ADVG: A PROVABLY CONVERGENT ACTOR–CRITIC ALGORITHM FOR REINFORCEMENT LEARNING IN CONTINUOUS TIME AND SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning in continuous time and state-action space poses a core challenge for today's RL research. Deterministic actor–critic methods offer a model-free approach by estimating how action changes affect expected return, while avoiding action-space integration in the policy update. For controlled diffusions with unknown dynamics, however, estimating these effects becomes difficult as shorter observation intervals reduce the signal-to-noise ratio. We propose AdvG, a novel deterministic policy gradient (DPG) algorithm using multi-step temporal difference error for value estimation while learning the advantage rate's action derivative from individual exploratory perturbations. By analyzing critic stability, we establish the first finite-sample joint actor and critic convergence guarantee in model-free DPG learning for nonlinear controlled diffusions. With exact critic features, we establish an $O(T^{-2/5})$ joint convergence rate, with T the expected total duration of collected RL trajectories. Furthermore, our algorithm inspires the construction of an actor gradient estimator with mean-squared error of order $O(T^{-2/3})$, achieving the minimax optimality over a family of diffusion models allowing nonlinear action effects.

1 INTRODUCTION

Reinforcement learning studies how an agent can improve its decisions through interaction with an unknown environment. Deep reinforcement learning has become an important approach for controlling dynamic systems, with algorithms such as deep deterministic policy gradient (DDPG) and twin delayed DDPG (TD3) learning policies directly from interaction (Lillicrap et al., 2016; Fujimoto et al., 2018; Haarnoja et al., 2018). These algorithms are formulated around discrete decision steps. Yet robotic manipulators, autonomous vehicles, and many physical systems evolve continuously, and learners must choose how frequently to observe them and update their actions: applying a discrete-time algorithm at a finer resolution can lead to instability and degraded performance rather than better control (Tallec et al., 2019; Park et al., 2021). This motivates studying reinforcement learning directly in continuous time, with guarantees that account for the observation interval.

For stochastic systems, the difficulty goes beyond numerical approximation. Over a short interval, the effect of an action on expected return can be difficult to distinguish from random state fluctuations. A useful learning target is the advantage rate, which measures the effect of briefly changing an action per unit physical time. The martingale framework (Jia & Zhou, 2022a;b; Cheng et al., 2025) enables model-free learning of values, advantage rates, and stochastic policies from diffusion trajectories. In particular, Cheng et al. (2025) develop a continuous-time deterministic policy-gradient formula and propose CT-DDPG. Their multi-step temporal-difference (TD) analysis shows how a fixed physical observation window can prevent value-update variance from diverging as observations become more frequent.

These developments provide a foundation for continuous-time deterministic actor–critic learning. However, a tractable policy-gradient formula and a bounded-variance update do not by themselves guarantee that learning converges or uses data optimally. In this paper, we focus on two questions: Can a continuous-time deterministic actor and its critic be proved to converge when trained together?

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.

What is the best policy-gradient accuracy achievable from a fixed amount of physical interaction, and can a model-free method attain it?

The convergence challenge. In a deterministic actor–critic algorithm, the actor improves its policy using a critic that estimates expected return and the effect of changing the action. An inaccurate critic can send the actor in the wrong direction, while each policy update changes the critic's target and the distribution of subsequent observations. A joint convergence analysis in single timescale must therefore explain when these errors can be controlled. This coupling has been studied for deterministic policy gradients and single-timescale actor–critic methods (Xiong et al. (2022); Zhao et al. (2025b)). Meanwhile, continuous-time theory provides complementary guarantees for policy evaluation (Kobeissi & Bach (2023)), idealized actor–critic flows (Zhou & Lu (2024)), policy optimization(Zhao et al. (2023)), and exploratory or fitted q-learning (Mou (2026)). Most recently, Optimal-PhiBE uses continuous-time PDE approximation from discrete transition data for model-free policy iteration (Zhu et al., 2025). However, a finite-sample theory of joint deterministic actor–critic learning for unknown nonlinear diffusions is still missing. Establishing joint convergence will give deterministic actor–critic learning, a central paradigm in continuous-control RL, a rigorous foundation for learning from finite data in unknown nonlinear stochastic systems.

The statistical-efficiency challenge. A convergence bound describes what a particular algorithm achieves, but does not establish whether another algorithm could use the same data more effectively. Minimax analysis addresses this question by identifying the smallest worst-case estimation error achievable by any learning procedure over a specified model class. The central difficulty is a bias–variance tradeoff: small action perturbations reveal the local response needed for policy improvement but are harder to distinguish from diffusion noise, whereas larger perturbations introduce bias from nonlinear action effects. Related efficiency questions have been studied for off-policy gradients in discrete time (Kallus & Uehara (2020a)) and for derivatives estimated from noisy function evaluations (Lam et al. (2020)). However, a sharp characterization of policy-gradient accuracy under a physical interaction budget is still missing for unknown nonlinear diffusions. In this setting, the gradient must be recovered from unknown diffusion dynamics, with critic learning and state sampling both charged to the same trajectory budget. An attainable limit on policy-gradient accuracy will set a principled benchmark for continuous-time RL, directing algorithm design toward the best use of limited physical interaction.

Our approach and contributions. We introduce AdvG, a model-free deterministic actor–critic algorithm for nonlinear controlled diffusions. The key idea is to use temporal information differently for value learning and action-derivative learning. Like CT-DDPG, AdvG estimates values using TD errors accumulated over a fixed physical window. For action-derivative learning, AdvG pairs each exploratory action perturbation with the TD residual from the same observation interval. This design (see Figure 1) retains the variance-reduction benefit of multi-step value estimation while learning the advantage-function gradient needed for policy improvement. We establish three novel results.

Critic stability. Under regularity, state-moment, and feature-coverage assumptions, we derive stability of the joint value and action-derivative critic (Theorem 4.1). We further quantify how approximation, exploration, and discretization errors affect policy-gradient accuracy.

finite-sample joint actor–critic convergence. We present the first joint actor and critic convergence guarantee in model-free DPG learning for nonlinear controlled diffusions (Theorem 4.2). For a sufficiently expressive critic function class, the expected squared stationarity and tracking errors, averaged over iterations, decay at rate $O(T^{-2/5})$, with T the expected physical duration of all collected trajectories.

Minimax-optimal gradient estimation and policy optimization. At a fixed policy, we establish a minimax mean-squared error of $\Theta(T^{-2/3})$ over a smooth diffusion class with known features that exactly represent the value and local action derivative (Theorem 5.1). A batch estimator based on AdvG's paired critic construction attains the upper bound, and the matching lower bound allows adaptive actions and even continuous observation of the state path. That is, neither more frequent observations nor a different estimator can improve the worst-case exponent over this class. Under an additional projected-gradient-dominance condition, a separate policy-improvement procedure also achieves the minimax $\Theta(T^{-2/3})$ expected return loss relative to the best policy in the prescribed family (Corollary 5.2).

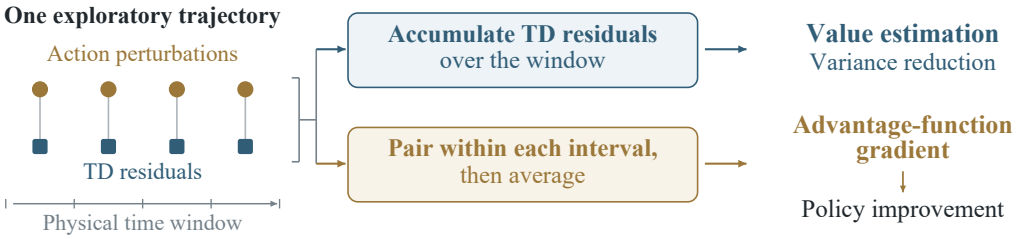


Figure 1: AdvG retains variance reduction through multi-step value estimation and learns the advantage-function gradient by pairing each action perturbation with the TD residual from the same observation interval.

2 Problem Formulation

Control objective. Throughout the paper we focus on continuous-time RL for controlled diffusions. Consider an agent that interacts with a stochastic system over an infinite horizon by choosing actions $A_t \in \mathbb{R}^{d_a}$ based on the observed state $X_t \in \mathbb{R}^{d_s}$, which evolves according to the SDE

$$dX_t = b(X_t, A_t)\,dt + \sigma(X_t, A_t)\,dW_t, \qquad X_0 \sim \nu_0, \tag{1}$$

with appropriate drift $b(\cdot,\cdot)$ and diffusion coefficients $\sigma(\cdot,\cdot)$, and W_t a standard d_w-dimensional Brownian motion in an appropriate probability space.

