$e_t^c := \mathbb{E}[h_t \mid \mathcal{F}_t] - \nabla f_t(w_t), \qquad \xi_t := h_t - \mathbb{E}[h_t \mid \mathcal{F}_t], \quad \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \le \sigma_c^2.$$

Thus e_t^c is a predictable bias, while ξ_t is conditionally centered. No independence from the actor direction is required. Set $\Delta_t = \|w_t - w_t^\star\|^2$, $\mathcal{E}_c = T^{-1}\sum_{t<T} \mathbb{E}\|e_t^c\|^2$, and suppose

$$0 < \beta_c \le \min\left\{ \frac{\mu_q}{2L_c^2}, \frac{1}{\mu_q} \right\}.$$

Then

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\Delta_t \le \underbrace{\frac{4\mathbb{E}\Delta_0}{\mu_q \beta_c T}}_{\textit{initialization}} + \underbrace{\frac{5\beta_c \sigma_c^2}{\mu_q}}_{\textit{noise}} + \underbrace{\frac{40\mathcal{E}_c}{\mu_q^2}}_{\textit{sampling/target bias}} + \underbrace{\frac{40L_{w^\star}^2 \eta^2}{\mu_q^2 \beta_c^2} \frac{1}{T}\sum_{t<T} \mathbb{E}\|g_t\|^2}_{\textit{actor-induced target drift}}. \tag{D.9}$$

For the finite-batch protocol, Lemma C.5 proves $\mathcal{E}_c \le E_c$.

Interpretation. The critic updates at each actual inner actor step and tracks the current-policy population minimizer. A minibatch from the stored rollout batch is centered on its empirical Monte Carlo loss. The discrepancy from the current-policy loss appears as the explicit bias floor $40\mathcal{E}_c/\mu_q^2$; it is not treated as martingale noise. The target-drift term scales as η^2/β_c^2.

Proof of Theorem D.2. Flatten $t = nK + k$ and write $e_t = w_t - w_t^\star$, $D_t = \mathbb{E}\|e_t\|^2$. All subsequent expectations are taken with respect to the stated empirical-data filtration.

Step 1: one-step recursion for the tracking error. Define the history-measurable frozen-policy step $u_t = e_t - \beta_c \nabla f_t(w_t)$ and the target displacement $d_t = w_t^\star - w_{t+1}^\star$. Then

$$e_{t+1} = u_t - \beta_c \xi_t + (d_t - \beta_c e_t^c). \tag{D.10}$$

The displacement d_t may depend on the current actor noise and be correlated with ξ_t.

Step 2: separate contraction, noise, and displacement. For $\alpha > 0$, Young's inequality and conditional centering give

$$\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] \le (1+\alpha)(\|u_t\|^2 + \beta_c^2\sigma_c^2) + 2(1+1/\alpha)\left(\mathbb{E}[\|d_t\|^2 \mid \mathcal{F}_t] + \beta_c^2\|e_t^c\|^2\right). \tag{D.11}$$

Conditional centering cancels the cross term between the measurable u_t and ξ_t, without requiring independence between the noise and the target displacement.

Step 3: contract the frozen-policy gradient step. Convexity, quadratic growth around the selected stationary minimizer, and smoothness imply

$$\langle \nabla f_t(w_t), e_t\rangle \ge \frac{\mu_q}{2}\|e_t\|^2, \qquad \|\nabla f_t(w_t)\| \le L_c\|e_t\|.$$

Consequently

$$\|u_t\|^2 \le (1 - \mu_q\beta_c + L_c^2\beta_c^2)\|e_t\|^2 \le (1 - \mu_q\beta_c/2)\|e_t\|^2. \tag{D.12}$$

The convex region must contain the current iterate and its selected stationary minimizer, as stated in the assumptions.

Step 4: control movement of the critic target. The minimizer Lipschitz property and the actor recursion give

$$\|d_t\|^2 \le L_{w^\star}^2\eta^2\|g_t\|^2. \tag{D.13}$$

This holds pathwise even when the two update directions share a minibatch.

Step 5: choose the contraction slack. Take $\alpha = \mu_q\beta_c/4$. Since $0 < \mu_q\beta_c \le 1$,

$$(1+\alpha)(1-\mu_q\beta_c/2) \le 1 - \mu_q\beta_c/4, \quad 1+\alpha \le 5/4, \quad 1+1/\alpha \le 5/(\mu_q\beta_c).$$

Substituting into equation D.11 and averaging yields

$$D_{t+1} \le (1-\mu_q\beta_c/4)D_t + \tfrac{5}{4}\beta_c^2\sigma_c^2 + \frac{10\beta_c}{\mu_q}\mathbb{E}\|e_t^c\|^2 + \frac{10L_{w^\star}^2\eta^2}{\mu_q\beta_c}\mathbb{E}\|g_t\|^2. \tag{D.14}$$

Step 6: telescope over actual updates. Sum equation D.14, discard the nonnegative D_T, and divide by $\mu_q\beta_c T/4$. This gives equation D.9. For reused Monte Carlo data, Lemma C.5 bounds the population discrepancy by E_c directly from the sampling protocol. □

E Closed-Loop Convergence Results

Proof of Theorem 3.1. We combine the actor and critic recursions, accounting explicitly for finite-batch error and trajectory mismatch.

Step 0: count actual updates and define averages. Let $T = NK$ and $t = nK + k$; both actor and critic execute every update counted here. Define

$$\bar{G} = T^{-1}\sum_{t<T}\mathbb{E}\|\nabla J(\theta_t)\|^2, \qquad \bar{\Delta} = T^{-1}\sum_{t<T}\mathbb{E}\Delta_t,$$
$$\bar{B} = T^{-1}\sum_{t<T}\mathbb{E}\|b_t\|^2, \qquad \bar{S} = T^{-1}\sum_{t<T}\mathbb{E}\|g_t\|^2, \tag{E.1}$$

where $b_t = \mathbb{E}[g_t \mid \mathcal{F}_t] - \nabla J(\theta_t)$ and the history includes the stored batch. The averages count only executed updates. Appendix H treats predictable stopping.

Step 1: open-loop actor progress. The smoothness argument in Theorem D.1 gives

$$\bar{G} \le A_0 + 4\bar{B}, \qquad A_0 = \frac{4(J_{\max} - J(\theta_0))}{\eta T} + 2L_J\eta\sigma_a^2. \tag{E.2}$$

Here the conditional bias uses the mean under resampling from the stored batch, which can differ from the population direction.

Step 2: control empirical, clipping, trajectory, and advantage errors. Lemma C.5 first compares the empirical clipped direction with its population counterpart. It then bounds clipping distortion with the random advantage sign inside the expectation, compares complete trajectory laws, and applies the on-policy GAE lemma. It proves

$$\bar{B} \le a\bar{\Delta} + F, \qquad a = 4G_\pi^2 L_V^2 \widetilde{C}_{\lambda,H}^2. \tag{E.3}$$

Here F is exactly equation 3.1. The factor four comes from the four-term squared-norm bound. In particular, the finite-batch contribution remains even with an exact critic, identical behavior/current policies, and inactive clipping. The trajectory comparison accounts for the behavior-policy continuation law, since a single action ratio does not make multistep GAE on-policy.

Step 3: combine actor bias and critic tracking. Combining the preceding bounds yields

$$\bar{G} \le A_0 + 4a\bar{\Delta} + 4F. \tag{E.4}$$

Step 4: open-loop critic tracking with the derived target bias. Theorem D.2 and the critic part of Lemma C.5 give

$$\bar{\Delta} \le U + d\bar{S}, \quad U = \frac{4\mathbb{E}\Delta_0}{\mu_q \beta_c T} + \frac{5\beta_c \sigma_c^2}{\mu_q} + \frac{40E_c}{\mu_q^2}, \quad d = \frac{40L_{w^\star}^2 \eta^2}{\mu_q^2 \beta_c^2}. \tag{E.5}$$

The critic uses stored Monte Carlo targets, whose batch and behavior-policy discrepancies are bounded by E_c.

Step 5: bound the update second moment. Conditional variance and $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ imply

$$\bar{S} \le \sigma_a^2 + 2\bar{G} + 2\bar{B} \le \sigma_a^2 + 2A_0 + 10a\bar{\Delta} + 10F.$$

Neither population-unbiased updates nor independent actor and critic minibatches are required.

Step 6: solve the critic inequality. Substitute into equation E.5. With $\rho = 10ad < 1$,

$$\bar{\Delta} \le \frac{U + d(\sigma_a^2 + 2A_0 + 10F)}{1-\rho}. \tag{E.6}$$

This is the critic bound in the theorem.

Step 7: conclude actor stationarity. Substitute equation E.6 into equation E.4 and collect terms:

$$\bar{G} \le \frac{(1-2ad)A_0 + 4F + 4aU + 4ad\sigma_a^2}{1-10ad} \le \frac{A_0 + 4F + 4aU + 4ad\sigma_a^2}{1-\rho}.$$

The second inequality discards a nonpositive term because $A_0 \ge 0$. This proves both bounds with the stated constants. □

F Rates and Rollout-Budget Consequences

Proof of Corollary 3.2. Apply Corollary C.7 to the actual conditional means. The resulting bounds are $\bar{B} \le a\bar{\Delta} + F_{\mathrm{str}}$ and $T^{-1}\sum_{t<T}\mathbb{E}\|u_t\|^2 \le E_{c,\mathrm{str}}$. The proof in Appendix E otherwise uses only these two inequalities, bounded conditional variances, and the stated smoothness and critic geometry. Substituting the structured residuals gives both claims. Finite complete-path support is needed only for the empirical-law comparison in the finite-support bridge. Bounded rewards, finite actions, and the tabular critic suffice for the population identities and KL comparison in the structured bridge. □

Proof of Corollary 3.3. For fixed problem constants and $1-\rho \ge 1-\rho_0 > 0$, the two bounds are at most constant multiples of

$$\frac{1}{\eta T} + \eta + \frac{1}{\beta_c T} + \beta_c + \left(\frac{\eta}{\beta_c}\right)^2 + v(B) + \kappa, \tag{F.1}$$

where $v(B) = S_{\mathrm{traj}}/B$ for the finite-support bridge and $v(B) = \log(eB)/B$ for a fixed structured problem. Indeed $A_0 = O((\eta T)^{-1} + \eta)$, $U = O((\beta_c T)^{-1} + \beta_c + v(B) + \kappa)$, and $d = O((\eta/\beta_c)^2)$; the remaining products are bounded by the same expression when these quantities are bounded. Under the stated schedule their orders are respectively $T^{-2/5}, T^{-3/5}, T^{-3/5}, T^{-2/5}, T^{-2/5}$. The prescribed batches give $v(B_T) = O(T^{-2/5})$ in either case. Choosing c_β below the critic restriction and then choosing c_η sufficiently small enforces all step restrictions and $\rho \le \rho_0$ for every $T \ge 1$; if $a = 0$ the coupling restriction is vacuous. This establishes the two rates. A sufficiently large $T = O(\varepsilon^{-5/2})$, rounded up to a multiple of fixed K, suffices. Substitution in $Q = TB_T/K$ gives the claimed counts and the extra logarithm in the structured case. These are averages of squared errors, not bounds on $\mathbb{E}\|\nabla J\|$ at the same numerical tolerance. With m trajectories per minibatch, the actor and critic each process mT trajectory samples, or HmT time-indexed sample contributions; batch reuse reduces fresh sampling without removing this optimization cost. □

A sufficient population-KL certificate. The rate corollary requires a vanishing population budget; a small empirical KL alone does not supply it. One sufficient deterministic rule is available when $\|\nabla \log \pi_\theta(a \mid s)\| \le G_\pi$ on every line segment joining within-epoch actor parameters and $\|g_t\| \le G_g$ along every update. Then $\|\theta_n^{(k)} - \bar\theta_n\| \le K\eta G_g$ and the fundamental theorem of calculus gives, pointwise in state,

$$\mathrm{KL}(\pi_{\bar\theta_n} \| \pi_{\theta_n^{(k)}}) \le G_\pi \|\theta_n^{(k)} - \bar\theta_n\| \le G_\pi G_g K \eta.$$

Thus fixed K and the schedule above give the sufficient budget $\kappa_T = G_\pi G_g K \eta_T = O(T^{-3/5})$. The score condition on these segments is an additional sufficient certificate, not a consequence of a bound only at the iterates. The examples in Appendices I and J verify the stronger quadratic certificate $\kappa = DK^2\eta^2$ globally.

Proposition F.1 (A fixed-rollout-budget bound)**.** *Suppose the synchronous proof supplies $F \le f_B/B + f_\kappa \kappa$ and $E_c \le e_B/B + e_\kappa \kappa$, with nonnegative constants, and an actual-iterate certificate $\kappa \le DK^2\eta^2$. Fix η, β_c, m satisfying the theorem, with $\rho = 10ad < 1$, and use common envelope, variance, and initialization constants over the candidate batch sizes and epoch lengths. For an exact fresh-rollout budget $Q = NB$ and $T = NK = KQ/B$, define*

$$P = \frac{4(J_{\max} - J(\theta_0))}{\eta} + \frac{16a\mathbb{E}\Delta_0}{\mu_q \beta_c},$$

$$V = 2L_J \eta \sigma_a^2 + \frac{20a\beta_c\sigma_c^2}{\mu_q} + 4ad\sigma_a^2,$$

$$C_B = 4f_B + \frac{160ae_B}{\mu_q^2}, \qquad C_\kappa = 4f_\kappa + \frac{160ae_\kappa}{\mu_q^2}.$$

Then

$$\bar G \le \frac{PB/(KQ) + C_B/B + V + C_\kappa DK^2\eta^2}{1-\rho}. \tag{F.2}$$

For positive coefficients, the continuous minimizers of the variable terms, holding the other quantities fixed, are

$$B_{\mathrm{cont}}(K) = \sqrt{\frac{C_B KQ}{P}}, \qquad K_{\mathrm{cont}}(B) = \left(\frac{PB}{2C_\kappa D\eta^2 Q}\right)^{1/3}. \tag{F.3}$$

These optimize this upper bound only. Actual choices must satisfy the integer, coverage, sampling, and stepsize requirements of the specified protocol.

Proof. Substitute the two residual bounds in equation 3.2, collect the $1/T$, constant, $1/B$, and κ terms, and use $T = KQ/B$ and the deterministic KL certificate. This gives equation F.2 with the displayed coefficients. For fixed K, differentiating $PB/(KQ) + C_B/B$ gives the first minimizer. For fixed B, differentiating $PB/(KQ) + C_\kappa D\eta^2K^2$ gives the second. Both functions have positive second derivatives on the positive axis. On bounded continuous feasible intervals, truncate the minimizer to the interval; on a discrete feasible set, evaluate the neighboring feasible values using equation F.2. Zero coefficients give the corresponding monotone or constant boundary cases. □

For the finite-support bridge one may use $f_B = 4R_{\max}^2 G_\pi^2 A_{\max}^2 S_{\mathrm{traj}}$ and $e_B = 2Q_c^2 S_{\mathrm{traj}}$, with the drift coefficients read from equation 3.1. For the structured bridge and $1 \le B \le Q$, replace ℓ_B by ℓ_Q in Corollary 3.2. Then

$$f_B(Q) = \frac{16R_{\max}^2 G_\pi^2 A_{\max}^2 M^2}{H^2}(\ell_Q + 3), \qquad e_B = \frac{16(W^2 + G_{\max}^2)}{H}$$

give a valid budget-dependent majorant to which the same proposition applies. The bound expresses a tradeoff: larger batches reduce uniform statistical error but permit fewer updates at fixed Q, K; greater reuse supplies updates but increases the certified drift. The stationary-point guarantee does not assert that these relaxed choices maximize return or that the upper bound is tight. If only the linear certificate $\kappa \le D_1 K\eta$ is verified, replace the last term by $C_\kappa D_1 K\eta$; its fixed-B continuous minimizer is $\sqrt{PB/(C_\kappa D_1 \eta Q)}$.

G The Classical Two-Time-Scale Limit

We derive Proposition 3.4 from the finite-batch bounds.

Proof of Proposition 3.4. Fix the finite trajectory support, horizon, discount, GAE and clipping parameters, initial actor, and all regularity, coverage, and noise constants across the family of runs. The initial expected critic error is uniformly bounded, and each run executes $T = N(T)K(T)$ joint updates using batches of size $B_T \to \infty$.