For policy $\mu : \mathbb{R}^{d_s} \to \mathbb{R}^{d_a}$, let X_t^μ be the solution to the SDE

$$dX_t^\mu = b(X_t^\mu, \mu(X_t^\mu))dt + \sigma(X_t^\mu, \mu(X_t^\mu))dW_t. \tag{2}$$

The goal is to maximize the expected accumulated discounted return over an infinite-time horizon and over appropriate admissible controls

$$\sup_\mu J(\mu), \qquad J(\mu) := \mathbb{E}_{\nu_0}\left[\int_0^\infty e^{-\beta t} r(X_t^\mu, \mu(X_t^\mu))\,dt\right], \tag{3}$$

with $r(x,a)$ the running reward, and $\beta > 0$ the discount rate. Under appropriate regularity conditions, it is well known (Fleming & Soner, 2006, Section III.9) that the optimal control lies in the class of stationary deterministic feedback policies.

In the RL setting where b, σ, and r are unknown, the agent must learn the optimal policy from the observed states and reward. Here we focus on admissible deterministic policies, whose state laws are well defined and nonexplosive and have finite expected reward. To facilitate the computation, we restrict ourselves to a parameterized family of admissible policies $\{\mu_\phi : \phi \in \Phi\}$, where $\Phi \subset \mathbb{R}^{d_\phi}$ is compact and convex. For example, μ_ϕ may be a neural network with parameters ϕ. The resulting optimization problem is

$$\max_{\phi\in\Phi} J(\phi), \qquad J(\phi) := J(\mu_\phi). \tag{4}$$

Below, we use X^ϕ to denote the corresponding state process under μ_ϕ.

Policy improvement. For a fixed ϕ, the value function is defined by

$$V^\phi(x) = \mathbb{E}_x\left[\int_0^\infty e^{-\beta t} r(X_t^\phi, \mu_\phi(X_t^\phi))\,dt\right], \qquad J(\phi) = \mathbb{E}_{\nu_0}[V^\phi(X_0)], \tag{5}$$

where $\mathbb{E}_x$ means that the process starts at x. Analogous to the advantage function in the discrete-time RL, consider the *advantage rate function* $q^\phi(x,a)$ in the continuous-time RL as in (Cheng et al., 2025). We then have

$$\mathcal{L}^a v(x) = b(x,a)^\top \nabla v(x) + \tfrac{1}{2}\,\mathrm{Tr}\left[\sigma(x,a)\sigma(x,a)^\top \nabla^2 v(x)\right],$$
$$q^\phi(x,a) = r(x,a) + \mathcal{L}^a V^\phi(x) - \beta V^\phi(x).$$

Here $\mathcal{L}^a$ is the diffusion generator. From the Bellman equation, we have $q^\phi(x, \mu_\phi(x)) = 0$.

Let ρ_ϕ be the occupancy measure of state under policy μ_ϕ such that $\rho_\phi(B) = \mathbb{E}_{\nu_0}\left[\int_0^\infty e^{-\beta t}\mathbf{1}\{X_t^\phi \in B\}\,dt\right]$ for any measurable set B. Under the regularity conditions (Appendix A), the deterministic policy-gradient identity (Cheng et al., 2025) gives

$$\nabla J(\phi) = \int \partial_\phi \mu_\phi(x)^\top g^\phi(x)\,\rho_\phi(dx), \tag{6}$$

where $\partial_\phi \mu_\phi$ is the Jacobian for policy parameterization, and the "advantage-function gradient" $g^\phi(x) := \nabla_a q^\phi(x, \mu_\phi(x))$ identifies local directions where the change in the policy's action improves return. This identity favorably allows the actor to improve its policy using estimates of the local advantage derivative, without knowing the dynamics.

Linear critic parameterization. To approximate the true value function V^ϕ, the critic is parameterized as V_θ, where θ is a trainable parameter. At iteration k, we update θ_k so that V_{θ_k} tracks V^{ϕ_k} as the actor parameters ϕ_k change. Since calculating the policy gradient requires the first-order action derivative at $\mu_\phi(x)$, we approximate the advantage rate locally by a function linear in the action deviation $a - \mu_\phi(x)$. For known state features $f : \mathbb{R}^{d_s} \to \mathbb{R}^{p_V}$ and $C : \mathbb{R}^{d_s} \to \mathbb{R}^{p_Q \times d_a}$, we apply the parameterization

$$V_\theta(x) = \theta^\top f(x), \qquad q_{\omega,\phi}(x,a) = \omega^\top C(x)(a - \mu_\phi(x)), \qquad z = (\theta, \omega) \in \mathcal{Z}, \tag{7}$$

where $\mathcal{Z} \subset \mathbb{R}^{p_V + p_Q}$ is compact and convex. This choice enforces $q_{\omega,\phi}(x, \mu_\phi(x)) = 0$ and provides a closed-form representation of the action gradient, which is $C(x)^\top \omega$. Specifically, we say features are **exact** if the ground truth V^ϕ and g^ϕ lie in the parameterized families.

Remark 2.1. Our convergence proof strategy can be adapted to fixed smooth nonlinear action features, e.g.,

$$q_{\omega,\phi}(x,a) = \omega^\top[\psi(x,a) - \psi(x, \mu_\phi(x))],$$

which also preserves the zero-advantage condition at the policy action. Here the features ψ remain fixed, and the paired test uses $\partial_a \psi(x, \mu_\phi(x))$ in place of $C(x)$. The extension requires additional smoothness, uniform coverage of these derivative features, and the corresponding moment, approximation, and policy-dependence bounds. We focus on Equation (7) here for ease of exposition.

3 AdvG: A CELL-PAIRING ALGORITHM FOR ACTOR-CRITIC LEARNING

We propose AdvG, a deterministic policy-gradient (DPG) algorithm designed for continuous-time RL. In this section, we introduce its key design choices, including the cell-pairing idea that helps us learn the action derivative of $q(x,a)$ (see a demonstration in Figure 1). We present the complete algorithmic procedure in Algorithm 1.

Multi-step TD learning. We adapt the multi-step TD learning routine following Cheng et al. (2025). Assume that our training samples are trajectory fragments with fixed length δ, and we call such fragments "windows". Each window will contain L observation intervals (we call them *cells*) of length $h > 0$, and $\delta = Lh$. We write X_i, A_i, and $r_i = r(X_i, A_i)$ for the initial state, held action, and observed reward rate in cell i, and $\gamma = e^{-\beta h}$ for its discount. For critic coefficients $z = (\theta, \omega)$, define the cell and window residuals as

$$D_i(z) = V_\theta(X_i) - \gamma V_\theta(X_{i+1}) - h\{r_i - q_{\omega,\phi}(X_i, A_i)\}, \qquad D^V(z) = \sum_{i=0}^{L-1} \gamma^i D_i(z). \tag{8}$$

With the true value and advantage functions, D_i has conditional mean zero up to discretization error. In our algorithm, we use D^V over a fixed physical window δ for learning the value function. This design effectively addresses the difficulty of vanishing signal-to-noise ratio in diffusion processes as $h \to 0$, allowing us to retain the variance reduction effect as proposed in (Cheng et al., 2025).

Learning the local action derivative. We introduce a key algorithmic idea, namely the "cell-pairing" mechanism designed for learning the correct advantage-rate derivative. During cell i, we

Algorithm 1 AdvG: Advantage Gradient inspired DPG Algorithm

Require: Reset distribution ν_0, discount $\beta > 0$, deterministic actor μ_ϕ, and fixed critic features f, C. Critic and actor constraint sets $\mathcal{Z}, \Phi$; initial $z_0 = (\theta_0, \omega_0) \in \mathcal{Z}$, $\phi_0 \in \Phi$. Observation interval $h > 0$, cells per window $L \geq 1$, batch size B, iterations K, exploration scale $0 < \sigma_e \leq 1$, critic and actor steps $\eta, \alpha > 0$, advantage weight $\chi > 0$.
Π_S denotes Euclidean projection onto S. Set $\gamma = e^{-\beta h}$ and window duration $\delta = Lh$.

for $k = 0, \ldots, K-1$ **do**
  Keep $(\phi, \theta, \omega) = (\phi_k, \theta_k, \omega_k)$ fixed for all collection and calculations below.
  for $b = 1, \ldots, B$ **do**
    Sample initial state.
    Draw G independently with $\Pr(G = j) = (1-\gamma)\gamma^j, j \geq 0$. Starting from a fresh state sampled from ν_0, follow μ_ϕ for G steps of duration h. Set X_0 to the terminal state.
    Collect one training window.
    for $i = 0, \ldots, L-1$ **do**
      Independently draw $U_i \sim N(0, \sigma_e^2 I)$ and set $A_i = \mu_\phi(X_i) + U_i$.
      Observe reward rate $r_i = r(X_i, A_i)$, hold A_i for time h, and observe X_{i+1}.
      Compute the cell residual: $D_i \leftarrow \theta^\top \{f(X_i) - \gamma f(X_{i+1})\} - h r_i + h\omega^\top C(X_i) U_i$.
    end for
    Compute update directions.
    Value: $m_b^V \leftarrow \frac{f(X_0)}{\delta} \sum_{i=0}^{L-1} \gamma^i D_i$.
    Advantage derivative (cell pairing): $m_b^Q \leftarrow \frac{1}{\delta \sigma_e^2} \sum_{i=0}^{L-1} \gamma^i C(X_i) U_i D_i$.
    Actor: $H_b \leftarrow \frac{h}{1-\gamma} \partial_\phi \mu_\phi(X_0)^\top C(X_0)^\top \omega$.
  end for
  Parameter update.
  $\widehat{m}^V \leftarrow \frac{1}{B} \sum_{b=1}^B m_b^V, \quad \widehat{m}^Q \leftarrow \frac{1}{B} \sum_{b=1}^B m_b^Q, \quad \widehat{H} \leftarrow \frac{1}{B} \sum_{b=1}^B H_b.$
  $z_{k+1} \leftarrow \Pi_{\mathcal{Z}} \left(z_k - \eta \begin{bmatrix} \widehat{m}^V \\ \chi \widehat{m}^Q \end{bmatrix} \right), \quad \phi_{k+1} \leftarrow \Pi_\Phi(\phi_k + \alpha \widehat{H}).$
end for
return actor and critic iterates $\{(\phi_k, z_k)\}_{k=0}^K$.