Step 1: random-iterate expectations equal time averages. The output index $\widehat{t}$ is uniform on $\{0, \dots, T-1\}$ and independent of the run. Hence

$$\mathbb{E}\|\nabla J(\theta_{\widehat{t}})\|^2 = T^{-1} \sum_{t<T} \mathbb{E}\|\nabla J(\theta_t)\|^2, \tag{G.1}$$

$$\mathbb{E}\Delta_{\widehat{t}} = T^{-1} \sum_{t<T} \mathbb{E}\Delta_t. \tag{G.2}$$

Step 2: apply the finite-time bounds. Use $A_{0,T}, U_T, d_T, F_T, E_{c,T}$ from Theorem 3.1. For sufficiently large T, the diminishing positive stepsizes meet its actor and critic restrictions. The bounds are

$$\bar{G}_T \le \frac{A_{0,T} + 4F_T + 4aU_T + 4ad_T\sigma_a^2}{1 - 10ad_T}, \tag{G.3}$$

$$\bar{\Delta}_T \le \frac{U_T + d_T(\sigma_a^2 + 2A_{0,T} + 10F_T)}{1 - 10ad_T}. \tag{G.4}$$

Since S_{traj} is fixed and $B_T \to \infty$, $\epsilon_{B_T} \to 0$. Together with $\kappa_T \to 0$, this proves $F_T \to 0$ and $E_{c,T} \to 0$. Thus both the finite-batch error and the error from stored behavior-policy targets vanish in the limit.

Step 3: control the coupling denominator. The ratio $\eta_T/\beta_{c,T} \to 0$ gives $d_T \to 0$ and $\rho_T = 10ad_T \to 0$. Therefore $(1-\rho_T)^{-1} \le 2$ eventually.

Step 4: every term in the critic bound vanishes. The initialization term tends to zero because $\beta_{c,T}T \ge \beta_{c,T}N(T) \to \infty$ and $\mathbb{E}\Delta_0$ is uniformly bounded. The critic noise term vanishes with $\beta_{c,T}$; the target-bias term vanishes with $E_{c,T}$. Thus $U_T \to 0$. Also $d_T\sigma_a^2 \to 0$ and $d_T F_T \to 0$. For the remaining product,

$$d_T A_{0,T} = O\left(\frac{\eta_T}{\beta_{c,T}^2 T} + \frac{\eta_T^3}{\beta_{c,T}^2}\right) \longrightarrow 0$$

under the stated schedules. Hence equation G.4 tends to zero.

Step 5: every term in the actor bound vanishes. The conditions $\eta_T T \to \infty$ and $\eta_T \to 0$ imply $A_{0,T} \to 0$. The preceding steps give $F_T, U_T, d_T \to 0$, so equation G.3 tends to zero as well. Apply equation G.1–equation G.2 to obtain the two claimed limits.

Conclusion. The result establishes stationarity and critic tracking in expectation; it does not imply policy optimality or monotonic improvement in return. Appendix I gives a concrete nonempty family of admissible schedules. □

H Predictable Early Stopping and Population KL Control

We extend the analysis to predictable early stopping using an implementable certificate for the population KL budget. An empirical KL statistic alone does not provide this certificate. There are N epochs and at most K update slots per epoch, so $T_{\max} = NK$. Let $I_{n,k} \in \{0, 1\}$ be measurable before the current minibatch draw, be zero after an epoch has stopped, and satisfy $I_{n,0} = 1$. Execute both updates when $I_{n,k} = 1$ and leave both parameters unchanged otherwise. Write $M = \sum_{n,k} I_{n,k}$, $R = \mathbb{E}M$, and $\chi = T_{\max}/R$, so $N \le R \le NK$. At each epoch, the B trajectories are conditionally i.i.d. given the pre-collection history. Subsequent minibatches are drawn uniformly with replacement from that stored batch, which is reused at most K times.

A sufficient population certificate. Suppose the score bound holds on the segments from the behavior actor to the current actor and to every possible next actor produced by a minibatch update. For every state, the fundamental theorem of calculus gives

$$\mathrm{KL}(\pi_{\bar\theta}(\cdot \mid s)\|\pi_\theta(\cdot \mid s)) \le G_\pi\|\theta - \bar\theta\|.$$

Every empirical actor direction has norm at most $G_g = R_{\max}G_\pi A_{\max}$. Consequently the predictable gate

$$G_\pi\big(\|\theta_n^{(k)} - \bar\theta_n\| + \eta G_g\big) \le \kappa \tag{H.1}$$

ensures that both the current actor and the actor after the next update satisfy the population budget, for every possible current minibatch. Assume $\eta G_\pi G_g \le \kappa$, so the first update is allowed. One may also require that the empirical KL of the current actor is below an empirical stopping threshold, using the stored batch, before executing the next step. This additional gate is predictable and can only stop the epoch sooner. If an executed step raises the empirical KL above the threshold, the epoch stops at the next slot; the population certificate still covers the executed step. Rejecting a proposal after inspecting its fresh stochastic gradient changes the conditional update law and is not covered by this rule.

Proposition H.1 (Finite-time control under predictable stopping)**.** *Under the synchronous theorem's premises and the preceding stopping protocol, define the update-weighted quantities*

$$\bar G_R = \frac{\mathbb{E}\sum_t I_t\|\nabla J(\theta_t)\|^2}{R}, \qquad \bar\Delta_R = \frac{\mathbb{E}\sum_t I_t\Delta_t}{R}.$$

Use the synchronous theorem's definitions with T replaced by R and ϵ_B replaced by $\chi\epsilon_B$ in F and E_c only. The resulting $A_{0,R}, F_R, E_{c,R}, U_R$ satisfy

$$\bar G_R \le \frac{A_{0,R} + 4F_R + 4aU_R + 4ad\sigma_a^2}{1 - 10ad}, \tag{H.2}$$

$$\bar\Delta_R \le \frac{U_R + d(\sigma_a^2 + 2A_{0,R} + 10F_R)}{1 - 10ad}. \tag{H.3}$$

No extra factor is needed on the population drift/clipping or martingale-noise terms.

Proof. Let $Z_n = \mathrm{TV}(\widehat P_n, P_{\bar\theta_n})$. The empirical-distribution bound holds conditional on the history before batch n is drawn, so $\mathbb{E}Z_n^2 \le \epsilon_B$. Although stopping depends on that batch,

$$\mathbb{E}\sum_{n,k} I_{n,k}Z_n^2 \le K\sum_n \mathbb{E}Z_n^2 \le NK\epsilon_B.$$

After dividing by R, the finite-batch term is at most $\chi\epsilon_B$. All trajectory and clipping population bounds hold at each executed step under equation H.1; their weighted sums therefore divide by R without a χ factor. This proves the active-update versions of the batch bridge.

Since I_t is predictable, multiplying the actor conditional progress inequality and critic conditional recursion by I_t preserves every conditional noise cancellation. On inactive slots, both parameters and

the current-policy critic minimizer remain unchanged, so there is no progress or contraction to count. Summing over all slots telescopes the actual endpoints and yields

$$\bar{G}_R \le A_{0,R} + 4\bar{B}_R, \quad \bar{B}_R \le a\bar{\Delta}_R + F_R, \quad \bar{S}_R \le \sigma_a^2 + 2\bar{G}_R + 2\bar{B}_R, \quad \bar{\Delta}_R \le U_R + d\bar{S}_R.$$

The same scalar elimination as in Appendix E proves the result. Predictability avoids conditioning the martingale law on a future acceptance event, while unchanged parameters on inactive slots contribute no critic contraction. □

Meaning of the output measure. The left sides are ratios of expectations, not expectations of a random ratio. They describe the occupation distribution of executed updates across runs, which weights a run proportionally to its number of executed updates. If instead a single run returns an index uniformly among its own M executed steps, then $M \ge N$ gives the explicit guarantees

$$\mathbb{E}\left[\frac{1}{M}\sum_t I_t \|\nabla J(\theta_t)\|^2\right] \le \frac{R}{N}\bar{G}_R, \qquad \mathbb{E}\left[\frac{1}{M}\sum_t I_t \Delta_t\right] \le \frac{R}{N}\bar{\Delta}_R.$$

Since $R/N \le K$, these also imply vanishing error under the classical schedules when K is fixed. If all NK slots execute, then $M = T = NK$ deterministically, the occupation and ordinary uniform-output measures coincide, and $\chi = 1$.

I An Explicit Finite-Batch Actor–Critic Example

We construct an analytic example with two-step trajectories, a learned tabular critic, and repeated PPO-Clip updates on a finite rollout batch. Population coverage and critic curvature hold along every realized iterate. Explicit bounds account for both empirical error and old-policy sampling. No projection, empirical KL acceptance test, or population-gradient oracle is used.

MDP, objective, and policy. Set $H = 2$ and $\gamma = 1$. There is one initial state s_0 and two states s_+, s_- at time one. Both actions belong to $\{-1, +1\}$, the first action a_0 determines $s_1 = s_{a_0}$, and the second action terminates the episode. The rewards are $r_0 = 0$ and

$$R(a_0, a_1) = r_1 = \frac{1 + a_0 a_1}{2} \in \{0, 1\}.$$

The terminal value is zero. For a scalar parameter $\theta \in \mathbb{R}$, use the same policy at all nonterminal states:

$$q(\theta) = \frac{1}{2}\tanh\theta, \qquad \pi_\theta(a \mid s) = \frac{1 + aq(\theta)}{2}. \tag{I.1}$$

This is a globally exploratory smooth policy: each action has probability in $[1/4, 3/4]$. Equivalently, it mixes the uniform policy and the binary softmax policy with equal weights. The four possible trajectories form a support Ω with $|\Omega| = 4$. Under the normalized objective used in the analysis,

$$J(\theta) = \frac{1}{2}\mathbb{E}_\theta R = \frac{1 + q(\theta)^2}{4}, \qquad J'(\theta) = \frac{q(\theta)q'(\theta)}{2}. \tag{I.2}$$

The objective is nonconstant. We take the fixed initial actor $\theta_0 = 1$. In particular, $J'(\theta_0) > 0$.

Critic, Monte Carlo targets, and updates. Use three parameters $w = (w_0, w_+, w_-)$, with $V_w(s_i) = w_i$, and initialize all three at $1/2$. Both Monte Carlo returns in a trajectory equal R. The current-policy population reference loss is

$$L_c(\theta, w) = \frac{1}{2}\mathbb{E}_\theta\big[(w_0 - R)^2 + (w_{a_0} - R)^2\big]. \tag{I.3}$$

This is the time-average squared loss. Its unique minimizer is the true, unnormalized value function:

$$w^\star(\theta) = \left(\frac{1 + q(\theta)^2}{2}, \frac{1 + q(\theta)}{2}, \frac{1 - q(\theta)}{2}\right). \tag{I.4}$$

The actor objective is normalized by the horizon, whereas the value targets are unnormalized returns.

At the start of epoch n, let $\bar{\theta} = \theta_n^{(0)}$ and draw B independent trajectories from $\pi_{\bar{\theta}}$. Keep their observed rewards and Monte Carlo targets throughout the epoch. Execute exactly $K \geq 2$ joint updates, with N, K deterministic. At each update, sample $m \geq 1$ trajectory indices independently with replacement from this batch. Use the current pre-update critic w to recompute raw GAE, with $\lambda = 1/2$:

$$\widehat{A}_0 = -w_0 + (1-\lambda)w_{a_0} + \lambda R, \qquad \widehat{A}_1 = R - w_{a_0}. \tag{I.5}$$

The labels are neither normalized nor truncated and are detached when differentiating the actor surrogate. The clipping threshold is $\delta = 1/5$. For one sampled trajectory, the actor direction is

$$g(\theta, w; \tau) = \frac{1}{2}\sum_{h=0}^{1} \alpha_h \frac{\pi_\theta(a_h \mid s_h)}{\pi_{\bar{\theta}}(a_h \mid s_h)} \partial_\theta \log \pi_\theta(a_h \mid s_h)\, \widehat{A}_h,$$

where α_h is the derivative selector of the clipped sample objective: it is one on the active unclipped branch, zero on the inactive branch, and belongs to $[0,1]$ at the corresponding boundary, using a fixed samplewise rule (we select the unclipped branch at ties). The actor update is $\theta^+ = \theta + \eta g_t$, where g_t is the average of the m sampled directions. The critic uses the average of the Monte Carlo gradients

$$\psi(w;\tau) = (w_0 - R)e_0 + (w_{a_0} - R)e_{a_0}, \qquad w^+ = w - \beta_c \frac{1}{m}\sum_{i=1}^{m} \psi(w;\tau_i). \tag{I.6}$$

Here e_i denotes a coordinate vector. Both directions use the same pre-update pair (θ, w). Their minibatch indices may be shared, so their noises need not be independent. After K actual updates, collect the next batch under the new actor.

Boundedness and regularity along all iterates. For $0 < \beta_c \leq 1$, each updated critic coordinate is a convex combination of its old value and sampled targets in $[0,1]$. For example, if f_a is the fraction of minibatch trajectories with first action a, then

$$w_a^+ = (1 - \beta_c f_a) w_a + \frac{\beta_c}{m} \sum_{i: a_{0,i} = a} R_i.$$

The same formula with $f_0 = 1$ applies to w_0. Consequently $w_t \in [0,1]^3$ for every realized batch and minibatch sequence. Equation I.5 then implies $|\widehat{A}_h| \leq 1$: the target $(1-\lambda)w_{a_0} + \lambda R$ is also in $[0,1]$.

The policy bounds hold for every $\theta \in \mathbb{R}$:

$$\frac{\pi_\theta(a \mid s)}{\pi_{\bar{\theta}}(a \mid s)} \leq 3, \qquad |\partial_\theta \log \pi_\theta(a \mid s)| = \left|\frac{a q'(\theta)}{1 + a q(\theta)}\right| \leq 1, \qquad |q'(\theta)| \leq \frac{1}{2}, \quad |q''(\theta)| \leq 1.$$

The product admits a tighter bound than the separate ratio and score envelopes:

$$\left|\frac{\pi_\theta(a \mid s)}{\pi_{\bar{\theta}}(a \mid s)} \partial_\theta \log \pi_\theta(a \mid s)\right| = \frac{|q'(\theta)|}{1 + a q(\bar{\theta})} \leq 1.$$

Thus $|g_t| \leq 1$ and $\|\psi(w_t;\tau)\|^2 \leq 2$. Conditioning on the full reused batch and the past, independent sampling with replacement gives the valid conditional variance bounds

$$\sigma_a^2 = \frac{1}{m}, \qquad \sigma_c^2 = \frac{2}{m}. \tag{I.7}$$

These are variances about the empirical conditional means, not assertions of population unbiasedness. Also $J \leq J_{\max} = 5/16$ and $|J''| \leq (|q'|^2 + |qq''|)/2 \leq 3/8$, so $L_J = 1$ is valid.

Writing $p_\pm = (1 \pm q)/2$, the critic Hessian is

$$\nabla_w^2 L_c(\theta, w) = \mathrm{diag}(1, p_+, p_-).$$

Therefore the loss is globally convex and $L_c = 1$ smooth, with quadratic-growth modulus $\mu_q = 1/4$ around the unique stationary minimizer equation I.4. The PL condition also holds with modulus $1/4$. The time-averaged state probabilities are $(1/2, p_+/2, p_-/2)$, whose minimum is at least $1/8$. Population coverage therefore holds uniformly over all actor iterates, although a finite empirical batch may omit some states. Since

$$\partial_\theta w^\star = (qq', q'/2, -q'/2), \qquad \|\partial_\theta w^\star\|^2 \leq \frac{3}{16},$$

we may take $L_{w^\star} = \sqrt{3}/4$. The value parameterization has $L_V = G_V = 1$. With the stated initialization, $\Delta_0 \leq 9/64$.

A deterministic implementable KL budget. Let $p = \pi_\theta(+1 \mid s)$ and $\bar{p} = \pi_{\bar{\theta}}(+1 \mid s)$. Using $\mathrm{KL} \le \chi^2$, $p(1-p) \ge 3/16$, and $|p - \bar{p}| \le |\theta - \bar{\theta}|/4$ gives

$$\mathrm{KL}(\pi_{\bar{\theta}}(\cdot \mid s) \| \pi_\theta(\cdot \mid s)) \le \frac{(p-\bar{p})^2}{p(1-p)} \le \frac{|\theta - \bar{\theta}|^2}{3}. \tag{I.8}$$

The policy is state independent, so this is also the required state-averaged KL bound. Because $|g_t| \le 1$, every pre-update and post-update actor within an epoch satisfies $|\theta - \bar{\theta}| \le K\eta$. It is therefore sufficient to choose

$$K^2\eta^2/3 \le \kappa. \tag{I.9}$$

This rule uses known constants and the prescribed epoch length, and certifies population KL for every realized run. There is no empirical accept/reject conditioning or random stopping in this example.

Finite-batch and old-policy residuals. Let $\widehat{P}_n$ be the empirical distribution on Ω and let $P_{\bar{\theta}}$ be its rollout law. Conditionally on the history before collecting the batch,

$$\mathbb{E}\big[\mathrm{TV}(\widehat{P}_n, P_{\bar{\theta}})^2\big] \le \min\left\{1, \frac{|\Omega|}{4B}\right\} =: \epsilon_B = \min\{1, 1/B\}.$$

Indeed, Cauchy–Schwarz bounds squared total variation by $|\Omega|\|\widehat{P}_n - P_{\bar{\theta}}\|_2^2/4$, whose expectation is at most $|\Omega|/(4B)$. The resulting finite-support bound is uniform over the data-dependent actor and critic iterates: for every bounded trajectory function f, the inequality $\|(\widehat{P}_n - P_{\bar{\theta}})f\| \le 2\|f\|_\infty \mathrm{TV}(\widehat{P}_n, P_{\bar{\theta}})$ holds pathwise. Thus adaptive batch reuse is covered while retaining the empirical-to-population discrepancy.

The generic bridge permits $R_{\max} = 3$, $G_\pi = L_V = G_V = A_{\max} = 1$, $V_{\max} = G_{\max} = 1$ and $Q_c = 4$. Direct calculation gives tighter bounds on three components for this example.

First, the actor integrand has envelope one, so its uniform squared empirical error is at most $4\epsilon_B = 4/B$. The four-term decomposition of the actor bias therefore contributes $16/B$. Its population drift and clipping contributions remain $8(39+2\sqrt{2})\kappa$. Second, equation I.12 below gives the exact current-policy GAE bias. Since $|q'| \le 1/2$ and $(e_+ - e_-)^2 \le 2\|e\|^2$, its square is at most $(1-\lambda)^2\Delta_t/32$. Multiplying by the four-term factor yields $a = (1-\lambda)^2/8 = 1/32$ at $\lambda = 1/2$.

Third, the empirical critic discrepancy admits a direct covariance bound. Write $\delta p_{ab} = \widehat{P}(a_0 = a, a_1 = b) - P_{\bar{\theta}}(a_0 = a, a_1 = b)$ and $\delta r = \delta p_{++} + \delta p_{--}$. Its initial-state component is $-\delta r$; its component for first action a is $w_a \delta p_{a,-a} - (1-w_a)\delta p_{a,a}$. Uniformly over $w \in [0,1]^3$,

$$\|(\widehat{P} - P_{\bar{\theta}})\psi_w\|^2 \le (\delta r)^2 + \sum_{a,b} (\delta p_{ab})^2.$$

Conditionally on the prebatch history, the expectation of this upper bound is

$$\frac{\mathrm{Var}(R) + 1 - \sum_{a,b} P_{\bar{\theta}}(a,b)^2}{B} \le \frac{1/4 + 3/4}{B} = \frac{1}{B}.$$

This pathwise envelope also controls a critic selected using the batch. For the old-policy component, regarding the population gradient as a function of q, its coordinates are

$$(w_0 - (1+q^2)/2,\ p_+ w_+ - p_+^2,\ p_- w_- - p_-^2).$$

Their derivatives in q have absolute bounds $1/2, 3/4, 3/4$, respectively, for all $w \in [0,1]^3$. Thus the squared population-gradient difference is at most $(11/8)|q - \bar{q}|^2$. Pinsker's inequality for the common state-independent Bernoulli policy gives $|q - \bar{q}|^2 \le 2\kappa$. Combining the two critic components using a factor two yields $E_c = 2/B + 11\kappa/2$.

Consequently the actual conditional biases obey

$$\frac{1}{T}\sum_{t<T} \mathbb{E}|b_t|^2 \le a\,\frac{1}{T}\sum_{t<T}\mathbb{E}\Delta_t + F, \qquad a = \frac{1}{32}, \tag{I.10}$$

$$F = \frac{16}{B} + 8(39 + 2\sqrt{2})\kappa,$$

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|e_t^c\|^2 \le E_c, \qquad E_c = \frac{2}{B} + \frac{11}{2}\kappa. \tag{I.11}$$

These residuals retain the empirical and old-policy target discrepancies. Substituting these constants into the argument in Appendix E gives the stated closed-loop bounds without assuming population unbiasedness.

Nontrivial feedback. The actor objective is nonconstant, and the critic is inexact at initialization. At an epoch start with $\bar{\theta} = \theta$, clipping is inactive and a population expectation of the raw-GAE direction gives

$$\mathbb{E}_\theta g(\theta, w; \tau) - J'(\theta) = \frac{(1-\lambda)q'(\theta)}{4}\big[(w_+ - w_+^\star) - (w_- - w_-^\star)\big]. \tag{I.12}$$

At the stated initialization this equals $-(1-\lambda)J'(\theta_0)/2 \neq 0$. Thus the learned critic changes the mean actor direction. Conversely, the nonconstant map equation I.4 moves when the actor changes. The formula is a population diagnostic of the coupling; the finite-batch discrepancies are separately retained in equation I.10–equation I.11. The example uses the clipped update but does not require clipping to activate at every admissible stepsize.

A nonempty stepsize regime and a vanishing-error family. The critic and actor restrictions are satisfied by

$$0 < \beta_c \leq \frac{1}{8}, \qquad 0 < \eta \leq \min\left\{\frac{1}{4}, \frac{\beta_c}{10}, \frac{\sqrt{3\kappa}}{K}\right\}.$$

The instance-specific bias coefficient gives

$$d = 120\frac{\eta^2}{\beta_c^2}, \qquad \rho = 10ad = \frac{75}{2}\frac{\eta^2}{\beta_c^2} \leq \frac{3}{8} < 1.$$

All constants hold along actual iterates without constraining θ to a compact set. For fixed $K \geq 2$, $0 < b \leq 1/8$, and $m \geq 1$, take T through multiples of K and set

$$\begin{gathered}\beta_{c,T} = bT^{-2/5}, \qquad \eta_T = \frac{b}{10}T^{-3/5}, \qquad \kappa_T = K^2\eta_T^2/3, \\ B_T = \lceil T^{2/5} \rceil, \qquad N(T) = T/K.\end{gathered} \tag{I.13}$$

Then $\rho_T = (3/8)T^{-2/5}$ and both $F, E_c = O(T^{-2/5})$. Corollary 3.3 gives

$$\frac{1}{T}\sum_{t<T}\mathbb{E}|J'(\theta_t)|^2 = O(T^{-2/5}), \qquad \frac{1}{T}\sum_{t<T}\mathbb{E}\Delta_t = O(T^{-2/5}), \qquad Q = NB = O(T^{7/5}/K).$$

The constants are uniform in T because the exploration floor is global and the critic cube is invariant. This is a synchronous example; it does not by itself establish the delayed or randomly stopped variants.

Numerical calibration of the complete bound. The elementary envelope for this objective is

$$\sup_\theta |J'(\theta)|^2 = \max_{x\in[-1,1]} \frac{x^2(1-x^2)^2}{64} = \frac{1}{432}.$$

Use $T = 10^6$, $K = 2$, $B = 10^6$, $m = 128$, $\eta = 10^{-4}$, and $\beta_c = 0.02$. The actual-iterate certificate is $\kappa = 4 \cdot 10^{-8}/3$, and $\rho = 0.0009375$. Writing $q_0 = \frac{1}{2}\tanh 1$, the exact initial quantities are $J_{\max} - J(1) = 1/16 - q_0^2/4$ and $\Delta_0 = q_0^4/4 + q_0^2/2$. Substitution in equation 3.2 with equation I.10–equation I.11 gives

$$\bar{G} < 0.002094 < 1/432 \approx 0.002315.$$

This analytic certificate improves on the elementary stationarity envelope; it is not a training result. It requires $Q = 5 \cdot 10^{11}$ fresh trajectories and $HQ = 10^{12}$ transitions; the bound is therefore not a practical sample requirement and does not establish a tight quantitative prediction.

J A Growing-Horizon Example with Polynomial Resources

The two-step example extends to every horizon $H \geq 2$ without introducing a separate state for each action prefix. The construction satisfies the assumptions along every realized update sequence. Applying the structured finite-batch bound gives polynomial resource bounds without an exponential trajectory-support factor. These upper bounds are conservative and do not provide practical tuning recommendations.

An $O(H)$-state chain and a shared policy. There is one initial state s_0 and two states $s_{h,+}, s_{h,-}$ at each time $h \in \{1, \dots, H-1\}$. Both actions lie in $\{-1, +1\}$; taking action a_h moves to s_{h+1,a_h} when $h < H-1$, and the final action terminates. Thus the state records only time and the immediately preceding action. Set $\gamma = 1$, terminal value zero, and

$$r_0 = 0, \qquad r_h = \mathbf{1}\{a_h = a_{h-1}\} \quad (1 \le h < H).$$

Use the same scalar policy at every state,

$$q(\theta) = \tfrac{1}{2}\tanh\theta, \qquad \pi_\theta(a \mid s) = \frac{1 + aq(\theta)}{2}, \qquad \theta_0 = 1.$$

Under a fixed policy, the actions are independent with mean q. Writing $c(\theta) = (1 + q(\theta)^2)/2$, the normalized objective is

$$J_H(\theta) = \frac{H-1}{H} c(\theta), \qquad J_H'(\theta) = \frac{H-1}{H} q(\theta) q'(\theta). \tag{J.1}$$

This objective is nonconstant uniformly in the horizon: $\sup_\theta J_H(\theta) - J_H(0) = (H-1)/(8H) \ge 1/16$ and $J_H'(1) \ge q(1)q'(1)/2 > 0$.

A jointly learned tabular critic. There are $d_H = 2H - 1$ critic coordinates, with $V_w(s) = w_s$ and fixed terminal value zero. Initialize

$$w_{0,0} = \frac{H-1}{2}, \qquad w_{0,h,+} = w_{0,h,-} = \frac{H-h}{2} \quad (1 \le h < H).$$

For the observed Monte Carlo return $G_h = \sum_{j=h}^{H-1} r_j$, the reference loss and its unique minimizer are

$$\mathcal{L}_c(\theta, w) = \frac{1}{H}\mathbb{E}_\theta \sum_{h=0}^{H-1} (w_{s_h} - G_h)^2, \tag{J.2}$$

$$w_0^\star(\theta) = (H-1)c(\theta),$$

$$w_{h,x}^\star(\theta) = \frac{1 + xq(\theta)}{2} + (H - 1 - h)c(\theta), \qquad x \in \{-1, +1\}. \tag{J.3}$$

In particular this critic realizes the unnormalized true values.

At each of N deterministic epochs, collect B independent trajectories under the epoch-start actor $\bar\theta$, retain their rewards and returns, and execute exactly $K \ge 2$ joint updates. Each update samples $m \ge 1$ complete trajectories independently with replacement from that batch. Recompute raw GAE with $\lambda = 1/2$ and the current pre-update critic, detach its labels in the actor derivative, and use PPO-Clip with $\delta = 1/5$. For one sampled path the directions are

$$g(\theta, w; \tau) = \frac{1}{H} \sum_{h=0}^{H-1} \alpha_h \frac{\pi_\theta(a_h \mid s_h)}{\pi_{\bar\theta}(a_h \mid s_h)} \partial_\theta \log \pi_\theta(a_h \mid s_h)\, \widehat{A}_h(w, \tau),$$

$$\psi(w; \tau) = \frac{2}{H} \sum_{h=0}^{H-1} (w_{s_h} - G_h) e_{s_h}. \tag{J.4}$$

Here α_h is the clipped sample derivative selector, choosing the unclipped branch at ties. The actor and critic update from the same pre-update pair by $\theta^+ = \theta + \eta g_t$ and $w^+ = w - \beta_c h_t$, where g_t, h_t are the minibatch averages. Their sampled indices may be shared.

Proposition J.1 (Polynomial resources for a growing-horizon chain)**.** *For every integer $H \ge 2$, fixed integers $K \ge 2, m \ge 1$, and $0 < b \le 1/8$, let $T = NK$ and choose*

$$\beta_{c,T} = bT^{-2/5}, \qquad \eta_T = \frac{b}{40H^3} T^{-3/5}, \qquad \kappa_T = 12H^2K^2\eta_T^2,$$

$$B_T = \left\lceil H^3 T^{2/5} \log(eHT) \right\rceil. \tag{J.5}$$

The process above satisfies all synchronous actor–critic premises along every realized sequence of batches and minibatches. In particular the population KL budget is certified by the deterministic step sizes. Define

$$\begin{aligned} A_H &= 20H^3/b + 2b/m, \\ F_H &= 101376 + 6b^2K^2, \\ U_H &= H^4/(4b) + 40bH^2/m + 5120 + 20b^2K^2, \\ P_H &= 2\big[A_H + 4F_H + 16U_H + 8H/m\big]. \end{aligned}$$

Then the actual T updates obey

$$\frac{1}{T}\sum_{t<T}\mathbb{E}|J'_H(\theta_t)|^2 \le P_H T^{-2/5}, \qquad \frac{1}{T}\sum_{t<T}\mathbb{E}\|w_t - w^\star(\theta_t)\|^2 \le P_H T^{-2/5}. \tag{J.6}$$

For $0 < \varepsilon \le 1$, both errors are at most ε when $T = K\lceil (P_H/\varepsilon)^{5/2}/K\rceil$. Counting newly collected trajectories as $Q = NB_T$, the resource upper bounds are

$$Q = O\left(\frac{H^3}{K}(P_H/\varepsilon)^{7/2}\log(eHP_H/\varepsilon)\right), \qquad HQ = O\left(\frac{H^4}{K}(P_H/\varepsilon)^{7/2}\log(eHP_H/\varepsilon)\right). \tag{J.7}$$

For fixed b, m, K, $P_H = O(H^4)$; hence these are respectively $\widetilde{O}(H^{17}\varepsilon^{-7/2})$ collected trajectories and $\widetilde{O}(H^{18}\varepsilon^{-7/2})$ environment steps.

Proof. We establish the premises along the update sequence and substitute the resulting constants into the synchronous bound.

Invariant critic region and bounded directions. All Monte Carlo targets lie in $[0, H]$. For a coordinate i, let f_i be the fraction of minibatch trajectories visiting its state. Its update is

$$w_i^+ = \left(1 - \frac{2\beta_c f_i}{H}\right) w_i + \frac{2\beta_c}{Hm}\sum_{j:s_i\in\tau_j} G_i(\tau_j).$$

The nonnegative weights sum to one when $0 < \beta_c \le H/2$. Therefore $[0, H]^{d_H}$ is invariant for every batch and minibatch realization, including batches missing some states. It contains the initialization and every $w^\star(\theta)$. Telescoping the matched-terminal raw GAE gives

$$\widehat{A}_h = -w_{s_h} + \sum_{\ell=0}^{H-1-h} 2^{-\ell} r_{h+\ell} + \frac{1}{2}\sum_{\ell=1}^{H-1-h} 2^{-(\ell-1)} w_{s_{h+\ell}}. \tag{J.8}$$

The two positive sums are at most 2 and H, respectively. Hence $-H \le \widehat{A}_h \le H+2$, so $A_{\max} = 2H$ is valid for $H \ge 2$.

For every real θ the action probabilities lie in $[1/4, 3/4]$, the likelihood ratios are at most three, and

$$|\partial_\theta \log \pi_\theta(a \mid s)| = \left|\frac{aq'(\theta)}{1 + aq(\theta)}\right| \le 1, \qquad |q'| \le \tfrac{1}{2}, \qquad |q''| \le 1.$$

Thus $|g_t| \le 6H$. Since a trajectory visits H distinct time-state coordinates and $|w_{s_h} - G_h| \le H$, the squared norm of its critic direction is at most $4H$. The conditional variances about the empirical means therefore satisfy

$$\sigma_a^2 = 36H^2/m, \qquad \sigma_c^2 = 4H/m. \tag{J.9}$$

These conditional means retain finite-batch bias.

Population coverage, critic geometry, and target motion. Writing $p_x = (1 + xq)/2$, the time-averaged state masses are $1/H$ at s_0 and p_x/H at $s_{h,x}$. Their minimum is at least $1/(4H)$, uniformly over all real actor parameters. The critic Hessian is diagonal with entries $2/H$ at the initial state and $2p_x/H$ at the other states. Consequently

$$\mu_q = \frac{1}{2H}, \qquad L_c = \frac{2}{H}, \qquad L_V = G_V = 1 \tag{J.10}$$

are valid global quadratic-growth, smoothness, and value regularity constants. The minimizer in equation J.3 is stationary and unique. Every coordinate of its derivative has absolute value at most $H/4$, since

$$(w_0^\star)' = (H-1)qq', \qquad (w_{h,x}^\star)' = xq'/2 + (H-1-h)qq'.$$

We may therefore take $L_{w^\star}^2 = H^3/8$. At initialization the coordinate errors are

$$w_{0,0} - w_0^\star = -(H-1)q(1)^2/2, \qquad w_{0,h,x} - w_{h,x}^\star = -xq(1)/2 - (H-1-h)q(1)^2/2.$$

Their absolute values are at most $H/8$, giving $\Delta_0 \le H^3/32$. Finally, $|J_H''| \le |q'|^2 + |qq''| \le 3/4$, so $L_J = 1$ is valid, and $J_{\max} - J_H(1) \le 1/8$.