apply action $A_i = \mu_\phi(X_i) + U_i$ with exploration noise $U_i \sim N(0, \sigma_e^2 I)$. Pairing each residual with its own variance-normalized perturbation gives the critic update

$$m_\phi^{\mathrm{p}}(z; \Xi) = \frac{1}{\delta} \begin{bmatrix} f(X_0) D^V(z) \\ \sum_{i=0}^{L-1} \gamma^i \sigma_e^{-2} C(X_i) U_i D_i(z) \end{bmatrix}, \qquad \bar{m}_\phi(z) = \mathbb{E} m_\phi^{\mathrm{p}}(z; \Xi). \tag{9}$$

Here Ξ is the observed window $\Xi := (X_0, A_0, r_0, X_1, A_1, r_1, \cdots, X_L)$. The two blocks learn value and the advantage-function gradient, respectively. For a fixed policy, AdvG seeks critic coefficients for which these two update directions have zero expectation. Since the residuals are affine in the critic coefficients, the corresponding population moment equation is

$$\bar{m}_\phi(z) = \mathbb{E}[m_\phi^{\mathrm{p}}(z; \Xi)] = A_\phi z - b_\phi = 0. \tag{10}$$

We denote its fitted root by z_ϕ^*. We will suppress the dependence of symbols in (10) on h, σ_e, and omit ϕ when the policy is fixed; when there's no need to distinguish between update rules, we write $m = m^{\mathrm{p}}$.

Martingale interpretation. The moment conditions (10) are motivated by the martingale property of the true value and advantage functions. According to Cheng et al. (2025), for a fixed actor, the process

$$M_t = e^{-\beta t} V^\phi(X_t) + \int_0^t e^{-\beta s} \{r(X_s, A_s) - q^\phi(X_s, A_s)\} \, ds \tag{11}$$

is a martingale along exploratory trajectories. Its increments are therefore orthogonal to square-integrable test weights known at their start. With the true value and advantage, $\gamma^i D_i$ discretizes

$-(M_{(i+1)h} - M_{ih})$. The two blocks in (9) test the window increment with $f(X_0)$ and each cell increment with $\sigma_e^{-2} C(X_i)U_i$, respectively. The weight $\sigma_e^{-2} C(X_i)U_i$ is known after drawing U_i, before observing X_{i+1}, so both are valid martingale tests. AdvG solves this finite collection of discretized orthogonality conditions by stochastic approximation; Theorem 4.1 bounds the bias from discretization, inexact features, and the local linear advantage model.

The residual D_i compares the predicted improvement $hq_{\omega,\phi}(X_i, A_i)$ with $hr_i + \gamma V_\theta(X_{i+1}) - V_\theta(X_i)$, the observed reward plus the estimated value change. Since $\nabla_\omega q_{\omega,\phi}(X_i, A_i) = C(X_i)U_i$, AdvG divides U_i by σ_e^2 and multiplies this feature by its own D_i before summing. For comparison, CT-DDPG (Cheng et al., 2025) updates its advantage network using $\nabla_\omega \frac{1}{2}[D^V(z)]^2$. In our parametrization, this produces the product of sums illustrated in Figure 2, including terms that pair a parameter gradient from one observation interval with a residual from another.

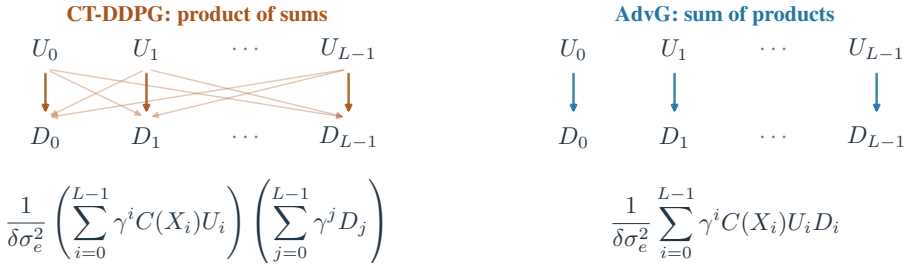


Figure 2: CT-DDPG multiplies sums of action features and residuals; AdvG sums matched products. Both methods retain multi-step value estimation.

Actor update direction. The critic supplies $C(x)^\top \omega$ to the policy-gradient formula. Let $d^0_{\phi,h}$ be the window-start law in (23), so $\frac{h}{1-\gamma} d^0_{\phi,h}$ approximates ρ_ϕ. The actor direction at state x is

$$H_{\phi,h}(z; x) = \frac{h}{1-\gamma} \partial_\phi \mu_\phi(x)^\top C(x)^\top \omega, \qquad \bar{H}_{\phi,h}(z) = \mathbb{E}_{X_0 \sim d^0_{\phi,h}} H_{\phi,h}(z; X_0). \tag{12}$$

Theorem 4.1 bounds the bias of $\bar{H}$ relative to $\nabla J(\phi)$.

Physical trajectory budget. Let T be the available budget for expected physical trajectory time, including the policy prefixes used to sample window-start states. For K iterations with B trajectories each, we require

$$\mathbb{E}[\text{total acquisition time}] = KB\Big(\underbrace{\frac{h}{e^{\beta h} - 1}}_{\text{policy prefix}} + \underbrace{\delta}_{\text{exploratory window}}\Big) \le T. \tag{13}$$

Additional Details. In the sampling procedure for Algorithm 1, we hold the selected action fixed during each step. We apply a weighting trick for critic updates: the matrix $W_\chi = \mathrm{diag}(I_{p_V}, \chi I_{p_Q})$ with sufficiently small $\chi > 0$ stabilizes the joint critic update without changing the root of our moment equation (10).

4 Convergence Analysis

The convergence analysis crucially depends on our algorithm design. Our zero-mean Gaussian perturbations prevent errors in the advantage estimate from affecting the expected value update, while cell-pairing gives a positive-definite matrix in the advantage equation under feature coverage. Choosing a sufficiently small weight χ for the advantage update relative to the value update ensures the stability of joint critic updates.

We now analyze the stability of the critic updates toward z^*_ϕ and the accuracy of the actor direction supplied by the fitted critic. Three types of conditions are imposed: regularity of dynamics, feature coverage, and function-class approximation precision, as stated precisely in Assumptions A.1–A.4.

In particular, let $e_V, e_Q \geq 0$ be uniform bounds on the errors in approximating V^ϕ and the local action derivative g^ϕ with the fixed critic features. Value error, including its first two state derivatives, is measured using a weight function $w(x) = 1 + \|x\|^2$ and its induced norm

$$\|v\|_{C^2_w} := \sup_x \left\{ \frac{|v(x)|}{w(x)} + \frac{\|\nabla v(x)\|}{\sqrt{w(x)}} + \|\nabla^2 v(x)\| \right\}.$$

Assumption (Critic approximation)**.** *For each policy μ_ϕ, we assume the existence of reference coefficients $\widetilde{z}_\phi = (\widetilde{\theta}_\phi, \widetilde{\omega}_\phi) \in \mathcal{Z}$, chosen independently of h, σ_e, which satisfy*

$$\begin{aligned} &\|V_{\widetilde{\theta}_\phi} - V^\phi\|_{C^2_w} \leq e_V, \\ &\sup_{0 \leq i < L} \left(\mathbb{E}\|C(X_i)^\top \widetilde{\omega}_\phi - g^\phi(X_i)\|^2\right)^{1/2} + \left(\int \|C(x)^\top \widetilde{\omega}_\phi - g^\phi(x)\|^2 \, \rho_\phi(dx)\right)^{1/2} \leq e_Q. \end{aligned} \tag{14}$$

We assume these bounds hold uniformly over $\phi \in \Phi$ and admissible h, σ_e, L, and the expectation is over the exploratory windows in Algorithm 1. When features are exact, we have $e_V = e_Q = 0$.

Theorem 4.1 (Critic stability)**.** *Under Assumptions A.1–A.4, fix a window duration $\delta = Lh$ and a sufficiently small critic weight $\chi > 0$. For sufficiently small h, σ_e, the moment equation (10) has a unique root z^*_ϕ. And there exists $\mu_\chi > 0$, independent of h, σ_e, such that*

$$\langle z - z', W_\chi\{\bar{m}(z) - \bar{m}(z')\}\rangle \geq \mu_\chi \|z - z'\|^2 \qquad \text{for all } z, z'. \tag{15}$$

The single-trajectory critic noise satisfies

$$\mathbb{E}\|W_\chi\{m(z;\Xi) - \bar{m}(z)\}\|^2 \leq \frac{C}{\delta\sigma_e^2}(1 + \|z - z^*_\phi\|^2).$$

The expected actor update direction $\bar{H}_{\phi,h}(z)$ approximates the true policy gradient with error

$$\|\bar{H}_{\phi,h}(z) - \nabla J(\phi)\| \leq C\|z - z^*_\phi\| + C\big[e_V + e_Q + \sigma_e^2 + h\big]. \tag{16}$$

(15) says that the mean critic update corrects coefficient errors in every direction. Together with the noise bound, it establishes stability as observations become finer while σ_e and δ remain fixed. Unlike earlier works of Ainsworth et al. (2021); Alacaoglu & Lyu (2023), rather than assuming the stability inequality, we derive it for the joint value and advantage-gradient critic from the diffusion dynamics and feature coverage, uniformly in h and allowing action-dependent diffusion. Bound (16) shows that fitting the critic brings its expected actor direction close to the true policy gradient, up to feature, exploration, and discretization errors. (See Corollaries A.6–A.7, Theorem B.1, and Corollary B.2 for detailed analysis.)

Simultaneous updates. For convergence analysis, we consider single-timescale actor critic updates, as commonly used in practice. We measure critic error by $\|z_k - z^*_{\phi_k}\|$ and policy progress by the projected gradient mapping $\mathcal{G}_\lambda(\phi) = \lambda^{-1}\{\Pi_\Phi(\phi + \lambda\nabla J(\phi)) - \phi\}$ for fixed $\lambda > 0$. This policy improvement step vanishes exactly at first-order stationary points of the constrained problem. If $\phi + \lambda\nabla J(\phi)$ lies in Φ, the mapping equals $\nabla J(\phi)$.

Theorem 4.2 (Joint actor and critic convergence)**.** *Under the conditions of Theorem 4.1, suppose the fitted critics z^*_ϕ belong to $\mathcal{Z}$. After K iterations with B fresh trajectories per iteration, with suitable learning rate η, α, Algorithm 1 yields*

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\mathcal{G}_\lambda(\phi_k)\|^2 + \|z_k - z^*_{\phi_k}\|^2\right] \\ &\quad \leq C\left\{\frac{1 + \|z_0 - z^*_{\phi_0}\|^2}{K\eta} + \frac{\eta}{B\delta\sigma_e^2} + \frac{1}{B} + \big[e_V + e_Q + \sigma_e^2 + h\big]^2\right\}. \end{aligned} \tag{17}$$

The constant C is independent of $h, \sigma_e, L, K, B, \eta$. Moreover, for sufficiently large physical budgets T, suitable hyperparameter configurations satisfy (13) give

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\mathcal{G}_\lambda(\phi_k)\|^2 + \|z_k - z^*_{\phi_k}\|^2\right] \leq CT^{-2/5} + C_{\text{approx}}(e_V, e_Q). \tag{18}$$

Here $C_{\text{approx}}(e_V, e_Q) \geq 0$ is approximation error, with $C_{\text{approx}}(0,0) = 0$ for exact features.

Theorem 4.2 gives the first finite-sample analysis of joint actor and critic convergence in model-free deterministic policy-gradient learning for nonlinear controlled diffusions, with general smooth nonlinear policies. Importantly, our result also allows inexact critic features and controls both errors throughout single-timescale learning. We highlight that bound (17) separates optimization error, sampling noise, and the effects of feature approximation, exploration, and discretization. Moreover, for exact features, (18) gives asymptotically vanishing iteration-averaged squared critic tracking error and projected policy-gradient norm. The proof is given in Corollary B.7.

Deterministic policy gradients underlie widely used continuous-control methods such as DDPG and TD3 (Silver et al., 2014; Lillicrap et al., 2016; Fujimoto et al., 2018), motivating research on their convergence guarantees that accommodate unknown stochastic dynamics. Compared with prior analyses of linear-quadratic models or diffusions with action-independent noise, Theorem 4.2 establishes an end-to-end guarantee for a broad class of unknown nonlinear dynamics with no prescribed parametric form. (See a detailed discussion on relevant research in Appendix F.)

5 Minimax optimality of policy gradient estimation

In this section, we show that the cell-pairing construction in AdvG yields a minimax-optimal policy-gradient estimator. For a fixed actor ϕ, we estimate $\nabla J(\phi)$ by solving the empirical AdvG critic equation and using the resulting critic in the policy-gradient formula.

Definition (Exact-feature diffusion class)**.** Fix the dimensions, discount factor $\beta > 0$, and known data $\mathcal{S} = (\nu_0, \mu, \phi, f, C)$. Let $\mathfrak{M}$ be the class of controlled diffusions with unknown drift, diffusion coefficient, and reward satisfying given regularity, moment, policy-evaluation, and feature-coverage conditions with common known bounds $\mathcal{K}$. The critic features exactly represent V^ϕ and g^ϕ in the sense of (28), using uniformly bounded coefficients independent of h, σ_e. All known data and bounds are fixed independently of T.

The detailed conditions to be satisfied are given in Definition D.1. Below, A ranges over all acquisition and estimation procedures with expected total trajectory duration at most T in every model, allowing adaptive actions and continuous path observations.

Theorem 5.1 (Minimax policy-gradient estimation)**.** *For any admissible exact-feature class $\mathfrak{M}$ defined above, let ϕ be the actor specified by $\mathcal{S}$ for estimation of $\nabla J(\phi)$. For all sufficiently large T and suitable parameter choices, there exists an AdvG-inspired estimator $\widehat{g}_T(\phi)$, with expected total trajectory duration at most T, satisfying*

$$\sup_{M \in \mathfrak{M}} \mathbb{E}_M \|\widehat{g}_T(\phi) - \nabla J_M(\phi)\|^2 \le C T^{-2/3}. \tag{19}$$

The constant C and the threshold on T depend only on the fixed class data.

There exists an admissible exact-feature class $\mathfrak{M}_\star$ with fixed data, evaluation actor $\phi_\star$, and a constant $c > 0$, independent of T, such that for all sufficiently large T,

$$c T^{-2/3} \le \inf_{\mathsf{A}} \sup_{M \in \mathfrak{M}_\star} \mathbb{E}_M \|\widehat{g}_{\mathsf{A}} - \nabla J_M(\phi_\star)\|^2. \tag{20}$$

Thus the paired-critic estimator attains the minimax rate on $\mathfrak{M}_\star$.

Algorithm 2 specifies this estimator $\widehat{g}_T(\phi)$; Corollaries D.4 and D.11 give the proofs.

Different from prior works of Kallus & Uehara (2020a) and Lam et al. (2020), our result establishes an attainable statistical limit with both critic fitting and state sampling procedures accounted for in the trajectory budget T. Also, the $2/3$ exponent is achieved from unknown SDE dynamics, a different setting from that in Lam et al. (2020).

Gradient-dominant diffusion class. Let Φ be the (known) compact convex set of actor parameters. Define $\mathfrak{C}_{\mathrm{gd}}$ as the class of controlled diffusion models that share the same known problem data other than the evaluation parameter ϕ, while satisfying the preceding diffusion-class definition with $e_V = e_Q = 0$ uniformly over $\phi \in \Phi$, with common bounds $\mathcal{K}$, and also satisfy the return regularity and gradient-dominance conditions stated below. The shared data, Φ, and all class constants are fixed independently of T.

For models in $\mathfrak{C}_{\mathrm{gd}}$, we require that J has a uniformly bounded range on Φ and an L_J-Lipschitz gradient on a neighborhood of Φ. In addition, assume that for known common $\lambda, \mu > 0$, the projected gradient controls the gap to the best return within the prescribed actor family:

$$\|\mathcal{G}_\lambda(\phi)\|^2 \geq 2\mu\{J^* - J(\phi)\}, \qquad J^* = \max_{\psi \in \Phi} J(\psi), \quad \phi \in \Phi. \tag{21}$$

This condition ensures that a small projected gradient mapping implies a small return gap, ruling out suboptimal stationary actors.