A pathwise certificate for the population KL budget. For $p = \pi_\theta(+1 \mid s)$ and $\bar{p} = \pi_{\bar{\theta}}(+1 \mid s)$, the elementary Bernoulli comparison gives

$$\mathrm{KL}(\pi_{\bar{\theta}} \| \pi_\theta) \le \frac{(p-\bar{p})^2}{p(1-p)} \le \frac{|\theta - \bar{\theta}|^2}{3}.$$

The last inequality uses $p(1-p) \ge 3/16$ and $|p - \bar{p}| \le |\theta - \bar{\theta}|/4$. Since $|g_t| \le 6H$, every pre-update and post-update actor in the epoch satisfies $|\theta - \bar{\theta}| \le 6HK\eta_T$. This proves the budget $\kappa_T = 12H^2K^2\eta_T^2$ in equation J.5 along every realized run. The policy is state independent, so the same inequality is the required state-averaged population KL certificate.

Structured finite-batch errors. There are $M_H = 2d_H = 4H - 2$ time-state-action cells. Raw GAE is $C_{1/2,H}$-Lipschitz in the critic infinity norm, with $C_{1/2,H} = 2 - 2^{-(H-1)} \le 2$. Apply Lemma C.6 with $W = G_{\max} = H$, $A_{\max} = 2H$, $R_{\max} = 3$, and $G_\pi = 1$. Its empirical actor and critic errors satisfy

$$\begin{aligned} e_a(B) &\le \frac{2304H^2}{B}\big[\log(24H) + 2H\log(B+2) + 3\big], \\ e_c(B) &\le \frac{16H}{B}. \end{aligned} \tag{J.11}$$

These are uniform bounds over the data-dependent iterates in the critic cube; they do not enumerate the 2^H possible action sequences. For the trajectory-level critic integrand, the bound above permits $Q_c = 2\sqrt{H}$. The score-baseline cancellation gives $\widetilde{C}_{1/2,H} = 1 - 2^{-(H-1)} \le 1$, so $a \le 4$. Corollary C.7 then gives

$$\begin{aligned} F &\le 4e_a(B) + 32H^2\kappa\big[(1+\sqrt{H})^2 + 36\big] \le 4e_a(B) + 672H^3\kappa, \\ E_c &\le 2e_c(B) + 16H^2\kappa. \end{aligned} \tag{J.12}$$

The old-policy continuation and stored-target discrepancies are included in the κ terms.

Let $L = \log(eHT)$. The chosen batch size satisfies $B_T \ge H^3T^{2/5}L$, $\log(B_T+2) \le 4L$, and $\log(24H) \le 3L$ for $H \ge 2, T \ge 1$. Hence the bracket in equation J.11 is at most $11HL$, and

$$e_a(B_T) \le 25344T^{-2/5}, \qquad F \le F_H T^{-2/5}.$$

The structured concentration bound removes the complete-trajectory support factor from this step.

Admissible steps and substitution. The critic restriction is $\beta_c \le \min\{\mu_q/(2L_c^2), 1/\mu_q\} = H/16$; it and cube invariance follow from $b \le 1/8$, $H \ge 2$. The actor step also satisfies $\eta_T \le 1/(4L_J)$. With the constants above,

$$d = \frac{40L_{w^\star}^2\eta_T^2}{\mu_q^2\beta_{c,T}^2} \le \frac{1}{80H}T^{-2/5}, \qquad \rho = 10ad \le \frac{1}{2H}T^{-2/5} \le \frac{1}{4}. \tag{J.13}$$

The open-loop terms satisfy

$$\begin{aligned} A_0 &= \frac{4(J_{\max} - J_H(1))}{\eta_T T} + 2L_J\eta_T\sigma_a^2 \le A_H T^{-2/5}, \\ U &= \frac{4\Delta_0}{\mu_q\beta_{c,T}T} + \frac{5\beta_{c,T}\sigma_c^2}{\mu_q} + \frac{40E_c}{\mu_q^2} \le U_H T^{-2/5}. \end{aligned}$$

For the last bound, the three terms are at most $H^4T^{-3/5}/(4b)$, $40bH^2T^{-2/5}/m$, and $(5120 + 20b^2K^2)T^{-2/5}$, respectively. The synchronous closure of Theorem 3.1, with $a \le 4$ and $(1-\rho)^{-1} \le 4/3$, gives

$$\bar{G}_T \le \tfrac{4}{3}\big[A_H + 4F_H + 16U_H + 36H/(5m)\big]T^{-2/5},$$
$$\bar{\Delta}_T \le \tfrac{4}{3}\big[U_H + 9H/(20m) + (2A_H + 10F_H)/(80H)\big]T^{-2/5}.$$

Both right sides are at most $P_H T^{-2/5}$. Finally $Q = (T/K)B_T$ and every collected path has H steps. Substituting the stated T proves equation J.7; the ceiling changes only numerical constants because $P_H \ge K^{2/5}$. □

Nonzero actor–critic feedback. The critic is neither exact initially nor independent of the actor. For a fixed on-policy pair (θ, w) at an epoch start, put $e_h(s) = V_w(s) - V^{\pi_\theta}(s)$ and $z_h = \partial_\theta \log \pi_\theta(a_h)$. The GAE error relative to the true critic is

$$-e_h(s_h) + (1-\lambda)\sum_{j=h+1}^{H-1} \lambda^{j-h-1} e_j(s_j).$$

The score has conditional mean zero, so its product with $e_h(s_h)$ averages to zero. For $j \ge h+2$, $s_j = s_{j,a_{j-1}}$ is independent of a_h under this shared state-independent policy, so those terms also average to zero. Only $j = h+1$ survives, and

$$\mathbb{E}_\theta[z_h e_{h+1}(s_{h+1})] = \frac{q'(\theta)}{2}\big[e_{h+1}(s_{h+1,+}) - e_{h+1}(s_{h+1,-})\big].$$

At an epoch start clipping is inactive. Consequently

$$\mathbb{E}_\theta g(\theta, w; \tau) - J'_H(\theta) = \frac{(1-\lambda)q'(\theta)}{2H}\sum_{j=1}^{H-1}\big[e_j(s_{j,+}) - e_j(s_{j,-})\big]. \tag{J.14}$$

At the stated initialization each bracket equals $-q(1)$, so this population actor bias is $-J'_H(1)/4 \ne 0$. Conversely the selected critic target equation J.3 changes with the actor. This diagnostic concerns the population mean; the stored-batch residuals remain in equation J.12.

The guarantee concerns expected stationarity and critic tracking for this analytic family and does not assert global policy optimality. In particular, fixing K and m still gives a polynomial sample bound while H grows; uniform vanishing along a joint sequence follows, for example, when $H = o(T^{1/10})$.

K ASYNCHRONOUS PPO-CLIP

K.1 ASYNCHRONOUS PPO-CLIP: PARAMETER-SERVER VS. DECOUPLED ROLLOUT–TRAINING

Large-model PPO-Clip systems can use *asynchronous* execution to improve throughput, mitigate stragglers, and accommodate heterogeneous hardware, such as many rollout workers paired with fewer training GPUs. Asynchrony introduces *staleness*: rollouts or gradients may be computed using parameters that lag behind the current model. We distinguish two asynchronous paradigms and explain their relation to the analysis.

Baseline: synchronous PPO-Clip. In the synchronous setting, outer iteration n collects a fresh batch under π_{θ_n} (and corresponding critic V_{w_n}), then performs K inner updates, producing (θ_{n+1}, w_{n+1}). In asynchronous settings, this "iteration barrier" is relaxed: the data used at some update step may have been generated by an older policy $\pi_{\theta_{n-\tau}}$ and/or evaluated by an older critic $V_{w_{n-\tau'}}$, where $\tau, \tau' \geq 0$ denote staleness delays.

K.1.1 MODEL I: PARAMETER-SERVER ASYNCHRONOUS PPO-CLIP

System architecture. In the *parameter-server (PS)* paradigm, a central server maintains the latest actor–critic parameters (θ, w). A set of workers interacts with environments (or prompts) to generate trajectories and compute per-minibatch gradients, while the server applies incoming updates asynchronously.

Asynchronous update rule. At (global) update index t, a worker pulls a *stale* snapshot $(\theta_{t-\tau_t}, w_{t-\tau'_t})$ from the server, computes a stochastic (sub)gradient of the PPO-Clip objective and a critic gradient using its assigned retained behavior-policy batch, and then pushes the gradients back to the server. The server applies these updates immediately:

$$\theta_{t+1} = \theta_t + \eta\, g_t(\theta_{t-\tau_t}, w_{t-\tau'_t}; \xi_t), \tag{K.1}$$

$$w_{t+1} = w_t - \beta_c\, h_t(\theta_{t-\tau_t}, w_{t-\tau'_t}; \xi_t), \tag{K.2}$$

where ξ_t denotes the random minibatch/trajectory data used by the worker and the delays τ_t, τ'_t capture staleness in the actor and critic snapshots, respectively. In PPO-Clip, the worker also maintains a behavior reference θ_{old} for importance ratios $r_{\theta,\theta_{\text{old}}}$ and clipping, which is the snapshot that generated the retained batch. It can precede the actor snapshot at which the current job is evaluated.

Effects of delayed evaluation. Evaluating updates at stale parameters and applying them to current parameters can introduce systematic *staleness bias* in addition to the other PPO error terms:

- **Policy-gradient mismatch:** the conditional update direction can differ from $\nabla J(\theta_t)$, even without clipping, because it is evaluated at $\theta_{t-\tau_t}$ using a potentially older behavior-policy batch.
- **Critic target drift under delay:** critic updates track $w^\star(\theta_{t-\tau_t})$ rather than $w^\star(\theta_t)$, adding delay-dependent terms to the tracking bound.
- **PPO-specific distortion under delay:** clipping and KL control compare the evaluation policy with the retained batch's behavior reference, which can differ from the current server policy.

The analysis bounds these effects through the *parameter drift* over the delay window, e.g., $\|\theta_t - \theta_{t-\tau_t}\|$ and $\|w_t - w_{t-\tau'_t}\|$.

K.1.2 MODEL II: DECOUPLED ROLLOUT–TRAINING ASYNCHRONOUS PPO-CLIP

System architecture. A second paradigm decouples *trajectory generation* (rollout) from *training* (optimization). A large pool of rollout workers continuously generates trajectories using periodically refreshed policy snapshots, and writes them into a replay-like buffer or queue. Separately, one or more trainers repeatedly sample batches from this buffer and perform PPO-Clip updates. Unlike off-policy RL, the buffer is *shallow* and is refreshed frequently; nonetheless, the trainer typically operates on *stale* data generated by older snapshots.

Data and parameter staleness. Let θ_t^{roll} denote the actor snapshot used by a rollout worker to generate a trajectory that later appears in the training batch at step t. The trainer updates the current parameters (θ_t, w_t) using samples generated under $\pi_{\theta_t^{\text{roll}}}$, where θ_t^{roll} may lag behind θ_t by a delay τ_t induced by buffering and queueing. The trainer update can be written as

$$\theta_{t+1} = \theta_t + \eta\, g_t(\theta_t; \theta_t^{\text{roll}}, w_t, \xi_t), \tag{K.3}$$

$$w_{t+1} = w_t - \beta_c\, h_t(\theta_t; \theta_t^{\text{roll}}, w_t, \xi_t), \tag{K.4}$$

where the PPO surrogate uses importance ratios $r_{\theta_t, \theta_t^{\text{roll}}}$ and clipping relative to the behavior policy θ_t^{roll}. In this paradigm, *the computation uses current parameters but stale data*, whereas in the PS paradigm the computation itself may use stale parameters.

Practical variants. Two implementation choices are common:

- **Snapshot-per-epoch rollouts:** rollouts are generated using a fixed snapshot for a window of time, and trainers perform multiple PPO epochs on that window (multi-epoch reuse on buffered data).
- **Streaming rollouts:** rollouts arrive continuously with mixed staleness; trainers enforce a small KL constraint and/or discard overly stale samples to approximate on-policy behavior.

Both variants can be modeled through an upper bound on the delay τ_t and a bound on the induced drift $\|\theta_t - \theta_t^{\text{roll}}\|$.

K.1.3 Comparison and modeling implications

The PS and decoupled paradigms induce different mathematical effects:

- **PS (stale compute):** gradients are evaluated at $(\theta_{t-\tau}, w_{t-\tau'})$ and applied at (θ_t, w_t), which can create a mismatch between the update direction and the current iterate.
- **Decoupled (stale data):** gradients are evaluated at current parameters but on data from an older behavior policy θ^{roll}, creating potential trajectory-policy mismatch and clipping distortion.

Recent parameter drift enters the staleness bounds and interacts with actor–critic coupling. The sufficient asynchronous conditions jointly constrain actor and critic stepsizes according to the coupling and delay bounds; the permitted critic stepsize decreases with its delay bound. The behavior-policy KL premise separately controls data mismatch.

Scope of the theorem. The theorem analyzes the single-gradient parameter-server model under the stated scheduling and sampling premises. Parameter drift is bounded through stepsizes and staleness caps, while the KL budget separately controls policy-distribution mismatch. Applying the result to a decoupled rollout–training system requires verification of the same premises.

Remark (PS staleness and sampling). Algorithm 1 describes the single-gradient read-delay model used in the theorem. A worker job may be computed before application, provided the application schedule and delays do not inspect that job's fresh minibatch noise; equivalently, its empirical mean remains the conditional mean given the pre-application history. Each job uses a critic snapshot taken before its gradient pair is computed. The theorem does not cover multi-step local fitting followed by accumulated parameter updates. The population KL premise can, for example, be guaranteed by a deterministic bound on behavior-snapshot age: if that age in actor updates is at most D_b, then $G_\pi \eta G_g D_b \le \kappa$ suffices, with $G_g = R_{\max} G_\pi A_{\max}$ and the score bound along connecting segments.

K.2 Asynchronous convergence guarantee

We extend the closed-loop finite-time guarantee in Theorem 3.1 to a specified *asynchronous* PPO-Clip model. Two forms of asynchrony arise in RL training: (i) a *training–generation decoupled* setting (Fu et al., 2025b), and (ii) a *parameter-server* setting (Nair et al., 2015; Mnih et al., 2016). In the training–generation decoupled setting, old-data effects can be bounded by the finite-batch trajectory comparison when its behavior-policy KL and data-assignment premises hold. The theorem concerns the single-gradient parameter-server model distinguished from decoupled rollout–training systems in Section K.1. Its additional challenge is *staleness*: rollouts, advantage estimates, and policy gradients may be computed using delayed actor and critic parameters and applied at a later server iterate.

Algorithm 1 Single-gradient parameter-server PPO-Clip with reused batches

Input: B trajectories per batch, K jobs per batch, N batches, $T = NK$.
Input: $H, \gamma, \lambda, \delta$, stepsizes η, β_c, delay limits τ_a, τ_w.
Fix the batch-to-job assignment and server application schedule independently of current job minibatch draws.
Initialize server parameters (θ_0, w_0).
At each scheduled batch collection, draw B independent trajectories under its behavior snapshot $\bar{\theta}_n$ and store their Monte Carlo returns.
for server jobs $t = 0, \dots, T-1$ (workers may compute concurrently) **do**
  Use causal actor/critic snapshots $(\theta_{t-\tau_t}, w_{t-\nu_t})$ with delays within the prescribed limits.
  Use the assigned retained batch $\mathcal{D}_n$ and its behavior reference $\bar{\theta}_n$.
  Maintain the stated population KL budget between $\bar{\theta}_n$ and $\theta_{t-\tau_t}$.
  Draw a fresh empirical minibatch for this job; actor/critic draws may coincide.
  Compute raw GAE using the pre-job critic $w_{t-\nu_t}$, with no normalization or advantage clipping.
  Compute one PPO-Clip actor gradient $\widetilde{g}_t$ at $\theta_{t-\tau_t}$, using weights γ^h and reference $\bar{\theta}_n$.
  Compute one Monte Carlo critic gradient h_t at $w_{t-\nu_t}$ from the stored returns.
  Apply the gradient pair at the server:
    $\theta_{t+1} \leftarrow \theta_t + \eta \widetilde{g}_t$, $w_{t+1} \leftarrow w_t - \beta_c h_t$.
end for
Return: server iterates (or an independently uniform server update index).

Asynchronous update model (parameter server). Let $t = 0, 1, \dots, T-1$ index *server-side* actor updates. This index t corresponds to the flattened index $t = nK + k$ in the synchronous outer–inner structure. At server update t, the parameter server maintains (θ_t, w_t). A worker evaluates a job at a possibly stale snapshot $(\theta_{t-\tau_t}, w_{t-\nu_t})$ using its assigned retained batch. The batch was generated by a behavior policy $\bar{\theta}_n$, which can precede this evaluation snapshot. Advantages are computed using the stale critic $V_{w_{t-\nu_t}}$. The server applies the delayed actor direction $\widetilde{g}_t$, where $\widetilde{g}_t$ is the empirical clipped-gradient estimate at $(\theta_{t-\tau_t}, w_{t-\nu_t})$ using the assigned retained batch. The critic gradient is evaluated at the same pair using stored Monte Carlo targets. Both gradients use the pre-job critic, with no intervening local optimization steps. The population KL bound and fixed job count are premises of the theorem; an empirical post-update stopping check alone does not establish them.