Corollary 5.2 (Minimax policy-return optimization)**.** *For every gradient-dominant diffusion family $\mathfrak{C}_{\mathrm{gd}}$, and all sufficiently large T, there exists an AdvG-inspired policy estimator $\hat{\phi}_T \in \Phi$, using expected trajectory time at most T, whose worst-case expected return loss is at most $CT^{-2/3}$. Moreover, there exists such a class $\mathfrak{C}_{\mathrm{gd},\star}$, with data and bounds fixed independently of T, for which*

$$cT^{-2/3} \leq \inf_{\mathsf{A}} \sup_{M \in \mathfrak{C}_{\mathrm{gd},\star}} \mathbb{E}_M[J_M^* - J_M(\widehat{\phi}_{\mathsf{A}})] \leq CT^{-2/3}, \tag{22}$$

for all sufficiently large T.

Algorithm 3 obtains $\hat{\phi}_T$ by repeatedly estimating policy gradients using Algorithm 2 and taking projected ascent steps. Theorems D.6 and D.12 give the proofs.

6 Numerical Experiments

We complement our theoretical analysis with experiments on coupled stochastic linear–quadratic control and a noisy pendulum. We compare AdvG with CT-DDPG, one-step DDPG, TD3, and SAC. Figure 3 shows effective policy learning by AdvG on both tasks at both observation intervals. AdvG retains competitive LQ performance under smaller discretization steps and substantially improves the initial policy on the noisy pendulum. In separate final-policy evaluations, AdvG achieves the highest mean discounted return among the five methods in all four settings. These results support the practical effectiveness of combining multi-step value learning with local action–residual pairing, and supports our algorithm's stability for the small-steplength setting. (See Appendix E for detailed experimental settings and final-policy comparisons.)

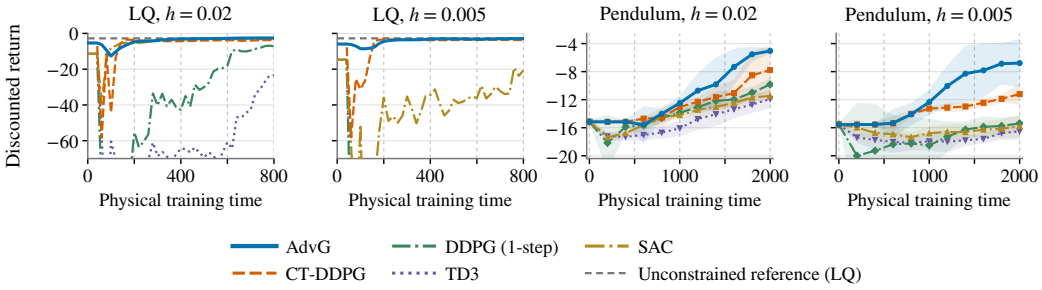


Figure 3: Discounted-return learning curves.

7 Conclusion

We developed AdvG, a deterministic actor–critic algorithm designed for model-free continuous time and space reinforcement learning. We introduce a novel method of estimating local action effects from matched action–residual pairs in DPG algorithms, which preserves the information needed for policy improvement. We establish the critic's stability as the observation interval shrinks and obtain the first finite-sample analysis of joint actor and critic convergence in model-free deterministic policy-gradient learning for RL in continuous time and space. Moreover, we prove a direct estimator based on the cell-pairing principle attains the minimax rate for fixed-policy gradient estimation.

STATEMENT ON AI USE

The authors used large language models for experiment execution, manuscript organization, exploration of proof strategies and calculations, and prose editing.

REFERENCES

Samuel Ainsworth, Kendall Lowrey, John Thickstun, Zaid Harchaoui, and Siddhartha Srinivasa. Faster policy learning with continuous-time gradients. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pp. 1054–1067, 2021. URL https://proceedings.mlr.press/v144/ainsworth21a.html.

Ahmet Alacaoglu and Hanbaek Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 458–489, 2023. URL https://proceedings.mlr.press/v202/alacaoglu23a.html.

Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34, 2022. URL https://jmlr.org/papers/v23/20-664.html.

Xuyang Chen and Lin Zhao. On the convergence of continuous single-timescale actor–critic. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9829–9859, 2025. URL https://proceedings.mlr.press/v267/chen25co.html.

Ziheng Cheng, Xin Guo, and Yufei Zhang. Deterministic policy gradient for reinforcement learning with continuous time and state, 2025. URL https://arxiv.org/abs/2509.23711.

Samuel N. Cohen, Jackson Hebner, Deqing Jiang, and Justin Sirignano. Neural actor-critic methods for Hamilton-Jacobi-Bellman PDEs: Asymptotic analysis and numerical studies, 2025. URL https://arxiv.org/abs/2507.06428. Version 2 revised May 2026.

Lorenzo Croissant, Marc Abeille, and Bruno Bouchard. Near-continuous time reinforcement learning for continuous state-action spaces. In *Proceedings of the 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pp. 444–498, 2024. URL https://proceedings.mlr.press/v237/croissant24a.html.

Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2 edition, 2006. doi: 10.1007/0-387-31071-1. URL https://doi.org/10.1007/0-387-31071-1.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596, 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Xuefeng Gao and Xun Yu Zhou. Square-root regret bounds for continuous-time episodic markov decision processes. *Mathematics of Operations Research*, 51(1):333–357, 2025. doi: 10.1287/moor.2022.0283.

Xin Guo, Anran Hu, and Yufei Zhang. Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls. *SIAM Journal on Control and Optimization*, 61(2): 755–787, 2023. doi: 10.1137/21M1414413. URL https://epubs.siam.org/doi/10.1137/21M1414413.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Yilie Huang, Yanwei Jia, and Xun Yu Zhou. Sublinear regret for a class of continuous-time linear–quadratic reinforcement learning problems, 2024. URL https://arxiv.org/abs/2407.17226.

Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55, 2022a. URL https://www.jmlr.org/papers/v23/21-0947.html.

Yanwei Jia and Xun Yu Zhou. Policy gradient and actor–critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50, 2022b. URL https://www.jmlr.org/papers/v23/21-1387.html.

Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. *Journal of Machine Learning Research*, 24(161):1–61, 2023. URL https://www.jmlr.org/papers/v24/22-0755.html.

Nathan Kallus and Masatoshi Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. In *Advances in Neural Information Processing Systems*, volume 33, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/75df63609809c7a2052fdffe5c00a84e-Abstract.html.

Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy policy gradients. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5089–5100, 2020b. URL https://proceedings.mlr.press/v119/kallus20c.html.

Ziad Kobeissi and Francis Bach. Temporal difference learning with continuous time and state in the stochastic setting, 2023. URL https://arxiv.org/abs/2202.07960. Version 3; first posted 2022.

Henry Lam, Haidong Li, and Xuhui Zhang. Minimax efficient finite-difference stochastic gradient estimators using black-box function evaluations, 2020. URL https://arxiv.org/abs/2007.04443.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.02971.

Wenlong Mou. Continuous-time reinforcement learning: Ellipticity enables model-free value function approximation, 2026. URL https://arxiv.org/abs/2602.06930.

Rémi Munos. Policy gradient in continuous time. *Journal of Machine Learning Research*, 7(27): 771–791, 2006. URL https://www.jmlr.org/papers/v7/munos06b.html.

Minh Nguyen. Decoupled continuous-time reinforcement learning via hamiltonian flow, 2026. URL https://arxiv.org/abs/2602.14587.

Seohong Park, Jaekyeom Kim, and Gunhee Kim. Time discretization-invariant safe action repetition for policy gradient methods, 2021. URL https://arxiv.org/abs/2111.03941.

Haruki Settai, Naoya Takeishi, and Takehisa Yairi. A temporal difference method for stochastic continuous dynamics. In *Advances in Neural Information Processing Systems*, volume 38, pp. 25469–25493, 2025. doi: 10.52202/085713-0856.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 387–395, 2014. URL https://proceedings.mlr.press/v32/silver14.html.

Corentin Tallec, Léonard Blier, and Yann Ollivier. Making deep q-learning methods robust to time discretization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6096–6104, 2019. URL https://proceedings.mlr.press/v97/tallec19a.html.

Wenpin Tang and Xun Yu Zhou. Regret of exploratory policy improvement and q-learning, 2024. URL https://arxiv.org/abs/2411.01302. Version 2 revised August 2026.

Boyu Wang, Xuefeng Gao, and Lingfei Li. Reinforcement learning for continuous-time optimal execution: Actor–critic algorithm and error analysis. *Finance and Stochastics*, 30(2):597–655, 2026. doi: 10.1007/s00780-026-00589-5. URL https://link.springer.com/article/10.1007/s00780-026-00589-5.

Huaqing Xiong, Tengyu Xu, Lin Zhao, Yingbin Liang, and Wei Zhang. Deterministic policy gradient: Convergence analysis. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2159–2169, 2022. URL https://proceedings.mlr.press/v180/xiong22a.html.

Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason D. Lee. Stochastic zeroth-order optimization under strongly convexity and lipschitz hessian: Minimax sample complexity. In *Advances in Neural Information Processing Systems*, volume 37, pp. 99564–99600, 2024. doi: 10.52202/079017-3159. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b444ad72520a5f5c467343be88e352ed-Abstract-Conference.html.

Hanyang Zhao, Wenpin Tang, and David Yao. Policy optimization for continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/2c53bc01e30711a08f6ac86919193022-Abstract-Conference.html.

Runze Zhao, Yue Yu, Ruhan Wang, Chunfeng Huang, and Dongruo Zhou. Instance-dependent continuous-time reinforcement learning via maximum likelihood estimation, 2025a. URL https://arxiv.org/abs/2508.02103.

Runze Zhao, Yue Yu, Adams Yiyue Zhu, Chen Yang, and Dongruo Zhou. Sample and computationally efficient continuous-time reinforcement learning with general function approximation. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 5030–5057, 2025b. URL https://proceedings.mlr.press/v286/zhao25a.html.

Mo Zhou and Jianfeng Lu. Solving time-continuous stochastic optimal control problems: Algorithm design and convergence analysis of actor-critic flow, 2024. URL https://arxiv.org/abs/2402.17208. Version 2 revised January 2025.

Yuhua Zhu, Yuming Zhang, and Haoyu Zhang. Optimal-PhiBE: A PDE-based model-free framework for continuous-time reinforcement learning, 2025. URL https://arxiv.org/abs/2506.05208.

A REGULARITY AND STABILITY OF THE CRITIC

The critic in AdvG learns two quantities with different roles. The value estimate V_θ predicts the return under the current policy. The slope estimate $C(x)^\top \omega$ approximates $g^\phi(x) = \nabla_a q^\phi(x, \mu_\phi(x))$, the derivative that tells the actor how changing its action affects return. How can these two estimates be learned together without losing stability as the observation interval h decreases? The key is how their expected updates depend on one another, which we will present in this chapter.

Notation and the starting-state distribution. Throughout, the actor parameter lies in a compact convex set $\Phi \subset \mathbb{R}^{d_\phi}$, and the critic parameter lies in a compact convex set $\mathcal{Z} \subset \mathbb{R}^{p_V+p_Q}$. Actor-parameter differentiability is understood on an open neighborhood of Φ. The window duration satisfies $\delta = Lh \in [\delta_{\min}, \delta_{\max}] \subset (0, \infty)$. Write $\gamma = e^{-\beta h}$ and take $0 < \sigma_e \leq 1$. We intentionally reuse C for sufficiently large positive constants, whose values may differ from one occurrence to the next. These constants are uniform over the allowed actors and sufficiently small h, σ_e; they may depend on the fixed model bounds and window range. The notation $C(x)$ denotes the slope-feature matrix.

For a function v, let $P_h^a v(x)$ be its expected value after starting at x and keeping action a fixed for time h. The policy transition and exploratory transition are

$$P_{\phi,h}^0 v(x) = P_h^{\mu_\phi(x)} v(x), \qquad P_{\phi,h}^{\sigma_e} v(x) = \mathbb{E}_U P_h^{\mu_\phi(x)+U} v(x), \qquad U \sim N(0, \sigma_e^2 I).$$

For a state law ν, νP denotes the law after one transition with P, and $\nu P v = \int P v(x) \nu(dx)$ whenever the integral exists. In particular, $\nu_0 (P_{\phi,h}^0)^j$ is the state law after j policy steps from the initial law ν_0. Algorithm 1 selects this prefix length with probability $(1-\gamma)\gamma^j$, so the first state in each exploratory window has the mixture distribution

$$d_{\phi,h}^0 = (1-\gamma) \sum_{j \geq 0} \gamma^j \nu_0 (P_{\phi,h}^0)^j. \tag{23}$$

This identity will give the discount inequality behind value-block stability. Complete trajectories are independent at a fixed actor; cells within a trajectory need not be independent. Write $A_i = \mu_\phi(X_i) + U_i$ and $r_i = r(X_i, A_i)$ for the exploratory window $X_0, \ldots, X_L$. Both update fields use parameters fixed before collection.

Recall the generator $\mathcal{L}^a$ and the continuous-policy occupation measure ρ_ϕ are defined in Section 2. We write $\mathcal{L}^{\mu_\phi} v(x) = \mathcal{L}^a v(x)|_{a=\mu_\phi(x)}$ and $\mathsf{P}_t^\phi v(x) = \mathbb{E}_x v(X_t^\phi)$ for continuous policy feedback. Unlike $(P_{\phi,h}^0)^j$, this last transition updates the action continuously.

From cell residuals to the weighted critic equation. For a fixed actor and critic coefficients $z = (\theta, \omega)$, the scalar residual from cell i in (8) is

$$D_i(z) = V_\theta(X_i) - \gamma V_\theta(X_{i+1}) - h r_i + h \omega^\top C(X_i) U_i, \qquad i = 0, \ldots, L-1. \tag{24}$$

It compares the predicted advantage increment $h\omega^\top C(X_i)U_i$ with the observed reward plus estimated value change $hr_i + \gamma V_\theta(X_{i+1}) - V_\theta(X_i)$. Both value evaluations use the same current coefficients θ.

Write $S_i = \sigma_e^{-2} C(X_i) U_i$ for the slope weight. For the observed window Ξ, the moment in (9) is $m_\phi = m_\phi^{\mathrm{p}}$, with value and slope blocks

$$m_\phi(z; \Xi) = \begin{bmatrix} m_{\phi,V}(z; \Xi) \\ m_{\phi,Q}(z; \Xi) \end{bmatrix} = \frac{1}{\delta} \sum_{i<L} \gamma^i \begin{bmatrix} f(X_0) \\ S_i \end{bmatrix} D_i(z), \qquad \bar{m}_\phi(z) = \mathbb{E} m_\phi(z; \Xi). \tag{25}$$

Thus the value equation weights every cell residual by $f(X_0)$, whereas the slope equation weights it by its own S_i. The matrix $W_\chi = \mathrm{diag}(I_{p_V}, \chi I_{p_Q})$ then multiplies the slope block by a fixed factor $\chi > 0$. It is a block weight applied to the critic vector, while $D_i(z)$ is the scalar residual used to construct that vector.

Let $\mathcal{F}_i^-$ be the history before drawing U_i, and let $\mathcal{F}_i^A = \mathcal{F}_i^- \vee \sigma(U_i)$ also include the action, before observing X_{i+1}. Both $f(X_0)$ and S_i are $\mathcal{F}_i^A$-measurable. Taking conditional expectation inside

(25) therefore gives

$$W_\chi \bar{m}_\phi(z) = \begin{bmatrix} \bar{m}_{\phi,V}(z) \\ \chi \bar{m}_{\phi,Q}(z) \end{bmatrix} = \frac{1}{\delta} \sum_{i<L} \gamma^i \mathbb{E}\left[\begin{bmatrix} f(X_0) \\ \chi S_i \end{bmatrix} \mathbb{E}[D_i(z) \mid \mathcal{F}_i^A] \right]. \tag{26}$$

This is how estimates of the conditional cell residual enter the weighted mean critic equation. Corollary A.7 controls the fluctuations of the sampled vector around this mean; Theorem B.1 bounds the mean at the reference critic to locate the fitted root.

Quadratic growth convention. We use a the standard weight $w(x) = 1 + \|x\|^2$. For $j \geq 2$, set

$$\|v\|_{C_w^j} = \sup_x \left\{ \frac{\|v(x)\|}{1+\|x\|^2} + \frac{\|\partial_x v(x)\|}{\sqrt{1+\|x\|^2}} + \sum_{k=2}^{j} \|\partial_x^k v(x)\| \right\}, \qquad \|v\|_{C_w^0} = \sup_x \frac{\|v(x)\|}{1+\|x\|^2}.$$

Here ∂_x^k denotes the kth state derivative. Thus values may grow quadratically, gradients linearly, and higher derivatives are bounded. This includes quadratic value functions without requiring bounded states. The weight is fixed, not a parameter of the model class. Norms are Euclidean or induced operator norms, $\mathrm{sym}(A) = (A + A^\top)/2$, and $\succeq$ denotes positive-semidefinite order.

We now state conditions on the dynamics, reward, policy, and features before proving stability of the expected critic equation. The assumptions on the true value and feature approximation will also be used in Section B.1 to compare the critic's actor direction with the true policy gradient.

Assumption A.1 (Smooth dynamics and state moments)**.** *The drift $b(x,a)$ and diffusion coefficient $\sigma(x,a)$ are three times continuously differentiable. Their derivatives of orders one through three are uniformly bounded. In particular, the coefficients are globally Lipschitz and grow at most linearly in state and action.*

Starting from ν_0, the state has a uniformly bounded sixteenth moment under every admitted policy:

$$\sup_{t \geq 0} \mathbb{E}_{\nu_0}[1 + \|X_t\|^{16}] \leq M.$$

This bound holds both when $A_t = \mu_\phi(X_t)$ is updated continuously and when the action is updated only at observation times and held constant between them, uniformly in ϕ and sufficiently small h.

The moment condition is a stability requirement on the controlled state process. It permits unbounded Gaussian states, including the stable LQG example in Section B.3.