Assumption 7 (Bounded delay)**.** *Assume the staleness is uniformly bounded: there exist finite integers (τ_a, τ_w) such that $0 \le \tau_t \le \min(t, \tau_a)$ and $0 \le \nu_t \le \min(t, \tau_w)$ for all t. Moreover, asynchronous actor update direction satisfies $\mathbb{E}\big[\|\widetilde{g}_t\|^2 \mid \mathcal{F}_t\big] \le \bar{\sigma}_a^2, \forall t$ for some finite constant $\bar{\sigma}_a^2 > 0$.*

Theorem K.1 (Closed-loop non-asymptotic convergence of asynchronous PPO-Clip)**.** *Consider the single-gradient parameter-server model described above, under the finite-batch protocol and Assumptions 1–6 and Assumption 7. Every batch contains B conditionally independent behavior-policy trajectories and is used for exactly K server jobs, with a deterministic assignment of batches to the $T = NK$ jobs. The application order and delays do not inspect a job's fresh minibatch noise; its empirical mean is the conditional mean given the pre-application history. The evaluation indices $s_t = t - \tau_t$, $r_t = t - \nu_t$ are chosen before the current minibatch draw, with $0 \le \tau_t \le \min(t, \tau_a)$ and $0 \le \nu_t \le \min(t, \tau_w)$. The behavior-policy KL budget is imposed relative to the actor θ_{s_t} at which the job is evaluated. Suppose also that $\nabla_w \mathcal{L}_c(\theta, w)$ is L_θ-Lipschitz in θ. The L_c and L_θ bounds hold on all mixed current/stale parameter pairs and connecting segments used in the delay comparison; a common Cartesian region containing the iterates suffices. Set $m = \min\{\mu_q/2, L_c\}$ and require*

$$0 < \eta \le \frac{1}{4L_J}, \qquad 0 < \beta_c \le \frac{m}{24L_c^2(\tau_w + 1)}.$$

Retain a, F, E_c, A_0 from Theorem 3.1, and define

$$U_{\mathrm{async}} = \frac{4\mathbb{E}\Delta_0}{m\beta_c T} + \frac{(6 + 8\tau_w)\beta_c \sigma_c^2}{m} + \frac{49E_c}{m^2}, \qquad d_{\mathrm{async}} = \frac{12L_{w^\star}^2 \eta^2}{m^2 \beta_c^2} + \frac{98L_\theta^2 \eta^2 \tau_a^2}{m^2}.$$

Let $c_a = 1 + \mathbf{1}\{\tau_a > 0\}$ and $c_w = 1 + \mathbf{1}\{\tau_a + \tau_w > 0\}$, and put

$$A_{\mathrm{async}} = c_a c_w (\tau_w + 1)a, \quad F_{\mathrm{async}} = c_a F, \quad q_{\mathrm{async}} = 2c_a a L_{w^\star}^2 \eta^2 (\tau_a + \tau_w)^2 + 2L_J^2 \eta^2 \tau_a^2,$$
$$h_{\mathrm{async}} = A_{\mathrm{async}} d_{\mathrm{async}} + q_{\mathrm{async}}, \qquad \rho_{\mathrm{async}} = 10 h_{\mathrm{async}}.$$

If $\rho_{\mathrm{async}} < 1$*, then*

$$\frac{1}{T}\sum\nolimits_{t<T}\mathbb{E}\|\nabla J(\theta_t)\|^2 \le \frac{A_0 + 4F_{\mathrm{async}} + 4A_{\mathrm{async}}U_{\mathrm{async}} + 4h_{\mathrm{async}}\sigma_a^2}{1-\rho_{\mathrm{async}}}, \tag{K.5}$$

$$\frac{1}{T}\sum\nolimits_{t<T}\mathbb{E}\Delta_t \le \frac{U_{\mathrm{async}} + d_{\mathrm{async}}(\sigma_a^2 + 2A_0 + 10F_{\mathrm{async}})}{1-\rho_{\mathrm{async}}}. \tag{K.6}$$

Relation to the synchronous bound. The asynchronous bound contains the same optimization, actor and critic noise, finite-batch, trajectory-shift, and clipping terms, along with terms for delayed critic updates and actor evaluation. At zero delay, $A_{\mathrm{async}} = a$, $F_{\mathrm{async}} = F$, and $q_{\mathrm{async}} = 0$, yielding the synchronous form. Theorem 3.1 gives tighter critic constants in this zero-delay case.

Asynchronous stability conditions. Two distinct sufficient restrictions are required: the delay-dependent critic stepsize bound and $\rho_{\mathrm{async}} < 1$. Even with a fixed actor, delayed critic gradients can oscillate if the critic stepsize is too large. Thus increasing the critic stepsize is useful only within its delay-dependent admissible range. The theorem concerns one gradient pair per server job and a nonanticipative sampling schedule; accumulated local parameter deltas and data-dependent arrival selection require additional analysis.

K.3 Auxiliary bounds for asynchronous convergence

K.3.1 A key auxiliary bound: server drift under bounded staleness

We first bound the distance between a current iterate and a stale snapshot. The averaged bound permits snapshot indices that depend on the observed history.

Lemma K.2 (Parameter-server staleness drift)**.** *Suppose the server actor recursion is*

$$\theta_{t+1} = \theta_t + \Delta\theta_t, \qquad \Delta\theta_t = \eta\widetilde{g}_t. \tag{K.7}$$

For deterministic $0 \le s \le t$*,*

$$\mathbb{E}\|\theta_t - \theta_s\|^2 \le (t-s)\sum_{j=s}^{t-1}\mathbb{E}\|\Delta\theta_j\|^2. \tag{K.8}$$

If $s_t \in \{\max(0, t-D), \dots, t\}$ *is a possibly random snapshot index, then*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\theta_t - \theta_{s_t}\|^2 \le \eta^2 D^2 \bar{S}, \qquad \bar{S} := \frac{1}{T}\sum_{t<T}\mathbb{E}\|\widetilde{g}_t\|^2. \tag{K.9}$$

The same averaged bound holds between any two indices in the preceding window of length D*. If* $\mathbb{E}\|\Delta\theta_t\|^2 \le \eta^2\bar{\sigma}_a^2$*, each such snapshot distance has expectation at most* $\eta^2 D^2 \bar{\sigma}_a^2$*.*

Proof. Telescoping and Cauchy–Schwarz give $\|\theta_t - \theta_s\|^2 \le (t-s)\sum_{j=s}^{t-1}\|\Delta\theta_j\|^2$. For a random snapshot, bound this pathwise by $D\sum_{j=\max(0,t-D)}^{t-1}\|\Delta\theta_j\|^2$. On summing over $t < T$, each increment occurs in at most D windows. The same argument applies to the subinterval joining two indices in a window. Taking expectations proves the assertions without an independence assumption on the indices. □

K.3.2 Asynchronous control of clip activation in the parameter-server model

The population clipping bound compares the evaluated stale actor with the worker's data-generating policy. The stored batch may have been collected before the actor snapshot was taken.

Lemma K.3 (PS-Async KL control of the clipping region)**.** *Fix behavior policy* p*, evaluated policy* q*, and* $\delta \in (0,1)$*, with common support as in Lemma C.4. If*

$$\mathbb{E}_{s\sim d^p}\mathrm{KL}(p(\cdot \mid s)\|q(\cdot \mid s)) \le \kappa, \tag{K.10}$$

then, for $(s,a) \sim d^p(s)p(a \mid s)$ *and* $r = q(a \mid s)/p(a \mid s)$*,*

$$\Pr(|r-1| \ge \delta) \le \frac{\sqrt{2\kappa}}{\delta}. \tag{K.11}$$

For bounded random advantage labels, the corresponding population squared gradient distortion is at most $2G_\pi^2 A_{\max}^2 \kappa(1 + 1/\delta)^2$*.*

Proof. Markov's inequality gives $\Pr(|r-1| \geq \delta) \leq \mathbb{E}|r-1|/\delta$. The identity $\mathbb{E}_{a\sim p}|r-1| = 2\mathrm{TV}(p,q)$, Pinsker's inequality, and Jensen's inequality give $\mathbb{E}|r-1| \leq \sqrt{2\kappa}$. The distortion claim follows from Corollary C.1, equation C.26. The clipped derivative is suppressed only on the appropriate advantage-sign branch, which is contained in the displayed ratio event. These population bounds are combined with a separate bound for empirical batch reuse below. □

K.3.3 Open-loop critic tracking with delayed gradients

Lemma K.4 (Delayed critic tracking with predictable bias)**.** *Let* $\theta_{t+1} = \theta_t + \eta\widetilde{g}_t$ *and*

$$w_{t+1} = w_t - \beta_c h_t, \qquad h_t = \nabla_w \mathcal{L}_c(\theta_{t-\tau_t}, w_{t-\nu_t}) + u_t + \xi_t,$$

where $0 \leq \tau_t \leq \min(t, \tau_a)$ *and* $0 \leq \nu_t \leq \min(t, \tau_w)$*. The iterates, delays, and* u_t *are* $\mathcal{F}_t$*-measurable, and*

$$\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0, \quad \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \leq \sigma_c^2, \quad \frac{1}{T}\sum_{t<T}\mathbb{E}\|u_t\|^2 \leq E_c.$$

Assume the convexity, quadratic-growth, stationary selected-minimizer, and L_c*-smoothness conditions of Assumption 4 hold on the relevant visited region. In addition, suppose*

$$\|\nabla_w \mathcal{L}_c(\theta, w) - \nabla_w \mathcal{L}_c(\theta', w)\| \leq L_\theta \|\theta - \theta'\|$$

for the mixed parameter pairs used in the proof. Both the L_c *and* L_θ *bounds hold on all such mixed current/stale pairs and connecting segments; a common Cartesian parameter region suffices. Put* $m := \min\{\mu_q/2, L_c\}$ *and assume* $0 < \beta_c \leq m/[24L_c^2(\tau_w+1)]$*. Then*

$$\bar{\Delta} \leq U_{\mathrm{async}} + d_{\mathrm{async}}\bar{S}, \tag{K.12}$$

where

$$U_{\mathrm{async}} := \frac{4\mathbb{E}\Delta_0}{m\beta_c T} + \frac{(6+8\tau_w)\beta_c\sigma_c^2}{m} + \frac{49E_c}{m^2}, \tag{K.13}$$

$$d_{\mathrm{async}} := \frac{12L_{w^\star}^2\eta^2}{m^2\beta_c^2} + \frac{98L_\theta^2\eta^2\tau_a^2}{m^2}, \tag{K.14}$$

and $\bar{\Delta} = T^{-1}\sum_{t<T}\mathbb{E}\|w_t - w^\star(\theta_t)\|^2$*. Actor and critic update noises may be correlated.*

Proof. Write $e_t = w_t - w^\star(\theta_t)$, $g_t^c = \nabla_w \mathcal{L}_c(\theta_t, w_t)$, and

$$q_t^c := \nabla_w \mathcal{L}_c(\theta_{t-\tau_t}, w_{t-\nu_t}) - g_t^c.$$

Convexity and quadratic growth imply $\langle e_t, g_t^c\rangle \geq m\Delta_t$; stationarity and gradient Lipschitzness imply $\|g_t^c\| \leq L_c\|e_t\|$. Moreover,

$$\|q_t^c\|^2 \leq 2L_c^2\|w_t - w_{t-\nu_t}\|^2 + 2L_\theta^2\|\theta_t - \theta_{t-\tau_t}\|^2. \tag{K.15}$$

Put $z_t = q_t^c + u_t$. Expanding around the current target, using conditional zero mean of ξ_t, and applying Young's inequality gives

$$\begin{aligned}\mathbb{E}[\|e_t - \beta_c h_t\|^2 \mid \mathcal{F}_t] \leq\ & \left(1 - \tfrac{3}{2}m\beta_c + 2\beta_c^2L_c^2\right)\Delta_t \\ & + \left(\tfrac{2\beta_c}{m} + 2\beta_c^2\right)\|z_t\|^2 + \beta_c^2\sigma_c^2 \\ \leq\ & (1 - m\beta_c)\Delta_t + \frac{4\beta_c}{m}\|z_t\|^2 + \beta_c^2\sigma_c^2.\end{aligned}$$

The last line uses $\beta_c \leq m/(4L_c^2)$ and $m\beta_c \leq 1$. To move the target to $w^\star(\theta_{t+1})$, apply the pathwise inequality

$$\|x - y\|^2 \leq (1 + m\beta_c/2)\|x\|^2 + (1 + 2/(m\beta_c))\|y\|^2.$$

Since $\|w^\star(\theta_{t+1}) - w^\star(\theta_t)\|^2 \leq L_{w^\star}^2\eta^2\|\widetilde{g}_t\|^2$, this yields

$$\begin{aligned}\mathbb{E}\Delta_{t+1} \leq\ & (1 - m\beta_c/2)\mathbb{E}\Delta_t + \frac{12\beta_c}{m}\mathbb{E}\|q_t^c\|^2 + \frac{12\beta_c}{m}\mathbb{E}\|u_t\|^2 \\ & + \tfrac{3}{2}\beta_c^2\sigma_c^2 + \frac{3L_{w^\star}^2\eta^2}{m\beta_c}\mathbb{E}\|\widetilde{g}_t\|^2.\end{aligned} \tag{K.16}$$

Here $\|q_t^c+u_t\|^2 \le 2\|q_t^c\|^2+2\|u_t\|^2$ was used. The target-shift argument does not require independent actor and critic noise.

Let $\bar{W} := T^{-1}\sum_{t<T}\mathbb{E}\|w_t - w_{t-\nu_t}\|^2$ and $\bar{\Theta} := T^{-1}\sum_{t<T}\mathbb{E}\|\theta_t - \theta_{t-\tau_t}\|^2$. Summing equation K.16 and using equation K.15 gives

$$\bar{\Delta} \le \frac{2\mathbb{E}\Delta_0}{m\beta_c T} + \frac{48L_c^2}{m^2}\bar{W} + \frac{48L_\theta^2}{m^2}\bar{\Theta} + \frac{24E_c}{m^2} + \frac{3\beta_c\sigma_c^2}{m} + \frac{6L_{w^\star}^2\eta^2}{m^2\beta_c^2}\bar{S}. \tag{K.17}$$

Set $\bar{H}_c := T^{-1}\sum_{t<T}\mathbb{E}\|h_t\|^2$. The window argument in Lemma K.2, applied to both recursions, gives

$$\bar{W} \le \beta_c^2\tau_w^2\bar{H}_c, \qquad \bar{\Theta} \le \eta^2\tau_a^2\bar{S}.$$

Conditional variance decomposition and a three-term norm bound give

$$\bar{H}_c \le 3L_c^2\bar{\Delta} + 6L_c^2\bar{W} + 6L_\theta^2\bar{\Theta} + 3E_c + \sigma_c^2.$$

Since $6L_c^2\beta_c^2\tau_w^2 \le 1/2$, absorption implies

$$\bar{W} \le \beta_c^2\tau_w^2\left(6L_c^2\bar{\Delta} + 12L_\theta^2\bar{\Theta} + 6E_c + 2\sigma_c^2\right).$$

Substituting into equation K.17 leaves the coefficient $\chi_c := 288\beta_c^2\tau_w^2L_c^4/m^2 \le 1/2$ multiplying $\bar{\Delta}$ on its right. Absorbing again gives

$$\bar{\Delta} \le \frac{4\mathbb{E}\Delta_0}{m\beta_c T} + \left(\frac{48}{m^2} + \frac{576\beta_c^2\tau_w^2L_c^2}{m^2}\right)E_c + \left(\frac{6\beta_c}{m} + \frac{192\beta_c^2\tau_w^2L_c^2}{m^2}\right)\sigma_c^2 + \left(\frac{96L_\theta^2}{m^2} + \frac{1152\beta_c^2\tau_w^2L_c^2L_\theta^2}{m^2}\right)\bar{\Theta} + \frac{12L_{w^\star}^2\eta^2}{m^2\beta_c^2}\bar{S}.$$

Finally, $\beta_c\tau_wL_c^2/m \le 1/24$ and $m \le L_c$ bound the respective coefficients by $49/m^2$, $(6+8\tau_w)\beta_c/m$, and $98L_\theta^2/m^2$. Using $\bar{\Theta} \le \eta^2\tau_a^2\bar{S}$ proves the result. □

K.4 Closed-Loop Convergence of Asynchronous PPO-Clip (Parameter-Server Model)

We combine the finite-batch bias bound, delayed-gradient critic tracking, and actor stationarity. Each server job applies one actor gradient and one critic gradient. The result does not cover accumulated parameter changes from multiple worker updates.

Proof of Theorem K.1. At job t, write $s_t = t-\tau_t$, $r_t = t-\nu_t$, $\bar{\theta}_t = \theta_{s_t}$, and $\bar{w}_t = w_{r_t}$. The job evaluates the actor and critic at these stale parameters; its stored batch has the behavior policy specified in the theorem. The N batches each receive exactly K jobs under the deterministic batch-to-job mapping, so $T = NK$. The filtration contains the batches already drawn, current and stale parameters, and chosen job indices, before the fresh minibatch draw.