Assumption A.2 (Reward, policy, and critic features)**.** *The policies are C^2 in x and C^1 in ϕ, with uniformly bounded $\partial_x \mu_\phi, \partial_x^2 \mu_\phi$ and $\sup_{\phi \in \Phi} \|\mu_\phi(0)\| < \infty$. The value features satisfy $\|f_r\|_{C_w^2} \leq C$, and*

$$\|C(x)\| + \|\partial_\phi \mu_\phi(x)\| \leq C(1 + \|x\|^2).$$

The reward satisfies $r \in C^2(\mathbb{R}^{d_s} \times \mathbb{R}^{d_a})$ and $\partial_a^3 r \in C^0$, with

$$|r(x,a)| \leq C(1 + \|x\|^2 + \|a\|^2), \qquad \|\nabla_{(x,a)} r(x,a)\| \leq C(1 + \|x\| + \|a\|),$$
$$\|\nabla^2_{(x,a)} r(x,a)\| + \|\partial_a^3 r(x,a)\| \leq C.$$

All bounds are uniform over the actor set.

Assumption A.3 (Value regularity)**.** *On an open neighborhood of Φ, the true value V^ϕ satisfies $(\beta - \mathcal{L}^{\mu_\phi}) V^\phi = r(\cdot, \mu_\phi)$. It is C^1 in ϕ, with $V^\phi, \nabla_\phi V^\phi \in C_w^4$ and the uniform bound*

$$\|V^\phi\|_{C_w^4} + \|\nabla_\phi V^\phi\|_{C_w^4} \leq C.$$

The return J has an L_J-Lipschitz gradient on a neighborhood of Φ.

These conditions concern the true value and its sensitivity to the policy parameters. In Section B.1, their smoothness controls the error from exploring around the policy action and from observing the state only at discrete times.

Assumption A.4 (Feature coverage, approximation, and policy dependence)**.**

(i) **Coverage.** *The sampled states must distinguish different coefficients in both feature families. Under $X_0 \sim d^0_{\phi,h}$ and the exploratory window,*

$$\mathbb{E}[f(X_0) f(X_0)^\top] \succeq \kappa_V I, \qquad \frac{h}{\delta} \sum_{i<L} \gamma^i \mathbb{E}[C(X_i) C(X_i)^\top] \succeq \kappa_Q I. \tag{27}$$

The constants κ_V, κ_Q are positive and independent of h, σ_e, ϕ. The corresponding upper moments are bounded.

(ii) **Approximation.** *The fixed features need only approximate the value and the action derivative used by the actor. For each actor there is a reference $\widetilde{z}_\phi = (\widetilde{\theta}_\phi, \widetilde{\omega}_\phi) \in \mathcal{Z}$, independent of h, σ_e, with distance at least $r_{\mathcal{Z}} > 0$ from $\partial\mathcal{Z}$, such that*

$$\|V_{\widetilde{\theta}_\phi} - V^\phi\|_{C^2_w} \le e_V,$$
$$\sup_{i<L}\big(\mathbb{E}\|C(X_i)^\top \widetilde{\omega}_\phi - g^\phi(X_i)\|^2\big)^{1/2} + \left(\int \|C^\top \widetilde{\omega}_\phi - g^\phi\|^2 d\rho_\phi\right)^{1/2} \le e_Q. \tag{28}$$

(iii) **Policy dependence.** *The expected critic equation is affine in its coefficients: $\bar{m}_\phi(z) = A_\phi z - b_\phi$, where $\bar{m}_\phi(z) = \mathbb{E}[m^{\mathrm{p}}_\phi(z;\Xi)]$. For fixed h, σ_e, its matrix and reward vector vary Lipschitzly with the actor:*

$$\|A_\phi - A_\psi\| + \|b_\phi - b_\psi\| \le C\|\phi - \psi\|, \qquad \phi, \psi \in \Phi. \tag{29}$$

This controls how far the critic's target moves after an actor update; no differentiation of the sampling law is required.

A coverage condition ensures that the observed states distinguish the feature coefficients; note that Gaussian action exploration alone does not guarantee it. We say the features are exact if $e_V = e_Q = 0$. They then represent V^ϕ and g^ϕ, even when $q^\phi(x, a)$ is nonlinear in a. The bound (29) controls how the expected critic equation changes when the actor moves.

A.1 Stability and sampling noise of the expected critic equation

Recall that $m(z;\Xi)$ is the observed critic equation assembled from cell residuals in (25). Since the critic is linear in $z = (\theta, \omega)$, its expectation has the form $\bar{m}(z) = Az - b$. Thus A describes how a coefficient error changes the expected update. The following theorem identifies the structure that makes this update stable. We use the transition estimates from Section A.2 at the steps where they are needed; their proofs are collected there.

Theorem A.5 (Block structure of the mean critic equation)**.** *Assume Assumptions A.1–A.2 and the coverage clause* (i) *of Assumption A.4. Fix ϕ, h, σ_e and write $m, \bar{m}, A, b$ for the sampled equation, its expectation, its matrix, and its reward vector. The matrix has the block form*

$$A = \begin{pmatrix} A_V & 0 \\ A_{QV} & A_Q \end{pmatrix}, \tag{30}$$

where A_V acts on value coefficients, A_Q acts on slope coefficients, and A_{QV} describes how value errors enter the slope equation. These blocks satisfy

$$\mathrm{sym}(A_V) \succeq \left\{\frac{1-\gamma^{L/2}}{\delta}\kappa_V - C\sigma_e^2\right\} I,$$
$$A_Q = \frac{h}{\delta}\sum_{i<L}\gamma^i \mathbb{E}[C(X_i)C(X_i)^\top] \succeq \kappa_Q I, \qquad \|A_{QV}\| \le K_{QV}. \tag{31}$$

Here $K_{QV} \ge 0$ is a fixed upper bound on the coupling block $\|A_{QV}\|$, uniform over the allowed actors, observation meshes, exploration strengths, and window durations. We retain this named constant to choose the relative update weight below.

Proof of Theorem A.5. Fix (ϕ, h, σ_e). We first identify which coefficients enter each equation, then bound the value and coupling blocks. The slope block follows directly from feature coverage.

Use the history $\mathcal{F}_i^-$ and the slope weight S_i defined above. In particular, X_i is known before U_i is drawn. The discounted value terms telescope across the window, so

$$D^V(z) = \theta^\top(f(X_0) - \gamma^L f(X_L)) - h\sum_{i<L}\gamma^i(r_i - \omega^\top C(X_i)U_i).$$

Substituting this identity and (24) into (25), the coefficients of (θ, ω) are

$$A_V = \delta^{-1}\mathbb{E}[f(X_0)(f(X_0) - \gamma^L f(X_L))^\top], \qquad A_{VQ} = \frac{h}{\delta}\sum_{i<L}\gamma^i\mathbb{E}[f(X_0)U_i^\top C(X_i)^\top],$$

$$A_{QV} = \delta^{-1}\sum_{i<L}\gamma^i\mathbb{E}[S_i(f(X_i) - \gamma f(X_{i+1}))^\top], \qquad A_Q = \frac{h}{\delta}\sum_{i<L}\gamma^i\mathbb{E}[S_iU_i^\top C(X_i)^\top].$$

Both $f(X_0)$ and $C(X_i)$ are $\mathcal{F}_i^-$-measurable, and

$$\mathbb{E}[U_i \mid \mathcal{F}_i^-] = 0, \qquad \mathbb{E}[U_iU_i^\top \mid \mathcal{F}_i^-] = \sigma_e^2 I.$$

Consequently,

$$\mathbb{E}[f(X_0)U_i^\top C(X_i)^\top \mid \mathcal{F}_i^-] = 0, \qquad \mathbb{E}[S_iU_i^\top C(X_i)^\top \mid \mathcal{F}_i^-] = C(X_i)C(X_i)^\top.$$

Hence $A_{VQ} = 0$, and (27) gives $A_Q \succeq \kappa_Q I$. We notice that A has a block-triangular structure: the slope coefficients do not affect the expected value update, while the reverse dependence remains in A_{QV}.

Discounting stabilizes the value equation. First remove exploration from the window and call the resulting block A_V^0. Write $d = d_{\phi,h}^0$, $P^0 = P_{\phi,h}^0$, and $g = u^\top f$: g is the change in the value estimate produced by a coefficient change u. Reindexing the geometric start law gives

$$\gamma^L d(P^0)^L = (1-\gamma)\sum_{n\geq L}\gamma^n\nu_0(P^0)^n \leq d. \tag{32}$$

The discounted distribution after L further policy cells is a tail of the mixture defining d. In particular, $\gamma^L\mathbb{E}_{d,P^0}g(X_L)^2 \leq \mathbb{E}_d g^2$. Cauchy–Schwarz therefore yields

$$\gamma^L|\mathbb{E}_{d,P^0}[g(X_0)g(X_L)]| \leq \gamma^{L/2}\,\mathbb{E}_d g^2, \qquad u^\top A_V^0 u \geq \frac{1-\gamma^{L/2}}{\delta}\mathbb{E}_d g^2.$$

Thus the discounted future value cannot cancel the whole current value signal. Coverage supplies $\mathbb{E}_d g^2 \geq \kappa_V\|u\|^2$, converting this fact into positivity in every value-coefficient direction.