Step 0: averaged quantities. Define

$$\bar{G} := T^{-1}\sum_{t<T}\mathbb{E}\|\nabla J(\theta_t)\|^2, \tag{K.18}$$

$$\bar{\Delta} := T^{-1}\sum_{t<T}\mathbb{E}\Delta_t, \quad \Delta_t := \|w_t - w^\star(\theta_t)\|^2, \tag{K.19}$$

$$\bar{S} := T^{-1}\sum_{t<T}\mathbb{E}\|\widetilde{g}_t\|^2. \tag{K.20}$$

Let $b_t := \mathbb{E}[\widetilde{g}_t \mid \mathcal{F}_t] - \nabla J(\theta_t)$ and $e_t := \mathbb{E}[\widetilde{g}_t \mid \mathcal{F}_t] - \nabla J(\theta_{s_t})$. Then

$$b_t = e_t + \nabla J(\theta_{s_t}) - \nabla J(\theta_t), \qquad \bar{B} := T^{-1}\sum_{t<T}\mathbb{E}\|b_t\|^2. \tag{K.21}$$

Step 1: open-loop actor stationarity. Part I of the proof of Theorem D.1 uses only the current conditional mean bias, smoothness, and conditional variance, so it applies here. For $\eta \le 1/(4L_J)$,

$$\bar{G} \le A_0 + 4\bar{B}, \qquad A_0 := \frac{4(J_{\max} - J(\theta_0))}{\eta T} + 2L_J \eta \sigma_a^2. \tag{K.22}$$

Step 2: clipping and finite-batch reuse at the stale actor. Apply Lemma C.5 at (θ_{s_t}, w_{r_t}), relative to the job's behavior policy. Its path-distribution and clipping bounds use the population KL premise between that behavior policy and θ_{s_t}; they do not require the current server policy to equal either one. The empirical total-variation bound is uniform in the parameters, so it permits adaptive choices of stale snapshots. Exactly K jobs per batch ensure that the averaged squared empirical distribution error is bounded by the same ε_B as in the synchronous bridge. Thus, with a, F, E_c defined in the theorem,

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|e_t\|^2 \le a\bar{R} + F, \qquad \bar{R} := \frac{1}{T}\sum_{t<T} \mathbb{E}\|w_{r_t} - w^\star(\theta_{s_t})\|^2. \tag{K.23}$$

The critic conditional-mean bias relative to $\nabla_w \mathcal{L}_c(\theta_{s_t}, w_{r_t})$ has average squared norm at most E_c. The argument retains the reused batch's empirical error after conditioning on the full history.

Step 3: relate stale critic errors to current tracking errors. Put $c_a := 1 + \mathbf{1}\{\tau_a > 0\}$ and $c_w := 1 + \mathbf{1}\{\tau_a + \tau_w > 0\}$. Lipschitzness of the selected critic minimizer gives

$$\|w_{r_t} - w^\star(\theta_{s_t})\|^2 \le c_w \Delta_{r_t} + 2L_{w^\star}^2 \|\theta_{r_t} - \theta_{s_t}\|^2.$$

For each realized delay sequence, any index j can occur as r_t at most $\tau_w + 1$ times. Both r_t and s_t lie in the preceding window of length $\tau_a + \tau_w$. Lemma K.2 therefore gives

$$\bar{R} \le c_w(\tau_w + 1)\bar{\Delta} + 2L_{w^\star}^2 \eta^2 (\tau_a + \tau_w)^2 \bar{S}. \tag{K.24}$$

Likewise, equation K.21, L_J-smoothness, and the window bound imply

$$\bar{B} \le c_a \frac{1}{T}\sum_{t<T} \mathbb{E}\|e_t\|^2 + 2L_J^2 \eta^2 \tau_a^2 \bar{S}.$$

The indicator factors equal one when the corresponding delays vanish.

Step 4: the asynchronous actor bias inequality. Combining the preceding bounds gives

$$\bar{B} \le A_{\text{async}} \bar{\Delta} + q_{\text{async}} \bar{S} + F_{\text{async}}, \tag{K.25}$$

where

$$A_{\text{async}} = c_a c_w (\tau_w + 1) a, \qquad F_{\text{async}} = c_a F,$$
$$q_{\text{async}} = 2c_a a L_{w^\star}^2 \eta^2 (\tau_a + \tau_w)^2 + 2L_J^2 \eta^2 \tau_a^2.$$

In particular, equation K.22 yields

$$\bar{G} \le A_0 + 4A_{\text{async}} \bar{\Delta} + 4q_{\text{async}} \bar{S} + 4F_{\text{async}}. \tag{K.26}$$

Step 5: open-loop critic tracking with delayed gradients. The empirical critic bridge supplies E_c for the predictable bias u_t in Lemma K.4. Its centered minibatch noise has conditional mean zero and variance at most σ_c^2 by the theorem's model. Applying that lemma gives

$$\bar{\Delta} \le U_{\text{async}} + d_{\text{async}} \bar{S}. \tag{K.27}$$

The additional restriction on β_c controls stale critic gradients even when the actor is fixed. The predictable critic bias contributes the explicit $49E_c/m^2$ floor.

Step 6: close the actor update second moment. Conditional variance decomposition gives

$$\bar{S} \le \sigma_a^2 + 2\bar{G} + 2\bar{B} \le \sigma_a^2 + 2A_0 + 10\bar{B}. \tag{K.28}$$

Let $\chi := A_{\text{async}} d_{\text{async}} + q_{\text{async}}$. Equations equation K.25 and equation K.27 imply

$$\bar{B} \le A_{\text{async}} U_{\text{async}} + F_{\text{async}} + \chi \bar{S}.$$

Consequently, with $\rho_{\text{async}} = 10\chi < 1$,

$$\bar{S} \le \frac{\sigma_a^2 + 2A_0 + 10A_{\text{async}} U_{\text{async}} + 10F_{\text{async}}}{1 - \rho_{\text{async}}}. \tag{K.29}$$

Step 7: conclude critic tracking. Substitute equation K.29 into equation K.27:

$$\bar{\Delta} \le \frac{(1 - 10q_{\mathrm{async}})U_{\mathrm{async}} + d_{\mathrm{async}}(\sigma_a^2 + 2A_0 + 10F_{\mathrm{async}})}{1 - \rho_{\mathrm{async}}}.$$

Dropping the nonpositive term $-10q_{\mathrm{async}}U_{\mathrm{async}}$ proves equation K.6.

Step 8: conclude actor stationarity. Use $\bar{G} \le A_0 + 4A_{\mathrm{async}}U_{\mathrm{async}} + 4F_{\mathrm{async}} + 4\chi\bar{S}$ and substitute equation K.29. After cancellation,

$$\bar{G} \le \frac{(1 - 2\chi)A_0 + 4A_{\mathrm{async}}U_{\mathrm{async}} + 4F_{\mathrm{async}} + 4\chi\sigma_a^2}{1 - \rho_{\mathrm{async}}}.$$

Dropping $-2\chi A_0$ proves equation K.5. At zero delay, $c_a = c_w = 1$, $q_{\mathrm{async}} = 0$, and $A_{\mathrm{async}} = a$. The bound then has the synchronous form, with the constants supplied by the delayed-critic lemma. □

L Experimental Details

L.1 Purpose, environment, and update protocol

Figures 1a–1c examine three aspects of the synchronous analysis: joint actor stationarity and critic parameter tracking, population GAE bias cancellation, and empirical gradient discrepancy under adaptive batch reuse. The two-step runs use the $H = 2, \gamma = 1$ example of Appendix I, with actions $a_0, a_1 \in \{-1, +1\}$, terminal reward $R = (1 + a_0 a_1)/2$, and zero terminal value. The shared policy has $q(\theta) = \frac{1}{2}\tanh\theta$ and $\pi_\theta(a \mid s) = (1 + aq(\theta))/2$. The normalized actor objective and the unnormalized critic target are, respectively,

$$J(\theta) = \frac{1 + q(\theta)^2}{4}, \quad J'(\theta) = \frac{q(\theta)q'(\theta)}{2}, \qquad w^\star(\theta) = \left(\frac{1 + q(\theta)^2}{2}, \frac{1 + q(\theta)}{2}, \frac{1 - q(\theta)}{2}\right).$$

Every training run starts from $\theta_0 = 1$ and $w_0 = (1/2, 1/2, 1/2)$, with critic coordinates ordered as the initial state, positive-action state, and negative-action state.

At each epoch we collect B independent complete trajectories under the frozen behavior policy and retain their realized Monte Carlo returns. At each of exactly K joint updates, we draw m trajectories independently with replacement from that batch. Actor and critic share these draws and use the same pre-update parameters. Raw GAE is recomputed from the current critic using equation I.5; the actor uses the sign-dependent clipping selector, choosing the unclipped branch at ties, and the critic uses equation I.6. We use plain gradient updates without advantage normalization, projection, or early stopping. Here K counts actual joint updates, not full passes through the batch. The critic cube and the population KL bound follow from the example's analytic invariance and stepsize certificates.

The four trajectory types permit exact multinomial compression: sample their counts from $\mathrm{Multinomial}(B, P_{\theta})$, then sample each minibatch's counts from $\mathrm{Multinomial}(m, \widehat{P})$. Because paths of a given type have identical rewards and gradients, this is exactly the distribution of B independent rollouts followed by replacement sampling of trajectory indices. This representation reduces computational cost while preserving the sample counts: $T = NK$, $Q = NB$, and $HQ = 2Q$.

L.2 Joint training diagnostics: Figure 1a

We run 32 independent seeds, 10000–10031, with

$$T = 8192, \quad B = 256, \quad K = 8, \quad m = 32, \quad \eta = 0.01, \quad \beta_c = 0.1, \quad \lambda = 0.5, \quad \delta = 0.2.$$

These settings satisfy the instance-specific restrictions: $\beta_c \le 1/8$, $\eta \le 1/4$, $\rho = 37.5(\eta/\beta_c)^2 = 0.375$, and $\kappa = K^2\eta^2/3 \approx 0.0021333$. Each seed represents $Q = 262{,}144$ fresh trajectories and $HQ = 524{,}288$ transitions. For each realized run we evaluate every pre-update iterate exactly and form

$$\overline{G}_t = \frac{1}{t}\sum_{u=0}^{t-1} |J'(\theta_u)|^2, \qquad \overline{\Delta}_t = \frac{1}{t}\sum_{u=0}^{t-1} \|w_u - w^\star(\theta_u)\|^2.$$

The plot averages these quantities over seeds; the critic metric is Euclidean parameter error, not a visitation-weighted value error. Its units differ from actor stationarity. All iterates enter the averages even though the curve is drawn at a subset of update indices. At $T = 8192$, the mean actor statistic is 2.8413×10^{-4} (95% interval $[2.8345, 2.8482] \times 10^{-4}$), and the mean critic statistic is 1.9790×10^{-3} (interval $[1.9499, 2.0063] \times 10^{-3}$). The mean final pre-update episode return is 0.61881. The largest observed within-epoch population KL, including post-update parameters, is 4.08×10^{-7}, below the deterministic certificate. No active clipping occurs in these training runs; the next diagnostic deliberately introduces drift to examine the clipped population gradient. Here active clipping means $\widehat{A} > 0, r > 1 + \delta$ or $\widehat{A} < 0, r < 1 - \delta$. Ratios outside the interval with the opposite advantage sign retain the unclipped gradient.

Substituting these training settings into the complete bound with equation I.10–equation I.11 gives actor and critic upper bounds of 7.5065 and 35.1910. These exceed the elementary envelopes $1/432$ and 3, respectively, and are therefore vacuous here. The measured reductions do not establish tightness of the bound or verify the separate vanishing-stepsize rate. The nonvacuous analytic calibration in Appendix I uses a much larger budget than the experiments.

L.3 Population cancellation and residual errors: Figure 1b

To isolate the effect of GAE, we hold $\theta = 1$, $w = (1/2, 1/2, 1/2)$, and $\delta = 0.1$, and vary $\lambda \in \{0, 0.1, \ldots, 1\}$. In the on-policy condition, $\bar{\theta} = \theta$. We enumerate the four paths to evaluate the population raw-GAE actor direction exactly. Writing $e = w - w^\star(\theta)$, its signed bias is

$$b_{\mathrm{GAE}} = \frac{(1-\lambda)q'(\theta)}{4}(e_+ - e_-).$$

The solid population curve displays its absolute value, which equals the RMS error of this deterministic quantity. At $\lambda = 1$ it is zero, despite the inaccurate critic. For each λ, we also draw 1,024 independent batches of $B = 64$ trajectories, using seeds 20000–21023, and compute the empirical conditional mean actor direction on each batch. Its plotted RMS error is $\sqrt{1024^{-1}\sum_i |\widehat{P}_i g - J'(\theta)|^2}$. Common batch seeds across λ give paired comparisons; batches within each setting are independent. At $\lambda = 1$, the measured RMS is 0.008439 (95% interval $[0.008106, 0.008786]$), while exact enumeration of $b_{\mathrm{GAE}}^2 + \mathrm{Var}_{P_\theta}(g)/B$ gives an RMS of 0.008704. Thus cancellation of the population critic contribution does not remove sampling error.

The dashed curve instead uses $\bar{\theta} = 0.5$ at the same current actor and critic, and evaluates the clipped population gradient under $P_{\bar{\theta}}$ by exact enumeration. Clipping is active at this parameter setting. Its absolute bias remains 0.08035 at $\lambda = 1$; the cancellation identity does not remove trajectory-distribution or clipping effects. This comparison isolates those residuals at fixed parameters; it does not use an iterate from the training runs above.

L.4 Adaptive batch discrepancies: Figure 1c

For each $B \in \{16, 32, 64, 128, 256, 512, 1024\}$, we draw 256 independent batches using seeds 30000–30255 and run a single epoch from the common initialization. The update parameters match the joint training configuration, with $K = 8$ in the main figure. Let $\phi_{\bar{\theta},\theta,w}$ denote the per-trajectory clipped actor direction and ψ_w the Monte Carlo critic gradient. At every pre-update iterate, we compute

$$z_k^a = \left|(\widehat{P} - P_{\bar{\theta}})\phi_{\bar{\theta},\theta_k,w_k}\right|^2, \qquad z_k^c = \left\|(\widehat{P} - P_{\bar{\theta}})\psi_{w_k}\right\|^2.$$

The plotted values are the means, over independent batches and minibatch draws, of $\max_{0\le k<K} z_k^a$ and $\max_{0\le k<K} z_k^c$. Later iterates depend on the same empirical batch used in the discrepancy, so this diagnostic includes adaptive selection. The maxima concern visited iterates, not suprema over the entire actor and critic classes. The example's pathwise envelopes hold uniformly over those classes and yield expectation bounds $4/B$ for the actor and $1/B$ for the critic, drawn as dotted lines.

For $B = 16$, the measured means are 1.7915×10^{-4} (actor) and 0.030325 (critic); for $B = 1024$, they are 3.1169×10^{-6} and 0.00045487. Both decrease over the measured batch-size range. Every observed update sequence satisfies its batch-dependent pathwise envelope. Additional comparisons use $K = 1$ and $K = 32$. For a fixed seed and B, these runs have common prefixes; the maximum over more iterates is necessarily nondecreasing in K and is not evidence that additional reuse harms learning. Clipping is inactive in all these adaptive runs. The $H = 2$ study illustrates adaptive empirical error but does not distinguish complete-path counting from the structured complexity bound at growing horizons.

L.5 Statistical analysis and reproducibility

For the two-step study, all shaded bands and error bars are pointwise 95% percentile bootstrap intervals with 2,000 resamples and bootstrap seed 20260925. We resample independent training seeds or independent batch replicates, never correlated inner iterates. For the RMS statistic, each resample averages squared errors before taking a square root. Population curves and analytic envelopes are deterministic and have no sampling intervals. All experimental settings, including the $K = 8$ condition shown in the figure, were specified in advance. The analysis includes all 32 training seeds, 1,024 replicates per λ, and 256 replicates per (B, K) setting, with no exclusions or replacements.

At each pre-update iterate, we evaluate the objectives, gradients, population KL, and clipping frequencies exactly. To distinguish the sources of actor bias, we decompose the empirical conditional-

mean error as

$$\widehat{P}\phi - J'(\theta) = (\widehat{P} - P_{\bar{\theta}})\phi + P_{\bar{\theta}}(\phi - \phi^{\mathrm{unclip}}) + \big(P_{\bar{\theta}}\phi^{\mathrm{unclip}} - P_{\theta}\phi^{\mathrm{on}}\big) + \big(P_{\theta}\phi^{\mathrm{on}} - J'(\theta)\big),$$

where ϕ^{unclip} retains behavior-policy ratios and ϕ^{on} uses current-policy scores and the same raw GAE. The critic decomposition similarly separates empirical and old-policy errors. Both decompositions concern conditional means, so minibatch sampling noise is treated separately from these bias terms.

Code is available in the anonymous repository linked in the reproducibility statement. The two-step experiments ran on an ARM64 macOS CPU using Python 3.12.14, NumPy 2.5.3, and double-precision arithmetic; figures used Matplotlib 3.11.2. The experiments took 52.3 seconds, excluding figure generation. These controlled measurements do not test neural or LLM performance, asynchronous updates, a necessary threshold at $\rho = 1$, the $T^{-2/5}$ rate, or the polynomial horizon-resource exponents.

L.6 Critic–clipping interaction during training

We examine whether critic error changes clipping decisions and distortion during training, while actor updates move the critic's population target. This complements the two-step study, whose training updates have inactive clipping.

Environment and protocol. We use the previous-action chain in Appendix J with $H = 8$, $\gamma = 1$, and zero terminal value. There is one initial state and two states per subsequent layer, recording the preceding binary action, for 15 nonterminal states. Rewards are $r_0 = 0$ and $r_h = \mathbf{1}\{a_h = a_{h-1}\}$ for $h \geq 1$. Unlike the shared scalar actor in that appendix, the experimental actor has an independent parameter at each state:

$$\pi_\theta(+1 \mid s) = 0.1 + 0.8\,\mathrm{sigmoid}(\theta_s).$$

All actor logits and learned tabular critic coordinates start at zero. The actor maximizes $J(\theta) = \mathbb{E}[G_0]/H$; reported returns are $\mathbb{E}[G_0]$, without horizon normalization.

We cross three critic modes—slow ($\beta_c = 0.02$), fast ($\beta_c = 0.2$), and an exact-current-value oracle—with $K \in \{16, 64\}$ and clipping enabled or disabled. Each of the 12 configurations has 16 replicates of $T = 2048$ actual actor updates. The actor stepsize is $\eta = 0.25$, the clipping half-width is $\delta = 0.05$, and $\lambda = 0.5$. Each epoch collects $B = 256$ independent trajectories under its frozen initial policy. Each update samples $m = 64$ trajectories with replacement from that batch, sharing actor and critic indices. Raw GAE is recomputed with the current pre-update critic and detached in the actor derivative. The critic loss averages squared residuals over trajectories and time positions, using stored, unnormalized Monte Carlo returns. Both gradients are evaluated before either plain ascent/descent update is applied. There is no advantage normalization, entropy bonus, gradient or value clipping, adaptive optimizer, KL rejection, early stopping, or extra critic update. Clipping ties select the unclipped derivative.

The oracle uses $V^{\pi_{\theta_t}}$ to recompute the same raw GAE on the stored trajectories; it neither refreshes the data nor substitutes true advantages. All $2^8 = 256$ action paths are enumerated for exact population diagnostics. Training still uses sampled finite batches and minibatches.

Clipping and interaction diagnostics. Let $c \in \{0, 1\}$ indicate whether clipping is enabled, and let $g_{\mathrm{clip}}(w)$ and $g_{\mathrm{unclip}}(w)$ denote population directions with the same current actor, behavior trajectory law, ratios, and raw-GAE convention. The applied population direction is

$$g_c(w) = g_{\mathrm{unclip}}(w) + c\,[g_{\mathrm{clip}}(w) - g_{\mathrm{unclip}}(w)].$$

For a trajectory label $\widehat{A}_h(w)$ and action ratio r_h, define

$$\mathcal{C}_h(w) = \{\widehat{A}_h(w) > 0,\ r_h > 1 + \delta\} \cup \{\widehat{A}_h(w) < 0,\ r_h < 1 - \delta\}.$$

Applied clipping activity averages $c\mathbf{1}_{\mathcal{C}_h(w)}$ over behavior trajectories and time positions; sampled activity averages it over the actual minibatch. With $w_t^\star = V^{\pi_{\theta_t}}$, the interaction vector is

$$I_t = c\Big([g_{\mathrm{clip}}(w_t) - g_{\mathrm{unclip}}(w_t)] - [g_{\mathrm{clip}}(w_t^\star) - g_{\mathrm{unclip}}(w_t^\star)]\Big).$$

Table 2: Configuration means over 16 replicates. Critic MSE and interaction norm are averaged over all pre-update checkpoints; final return is measured after 2048 updates. Disabled-clipping interaction is zero by definition.

Critic	K	Clipping	Final return	Critic MSE	mean $\|I_t\|$
Slow	16	Enabled	6.03452	0.935176	0
Slow	16	Disabled	6.03452	0.935176	0
Slow	64	Enabled	6.01705	0.923023	0.00204463
Slow	64	Disabled	6.03721	0.972968	0
Fast	16	Enabled	6.02924	0.059509	0
Fast	16	Disabled	6.02924	0.059509	0
Fast	64	Enabled	5.99900	0.066098	0.00042109
Fast	64	Disabled	6.02917	0.067160	0
Oracle	16	Enabled	6.02953	0	0
Oracle	16	Disabled	6.02953	0	0
Oracle	64	Enabled	5.99662	0	0
Oracle	64	Disabled	6.02991	0	0

Mask disagreement averages $c\,|\mathbf{1}_{\mathcal{C}_h(w_t)} - \mathbf{1}_{\mathcal{C}_h(w_t^\star)}|$. These counterfactuals change only the critic at the same checkpoint. Disabled-clipping zeros follow from $c = 0$; they do not imply that the hypothetical event $\mathcal{C}_h$ is absent. Oracle interaction is zero by construction.

Target movement and error measures. We measure the reverse direction of coupling by

$$M_t = \|V^{\pi_{\theta_{t+1}}} - V^{\pi_{\theta_t}}\|_2^2,$$
$$E_t = \|w_{t+1} - V^{\pi_{\theta_{t+1}}}\|_2^2 - \|w_{t+1} - V^{\pi_{\theta_t}}\|_2^2.$$

Here E_t isolates target movement at a fixed updated critic; it is not the total one-step tracking-error change. Negative values mean target movement reduces that error. For the oracle, resetting the critic to the new exact value gives $E_t = -M_t$. Critic MSE divides squared Euclidean tracking error by 15 and is unweighted over states. Stationarity is $\|\nabla J(\theta_t)\|_2^2$.

Results. Table 2 reports all configurations. At $K = 64$ with clipping, the mean interaction norms are 0.00204463 for slow and 0.000421091 for fast critics. Their paired difference is 0.00162354, with 95% bootstrap interval $[0.00159658, 0.00165281]$. Mask disagreement is respectively 1.59182% and 0.136202%. Both learned critics have positive mean interaction in every replicate. At $K = 16$, clipping is inactive and paired clipped/unclipped traces coincide.

The performance ordering is mixed. At $K = 64$, the slow critic tracks less accurately but clips less (3.56532% versus 5.07150% population activity) and attains a slightly higher final return than the fast critic: a paired gain of 0.0180456 $[0.0148847, 0.0212563]$, about 0.3% of the fast return. Clipping reduces final return relative to its disabled control by 0.0201616, 0.0301689, and 0.0332968 for slow, fast, and oracle modes, respectively. The oracle has 5.11585% clipping activity despite zero interaction. These observations do not give a universal performance ranking of critic speeds.

Mean target movement is 2.01810×10^{-5}, 1.93706×10^{-5}, and 1.94415×10^{-5} for the three clipped $K = 64$ modes. The learned critics' mean signed effects are positive (0.0303849 and 0.00532343), but negative effects occur at 0.787354% and 12.4725% of checkpoints. The oracle's negative effect follows from its exact reset.

Statistics, resources, and scope. Time averages first use all 2048 pre-update checkpoints within each replicate. Intervals use 2000 percentile bootstrap resamples of the 16 replicate summaries (seed 20260926, NumPy PCG64). Contrasts pair replicate IDs within a fixed K; correlated updates are not independent bootstrap observations. Rollout and minibatch seeds are $50000 + 2j$ and $50001 + 2j$, $j = 0, \ldots, 15$, with common random streams across matched configurations. Intervals are marginal, and curve bands are pointwise. Curves use nonoverlapping 64-update bins.

Each replicate processes $Tm = 131072$ trajectories per learned component. The $K = 16$ and $K = 64$ settings collect respectively 32768 and 8192 fresh trajectories, or 262144 and 65536 transitions. Thus cross-K comparisons match updates but not fresh-data budgets. The oracle is a

privileged diagnostic control. Execution used CPU float64 with one PyTorch thread on an AMD EPYC 9B14 under Windows 11, Python 3.12.14, and PyTorch 2.10.0; the main run took 183.79 seconds. The modified actor and initialization require their own theoretical constants; Appendix J's certificate is not transferred to these settings. The study supports interacting mechanisms, without testing the convergence rate, coupling condition, bound tightness, horizon scaling, or generalization to neural PPO.

M DISCUSSION

This appendix discusses the roles and limitations of the assumptions in the finite-time analysis and the interpretation of the coupling index ρ.

M.1 DISCUSSION OF ASSUMPTIONS

The assumptions retain several features of PPO-Clip training—mini-batch reuse, clipping, KL control, and a learned critic—while permitting finite-time bounds with explicit dependence on hyperparameters. The geometric assumptions concern the visited parameter region and the connecting segments used in the proofs. KL control limits policy-distribution mismatch; it does not itself bound parameter distance or keep the iterates near initialization. The statistical bridge requires finite trajectory support or the structured tabular conditions of Corollary 3.2; positive action coverage and the conditional sampling protocol are separate premises. These restrictions make the result a finite-sample specialization, not a guarantee for arbitrary neural PPO implementations.

Smoothness of the actor objective. Assumption 1 imposes L_J-smoothness on the underlying reinforcement learning objective $J(\theta)$. This is the canonical regularity condition used in non-asymptotic analyses of stochastic gradient methods, enabling a one-step descent/ascent inequality and a telescoping argument. Smoothness is assumed for the *true objective* $J(\theta)$ rather than for the PPO-Clip surrogate, which is non-smooth by construction due to clipping. In our analysis, this non-smoothness is handled explicitly via event localization and clipping-activity bounds (Appendix C), allowing the smoothness of J to be used without approximating or smoothing the surrogate.

Variance control of stochastic actor updates. Assumption 2 bounds the conditional variance of the stochastic actor update direction. In the smoothness argument, this variance bound is combined with a small actor stepsize to absorb the conditional mean part of the update; in the closed-loop argument, it gives the variance–mean decomposition used to control the averaged update magnitude. Such bounds are standard in stochastic optimization and RL theory and formalize the requirement that minibatch gradient estimators do not exhibit heavy-tailed behavior on the visited region. In the analyzed protocol, bounded raw advantages, bounded scores, and the separate ratio envelope $r \le R_{\max}$ bound each actor contribution by $R_{\max} G_\pi A_{\max}$. The critic contribution is bounded by $Q_c = 2(V_{\max} + G_{\max}) G_V$. Thus $R_{\max}^2 G_\pi^2 A_{\max}^2$ and Q_c^2 suffice as conditional variance bounds σ_a^2 and σ_c^2 for single-sample draws. Uniform sampling with replacement then gives the conditional variance bounds in Lemma C.5. Average KL control alone does not imply a bounded likelihood ratio. Bounded neural-network logits alone also do not bound the parameter derivatives of those logits; the score bound remains an explicit premise.

Bounded advantages. Assumption 3 bounds the raw GAE labels used by the theorem. A finite horizon and bounded rewards and critic values provide such an envelope. Practical advantage normalization and clipping are different transformations; their effects on the actor mean are not included in the present raw-GAE result. Analytically, bounded advantages serve two essential roles: (i) they ensure that the PPO surrogate gradient has controlled magnitude, and (ii) they allow clipping-induced distortion to be quantified explicitly through the clip-activation event and its probability. Without bounded advantages, the clipped surrogate could still produce arbitrarily large gradients even when likelihood ratios are well controlled.

KL trust-region control. Assumption 6 is a population KL premise motivated by the empirical KL checks used in PPO practice. An empirical check or penalty alone does not establish this premise. Appendix H gives a sufficient population certificate and a predictable stopping rule under which

the finite-time bound remains valid. In our analysis, the KL budget plays a central dual role: (i) it controls within-epoch drift induced by multi-epoch reuse (placing PPO in a "weakly off-policy" regime), and (ii) it provides quantitative control over clipping activity by bounding the probability that likelihood ratios exit the clipping interval. These two controls determine how the KL budget enters the non-asymptotic PPO-Clip bounds.

Bounded return / objective. When a uniform bound $J(\theta) \le J_{\max}$ is invoked, it is not implied by smoothness. Rather, it is a standard RL assumption that follows from bounded rewards under a finite horizon, a condition that is typically enforced in RLHF through reward clipping or normalization. The corresponding return assumption requires a uniform bound on the region of parameter space visited by the algorithm.

Value realization and coverage. For the Monte Carlo loss, the conditional-variance decomposition shows that a realizable minimizer agrees with the true value on every state of positive visitation probability. The coverage premise in Section 2.4 makes these all relevant states; Appendix I verifies this identification in an explicit example.

Why local critic convexity and quadratic growth are used. Our analysis relies on a local convexity and quadratic-growth condition on the critic loss (Assumption 4). This condition is used to obtain the parameter-space tracking contraction needed by the actor–critic feedback argument. In PPO-Clip, the critic is not optimized against a fixed objective: as the policy parameters evolve, the population-optimal critic $w^\star(\theta)$ becomes a *moving target*. When $a > 0$, controlling the tracking error limits its contribution to the population actor-bias bound.

The quadratic-growth condition converts critic excess loss into distance to the selected critic optimum, while local convexity implies the one-point inequality

$$f_\theta(w) - f_\theta(w^\star(\theta)) \le \langle \nabla_w f_\theta(w),\, w - w^\star(\theta) \rangle, \qquad f_\theta(w) := \mathcal{L}_c(\theta, w).$$

Together with smoothness and the stated stepsize restriction, these properties imply that one critic gradient step contracts the expected tracking error up to conditional noise, the empirical-to-population critic bias, and actor-induced target drift. This is the precise stability property needed to close the finite-time actor–critic inequalities.

Without such a local contraction condition, a smooth nonconvex critic objective may not correct critic error quickly enough relative to the movement of $w^\star(\theta)$ induced by policy updates under constant stepsizes. Critic lag can then accumulate and propagate into the actor update through advantage estimation. A different analysis would require additional structure or a different error measure; the present parameter-distance proof does not establish tracking without its contraction premises.

Relaxing the critic-geometry assumption via loss-gap tracking. The local convexity and quadratic-growth assumption is not the only possible route to a closed-loop analysis. One can relax the parameter-space contraction requirement by redefining the critic tracking error in terms of the critic optimization loss rather than distance to the selected minimizer. Concretely, for $f_\theta(w) := \mathcal{L}_c(\theta, w)$, define

$$\Delta_t^{\mathrm{loss}} := f_{\theta_t}(w_t) - f_{\theta_t}(w^\star(\theta_t)).$$

For a fixed population objective, a Polyak–Łojasiewicz-type condition and smoothness give a function-value contraction when the critic step is decomposed into its population gradient, predictable bias, and centered noise. Reused data retain a bias floor as well as the noise floor. A changing objective requires a further drift bound; PL alone does not supply it. Rather than a direct contraction of $\|w_t - w^\star(\theta_t)\|^2$, this route seeks a recursion for Δ_t^{loss}.

To close the actor–critic loop under this alternative metric, one then needs a calibration step connecting critic loss gap to value-estimation error. For example, one may assume a local value-error bound of the form

$$\|V_{w_t} - V^{\pi_{\theta_t}}\|_\infty^2 \le C_{\mathrm{val}}\, \Delta_t^{\mathrm{loss}},$$

or an analogous $L_2(d^{\pi_{\theta_t}})$ bound together with a concentrability condition. Such a condition replaces the parameter-space Lipschitz bound $\|V_w - V_{w^\star(\theta)}\|_\infty \le L_V \|w - w^\star(\theta)\|$. The resulting expected actor-bias bound replaces $a\mathbb{E}\Delta_t$ by $4G_\pi^2 \widetilde{C}_{\lambda,H}^2 C_{\mathrm{val}} \mathbb{E}\Delta_t^{\mathrm{loss}}$, while retaining the finite-batch, trajectory-mismatch, and clipping terms in F.

The remaining requirement is control of the moving-objective drift. In parameter-distance tracking, policy drift enters through

$$\|w^\star(\theta_{t+1}) - w^\star(\theta_t)\| \le L_{w^\star}\|\theta_{t+1} - \theta_t\|.$$

For loss-gap tracking, one instead has to compare $f_{\theta_{t+1}}(w_{t+1}) - f_{\theta_{t+1}}(w^\star(\theta_{t+1}))$ with $f_{\theta_t}(w_{t+1}) - f_{\theta_t}(w^\star(\theta_t))$. This requires additional Lipschitz or smoothness assumptions on the dependence of the critic loss and its optimal value on θ, and it produces several extra drift terms proportional to policy movement, such as $\eta\|g_t\|$ or $\eta^2\|g_t\|^2$ after standard Young-type inequalities. Bounds on these drift terms would permit a corresponding closed-loop argument with a coupling condition adapted to the loss-gap metric. The calibration and drift estimates are substantive additional assumptions, not consequences of PL or of the main theorem. Appendix N develops this conditional route. The main result uses parameter-distance tracking and its explicit contraction premises.

Interpreting the condition in LLM PPO. Local convexity and quadratic growth are substantive restrictions on a neural critic. The following considerations motivate examining local geometry, but do not establish these assumptions for an LLM:

- **Policy-distribution locality.** The population KL budget controls mismatch from each batch's behavior policy. It does not prevent cumulative parameter drift across batches or imply critic convexity. The required critic region and connecting segments must be checked separately.
- **Post-training initialization.** PPO is typically applied during post-training, starting from a supervised fine-tuning (SFT) checkpoint. This motivates studying the region reached from that checkpoint, but an SFT initialization does not guarantee high reward, local convexity, or quadratic growth of the subsequent critic loss.
- **Over-parameterization and regression structure.** Modern LLM critics operate in a highly over-parameterized regime and are trained via squared-error regression. Linearized regression models can motivate restricted geometric assumptions, but width and squared-error loss alone do not imply the selected-minimizer quadratic-growth condition. Parameter redundancies and insufficient state coverage can violate it.

Finite-batch scope and structured specialization. In the autoregressive setting, a finite prompt distribution, finite actions and horizon, and a deterministic reward model give finite trajectory support. The baseline cardinality-dependent bound can nevertheless scale exponentially with H. The structured specialization in Corollary 3.2 instead controls the clipped-gradient class through the numbers of critic coordinates and time–state–action cells. Its envelope is required on the entire critic cube and its clipping-tie rule selects the unclipped derivative. Appendices I and J verify the premises in explicit examples, including a growing-horizon family with polynomial constants. This remains a structured guarantee; validity for a particular neural PPO implementation requires separate verification.

M.2 Stability lens: the coupling index ρ

When $aL_{w^\star}^2 > 0$, the sufficient synchronous condition $\rho < 1$ limits the actor-to-critic stepsize ratio, with a threshold determined by the problem constants. When $a = 0$, including $\lambda = 1$ or $H = 1$ with matched terminal values, $\rho = 0$ imposes no time-scale restriction; critic tracking still depends on U and d.

Where ρ comes from. At a high level, the open-loop actor bound takes the form

$$\bar{G} \le A_0 + 4F + 4a\bar{\Delta},$$

while the open-loop critic tracking bound takes the form

$$\bar{\Delta} \le U + d\bar{S}, \qquad d = \frac{40L_{w^\star}^2\eta^2}{\mu_q^2\beta_c^2}.$$

Closing the loop requires controlling $\bar{S}$ (the average squared actor update magnitude) in terms of $\bar{G}$ and the actor bias, and the actor bias itself depends on $\bar{\Delta}$. Here U includes the critic finite-batch and trajectory-mismatch floor $40E_c/\mu_q^2$. After substitution, $\bar{\Delta}$ appears on both sides with coefficient $\rho = 10ad$. Thus $\rho < 1$ ensures the coupled inequalities can be solved with a finite amplification factor.

Interpretation. When $\rho \ll 1$, feedback produces a small amplification factor in these inequalities; the numerical guarantee also depends on the sizes of the other constants and error floors. As $\rho \to 1^-$, the factor $(1-\rho)^{-1}$ becomes large. This deterioration of a sufficient upper bound does not identify an actual stability transition in the algorithm. The divergence of the bound as $\rho \to 1^-$ should be interpreted as a loss of *uniform* finite-time control, rather than immediate divergence of the iterates. When $\rho \geq 1$, the scalar closure does not provide the stated guarantee. The theorem makes no conclusion about convergence or divergence in that regime, or about which implementation changes would restore stability.

Practical implication. For fixed regularity constants, reducing η/β_c reduces the synchronous coupling coefficient. Increasing β_c is subject to its contraction restriction and raises the critic noise floor; in the asynchronous theorem its permitted size also decreases with critic delay. Tightening κ reduces the explicit population mismatch floors, while increasing B reduces the finite-batch floor. These are implications for the proved bound, not necessary tuning rules or guarantees for alternative update schedules.

N Extensions and Alternative Critic Error Measures

This appendix examines the critic assumptions underlying the parameter-distance tracking error

$$\Delta_t := \|w_t - w^\star(\theta_t)\|^2$$

and considers two extensions: tracking the critic loss gap under additional calibration and drift assumptions, and including critic approximation error in the closed-loop bounds.

N.1 Why the strongest assumptions concern the critic landscape

Among the assumptions used in the paper, the strongest ones are the local regularity assumptions imposed on the critic landscape. In particular, Assumption 4 requires that, on the visited critic region and for each fixed policy parameter θ, the critic loss

$$f_\theta(w) := \mathcal{L}_c(\theta, w)$$

is locally convex, has Lipschitz-continuous gradients in w, satisfies quadratic growth around a selected minimizer $w^\star(\theta)$, and has a selected minimizer map $w^\star(\theta)$ that is Lipschitz in θ. These conditions are stronger than the actor smoothness, bounded-variance, bounded-advantage, and KL-control assumptions, because they impose geometric structure on a potentially nonconvex learned value-function objective.

This critic geometry yields a moving-target contraction for the parameter-distance tracking error $\Delta_t = \|w_t - w^\star(\theta_t)\|^2$:

$$\mathbb{E}\Delta_{t+1} \leq (1 - \mu_q\beta_c/4)\mathbb{E}\Delta_t + \tfrac{5}{4}\beta_c^2\sigma_c^2 + \frac{10\beta_c}{\mu_q}\mathbb{E}\|e_t^c\|^2 + \frac{10L_{w^\star}^2\eta^2}{\mu_q\beta_c}\mathbb{E}\|g_t\|^2,$$

where $e_t^c = \mathbb{E}[h_t \mid \mathcal{F}_t] - \nabla_w\mathcal{L}_c(\theta_t, w_t)$ is the predictable critic bias. Averaging this recursion and applying the finite-batch bridge gives

$$\bar{\Delta} \leq \frac{4\mathbb{E}\Delta_0}{\mu_q\beta_c T} + \frac{5\beta_c\sigma_c^2}{\mu_q} + \frac{40E_c}{\mu_q^2} + \frac{40L_{w^\star}^2\eta^2}{\mu_q^2\beta_c^2}\bar{S} = U + d\bar{S}.$$

When combined with the actor bias bound

$$\mathbb{E}\|b_t\|^2 \leq a\mathbb{E}\Delta_t + F,$$

this produces the closed-loop amplification factor

$$\rho = 10ad, \qquad a = 4G_\pi^2 L_V^2 \widetilde{C}_{\lambda,H}^2, \qquad d = \frac{40L_{w^\star}^2\eta^2}{\mu_q^2\beta_c^2}.$$

These parameter-space assumptions provide a sufficient route to controlling actor–critic coupling, the GAE/horizon factor $\widetilde{C}_{\lambda,H}$, KL-controlled off-policy drift, and clipping distortion.

An alternative is critic-loss tracking under a Polyak–Łojasiewicz (PL) condition. PL and smoothness control a frozen objective; extending this control to the coupled process also requires a moving-objective drift bound and calibration from loss gap to value error. These are additional premises beyond the main assumptions.

N.2 Tracking the critic loss gap instead of parameter distance

Define the critic optimal value and loss-gap tracking error by

$$f_\theta^\star := \inf_w f_\theta(w), \qquad \Delta_t^{\mathrm{loss}} := f_{\theta_t}(w_t) - f_{\theta_t}^\star.$$

This definition avoids requiring a locally unique or Lipschitz-selected minimizer in parameter space. Instead, the critic part of the analysis can be based on local smoothness and a PL inequality:

$$\frac{1}{2}\|\nabla_w f_\theta(w)\|^2 \geq \mu_{\mathrm{PL}}\big(f_\theta(w) - f_\theta^\star\big)$$

on the visited region. Let the critic update have conditional mean $\nabla_w f_{\theta_t}(w_t) + e_t^c$ and centered conditional variance at most σ_c^2. For $\beta_c \leq \min\{1/(4L_c), 1/\mu_{\mathrm{PL}}\}$, smoothness and Young's inequality give the frozen-objective bound

$$\mathbb{E}\big[f_{\theta_t}(w_{t+1}) - f_{\theta_t}^\star \mid \mathcal{F}_t\big] \leq (1 - \mu_{\mathrm{PL}}\beta_c)\Delta_t^{\mathrm{loss}} + \tfrac{5}{4}\beta_c\|e_t^c\|^2 + \tfrac{1}{2}L_c\beta_c^2\sigma_c^2,$$

provided smoothness holds along the critic step. Indeed the descent calculation leaves at most $-\frac{1}{2}\beta_c\|\nabla f_{\theta_t}(w_t)\|^2 + \frac{5}{4}\beta_c\|e_t^c\|^2$ before applying PL. The empirical-to-population bias does not vanish merely because the critic loss is PL.

The main difference is the moving-objective drift. In parameter-distance tracking, actor movement enters through

$$\|w^\star(\theta_{t+1}) - w^\star(\theta_t)\| \leq L_{w^\star}\|\theta_{t+1} - \theta_t\| = L_{w^\star}\eta\|g_t\|.$$

With loss-gap tracking, there is no need to compare minimizers directly. Instead one must compare two different critic objectives:

$$\Delta_{t+1}^{\mathrm{loss}} = f_{\theta_{t+1}}(w_{t+1}) - f_{\theta_{t+1}}^\star.$$

Insert and subtract the old objective:

$$\begin{aligned}\Delta_{t+1}^{\mathrm{loss}} = {} & \big(f_{\theta_t}(w_{t+1}) - f_{\theta_t}^\star\big) \\ & + \big(f_{\theta_{t+1}}(w_{t+1}) - f_{\theta_t}(w_{t+1})\big) + \big(f_{\theta_t}^\star - f_{\theta_{t+1}}^\star\big).\end{aligned}$$

The conditional expectation of the first term satisfies the biased PL bound above. The second and third terms are drift terms caused by policy movement. To control them, one needs regularity of the critic objective with respect to θ, for example local Lipschitz or smooth dependence of $f_\theta(w)$ and $f_\theta^\star$ on θ along the visited region. A typical bound would have the form

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}^{\mathrm{loss}} \mid \mathcal{F}_t] \leq {} & (1 - \mu_{\mathrm{PL}}\beta_c)\Delta_t^{\mathrm{loss}} + \tfrac{5}{4}\beta_c\|e_t^c\|^2 \\ & + \tfrac{1}{2}L_c\beta_c^2\sigma_c^2 + D_t^{\mathrm{move}},\end{aligned}$$

where D_t^{move} is a history-measurable upper bound on the conditional expectation of the sum of the two objective-change terms. A uniform Lipschitz bound gives a term proportional to $\eta\mathbb{E}[\|g_t\| \mid \mathcal{F}_t]$. A quadratic bound requires a separate estimate; it does not follow from PL alone. If such linear or quadratic movement control is established, averaging produces respectively a linear movement term of order

$$\frac{\eta}{\mu_{\mathrm{PL}}\beta_c}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|g_t\|$$

or, in the quadratic-drift case, a term comparable to

$$\frac{\eta^2}{\mu_{\mathrm{PL}}\beta_c}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|g_t\|^2.$$

The predictable critic bias additionally contributes at most $5E_c/(4\mu_{\mathrm{PL}})$ after averaging. Young's inequality can convert a linear movement term into squared-update terms plus a residual, but that residual and the resulting coupling condition must be retained.

On the actor side, parameter-distance tracking uses

$$\|V_w - V_{w^\star(\theta)}\|_\infty \le L_V \|w - w^\star(\theta)\|$$

to turn parameter tracking into an advantage-estimation error. If the critic is tracked by loss gap, this step must be replaced by a calibration assumption such as

$$\|V_{w_t} - V^{\pi_{\theta_t}}\|_\infty^2 \le C_{\mathrm{val}} \Delta_t^{\mathrm{loss}}$$

or an $L_2(d^{\pi_{\theta_t}})$ version plus a concentrability condition. The actor bias bound then becomes

$$\mathbb{E}\|b_t\|^2 \le a_{\mathrm{loss}} \mathbb{E}\Delta_t^{\mathrm{loss}} + F, \qquad a_{\mathrm{loss}} := 4G_\pi^2 \widetilde{C}_{\lambda,H}^2 C_{\mathrm{val}}.$$

If a separately proved drift estimate gives $\bar{\Delta}^{\mathrm{loss}} \le U_{\mathrm{loss}} + d_{\mathrm{loss}} \bar{S}$, the same scalar elimination as in the main proof uses

$$\rho_{\mathrm{loss}} = 10 a_{\mathrm{loss}} d_{\mathrm{loss}} < 1.$$

For example, the additional premise $D_t^{\mathrm{move}} \le C_{\mathrm{move}} \eta^2 \mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t]$ would give $d_{\mathrm{loss}} = C_{\mathrm{move}} \eta^2 / (\mu_{\mathrm{PL}} \beta_c)$. This quadratic drift estimate remains an additional premise. In particular, the β_c^{-2} dependence of the parameter-distance bound does not follow from PL alone. Linear movement control instead introduces the additional residuals discussed above.

Critic-loss tracking therefore offers a possible relaxation of local parameter contraction, conditional on the additional calibration and drift estimates. A closed-loop bound based on this metric would depend on calibration constants, objective-drift constants, and possibly additional movement floors. The parameter-distance result follows from the explicit contraction premises of the main theorem; the loss-gap route outlined here requires these further estimates.

N.3 Including critic approximation error

The main theorem assumes a realizable critic: the selected population critic satisfies

$$V_{w^\star(\theta)} = V^{\pi_\theta}$$

on the visited region. This removes irreducible value-approximation error, leaving tracking lag and the finite-batch, clipping, and trajectory-mismatch floors. The following extension allows an imperfect critic class while retaining the other geometry, coverage, sampling, terminal-boundary, and boundedness premises.

Let

$$\varepsilon_{\mathrm{app}}^2(\theta) := \|V_{w^\star(\theta)} - V^{\pi_\theta}\|_\infty^2, \qquad \bar{\varepsilon}_{\mathrm{app}}^2 := \sup_{\theta \in \mathcal{R}_\theta} \varepsilon_{\mathrm{app}}^2(\theta).$$

Then the value-estimation error decomposes as

$$\begin{aligned} \|V_{w_t} - V^{\pi_{\theta_t}}\|_\infty^2 &\le 2\|V_{w_t} - V_{w^\star(\theta_t)}\|_\infty^2 + 2\|V_{w^\star(\theta_t)} - V^{\pi_{\theta_t}}\|_\infty^2 \\ &\le 2L_V^2 \Delta_t + 2\bar{\varepsilon}_{\mathrm{app}}^2. \end{aligned}$$

Substituting this into the GAE term of the four-term actor-bias decomposition gives

$$\mathbb{E}\|b_t\|^2 \le a_{\mathrm{app}} \mathbb{E}\Delta_t + F_{\mathrm{app}}, \qquad a_{\mathrm{app}} := 2a,$$

$$F_{\mathrm{app}} := F + 8G_\pi^2 \widetilde{C}_{\lambda,H}^2 \bar{\varepsilon}_{\mathrm{app}}^2.$$

The factor eight combines the factor two in the value-error decomposition and the factor four in the actor-bias decomposition. The critic recursion still tracks the selected population minimizer, so U, d, and E_c are unchanged. The increased tracking coefficient requires the updated condition $\rho_{\mathrm{app}} := 10 a_{\mathrm{app}} d < 1$. Scalar elimination then gives

$$\bar{G} \le \frac{A_0 + 4F_{\mathrm{app}} + 4a_{\mathrm{app}} U + 4a_{\mathrm{app}} d\sigma_a^2}{1 - \rho_{\mathrm{app}}},$$

$$\bar{\Delta} \le \frac{U + d(\sigma_a^2 + 2A_0 + 10F_{\mathrm{app}})}{1 - \rho_{\mathrm{app}}}.$$

Thus the approximation floor also enters critic tracking through the actor update magnitude. This extension retains the learned-critic dynamics while adding irreducible value-approximation error to the closed-loop bounds.

An open-loop alternative assumes directly that the critic used by the actor satisfies

$$\|V_{w_t} - V^{\pi_{\theta_t}}\|_\infty^2 \le \varepsilon_{\mathrm{crit}}^2 \qquad \text{for all } t.$$

Then the finite-batch actor proof gives $\bar{G} \le A_0 + 4F + 16G_\pi^2 \widetilde{C}_{\lambda,H}^2 \varepsilon_{\mathrm{crit}}^2$, and requires no critic tracking theorem. This formulation treats critic quality as an exogenous oracle property. The closed-loop analysis instead quantifies its dependence on critic optimization, actor drift, and the actor–critic timescale ratio.