It remains to check that exploration preserves this positivity. Writing $P^{\sigma_e} = P_{\phi,h}^{\sigma_e}$, we change the action law one cell at a time:

$$(P^{\sigma_e})^L - (P^0)^L = \sum_{j<L}(P^{\sigma_e})^j(P^{\sigma_e} - P^0)(P^0)^{L-1-j}.$$

In each summand, $(P^0)^{L-1-j}f_r$ is the expected value of the fixed feature f_r after $L-1-j$ target-policy cells. Lemma A.9 bounds the change caused by centered exploration in one cell, while Lemma A.8 controls its propagation:

$$|(P^{\sigma_e} - P^0)(P^0)^{L-1-j}f_r(x)| \leq Ch\sigma_e^2 w(x), \qquad (P^{\sigma_e})^j w(x) \leq Cw(x).$$

Hence

$$\left\|(P^{\sigma_e})^j(P^{\sigma_e} - P^0)(P^0)^{L-1-j}f_r\right\|_{C_w^0} \leq Ch\sigma_e^2.$$

There are L such changes, so their total size is $O(\delta\sigma_e^2)$. The normalization by δ in A_V cancels this window length:

$$\|((P^{\sigma_e})^L - (P^0)^L)f\|_{C_w^0} \leq C\delta\sigma_e^2, \qquad \|A_V - A_V^0\| \leq C\sigma_e^2.$$

For the second bound, Lemma A.8 and the feature growth assumption give $\mathbb{E}[\|f(X_0)\|w(X_0)] \leq C$. Together with the preceding lower bound on A_V^0, it proves the value-block inequality in (31).

Finally, we show value errors have a bounded effect on the slope equation. The same centering gives $\mathbb{E}[S_i f(X_i)^\top] = 0$, so only the next-state feature contributes to A_{QV}. To replace the score by an action derivative, we use Gaussian integration by parts: for a continuously differentiable $\psi : \mathbb{R}^{d_a} \to \mathbb{R}^p$ whose value and Jacobian have polynomial growth,

$$\mathbb{E}[\sigma_e^{-2}U_i\psi(U_i)^\top \mid \mathcal{F}_i^-] = \mathbb{E}[\partial_u\psi(U_i)^\top \mid \mathcal{F}_i^-]. \tag{33}$$

The Gaussian density satisfies $\nabla p_{\sigma_e}(u) = -\sigma_e^{-2} u p_{\sigma_e}(u)$, and its decay makes the boundary term zero. Conditioning fixes X_i, so the identity also applies when ψ depends on X_i. Together with the Markov property, it gives

$$\mathbb{E}[S_i f(X_{i+1})^\top \mid \mathcal{F}_i^-] = C(X_i)\mathbb{E}_U \left[\partial_a [P_h^a f](X_i)^\top\big|_{a=\mu_\phi(X_i)+U}\right].$$

Lemma A.9 supplies the factor h:

$$\|\partial_a P_h^a f(X_i)|_{a=\mu_\phi(X_i)+U}\| \le Chw(X_i)(1+\|U\|^2).$$

Using Lemma A.8 to average over X_i then gives

$$\|\mathbb{E}[S_i f(X_{i+1})^\top]\| \le Ch, \qquad \|A_{QV}\| \le \frac{Ch}{\delta}\sum_{i<L}\gamma^{i+1} \le K_{QV}.$$

The $L = \delta/h$ contributions therefore remain bounded after division by δ. This proves the remaining bound in (31). □

The coupling need not be small by itself. We control it by reducing the weight of the slope update relative to the value update. A fixed weight suffices because Theorem A.5 bounds the coupling uniformly in h and σ_e.

Corollary A.6 (Stability and the critic root)**.** *Under the hypotheses of Theorem A.5, choose* $0 < \sigma_{e,0} \le 1$ *so the value bound is at least* $a_V I$, $a_V > 0$, *and* $0 < \chi \le \min\{1, a_V\kappa_Q/K_{QV}^2\}$, *omitting the second restriction when* $K_{QV} = 0$. *For* $0 < \sigma_e \le \sigma_{e,0}$ *and* $W_\chi = \mathrm{diag}(I, \chi I)$, *there are* $\mu_\chi, L_\chi > 0$, *independent of* h, σ_e, *such that*

$$\begin{aligned}\langle z - z', W_\chi\{\bar m(z) - \bar m(z')\}\rangle &\ge \mu_\chi\|z - z'\|^2,\\ \|W_\chi\{\bar m(z) - \bar m(z')\}\| &\le L_\chi\|z - z'\|.\end{aligned} \tag{34}$$

Together, these inequalities imply contraction of the expected update for sufficiently small critic steps. There is a unique, uniformly bounded fitted critic $z^\star = A^{-1}b$, *characterized by* $\bar m(z^\star) = 0$.

Proof of Corollary A.6. The window range is fixed, so $\inf_{\delta\in[\delta_{\min},\delta_{\max}]} \kappa_V(1-e^{-\beta\delta/2})/\delta > 0$. Theorem A.5 therefore permits a fixed $\sigma_{e,0}$ for which $\mathrm{sym}(A_V) \succeq a_V I$. The formulas for A_V, A_Q, the feature moments, and $\delta \ge \delta_{\min}$ give $\|A_V\| \le K_V$ and $\|A_Q\| \le K_Q$. Write a coefficient error as $e = (u, v)$, with value part u and slope part v. In the quadratic form below, the two diagonal blocks supply positive terms, while the coupling can have either sign. Young's inequality and the choice of χ absorb that coupling into half of each positive term:

$$\langle e, W_\chi A e\rangle \ge a_V\|u\|^2 - \chi K_{QV}\|u\|\|v\| + \chi\kappa_Q\|v\|^2 \ge \frac{a_V}{2}\|u\|^2 + \frac{\chi\kappa_Q}{2}\|v\|^2.$$

Thus (34) holds with $\mu_\chi = \frac{1}{2}\min\{a_V, \chi\kappa_Q\}$ and $L_\chi = K_V + \chi K_{QV} + \chi K_Q$. These bounds give

$$\|e - \eta W_\chi A e\|^2 \le (1 - 2\mu_\chi\eta + L_\chi^2\eta^2)\|e\|^2,$$

which is a contraction for $0 < \eta < 2\mu_\chi/L_\chi^2$. They also imply that A is invertible and $\|(W_\chi A)^{-1}\| \le \mu_\chi^{-1}$. Multiplying the equation by W_χ changes its update scales but not its solution.

To bound that solution, it remains to bound the reward vector. Its blocks are

$$b_V = \frac{h}{\delta}\sum_{i<L}\gamma^i\mathbb{E}[f(X_0)r_i], \qquad b_Q = \frac{h}{\delta}\sum_{i<L}\gamma^i\mathbb{E}[S_i r_i].$$

The reward and feature moments bound $\|b_V\|$. For b_Q, (33) gives

$$\mathbb{E}[S_i r_i \mid \mathcal{F}_i^-] = C(X_i)\mathbb{E}_U\nabla_a r(X_i, \mu_\phi(X_i) + U).$$

The reward-derivative bound and (40) imply $\|\mathbb{E}[S_i r_i]\| \le C$. Since $hL/\delta = 1$, we obtain $\|b\| \le C$ and

$$\|z^\star\| = \|(W_\chi A)^{-1}W_\chi b\| \le C.$$

□

Apart from stability of the expected update, we also seek control control the noise in an observed update. We now prove a variance bound that grows as the action perturbations shrink, but stays uniform in the observation mesh at a fixed physical window length.

Corollary A.7 (Critic sampling noise)**.** *Under the hypotheses and with the fixed weight χ of Corollary A.6, let $\Sigma^2_{\sigma_e,\delta} = (\delta\sigma_e^2)^{-1}$. For every deterministic critic coefficient vector z,*

$$\mathbb{E}\|W_\chi\{m(z;\Xi) - \bar{m}(z)\}\|^2 \le C\Sigma^2_{\sigma_e,\delta}(1 + \|z - z^\star\|^2). \tag{35}$$

For an average $\widehat{m}_B$ of B conditionally independent complete trajectories, given a history $\mathcal{G}$ fixing (ϕ, z),

$$\mathbb{E}[\|W_\chi\{\widehat{m}_B(z) - \bar{m}(z)\}\|^2 \mid \mathcal{G}] \le \frac{C\Sigma^2_{\sigma_e,\delta}}{B}(1 + \|z - z^\star\|^2). \tag{36}$$

In particular, on the fixed compact critic set $\mathcal{Z}$,

$$\mathbb{E}[\|W_\chi\{\widehat{m}_B(z) - \bar{m}(z)\}\|^2 \mid \mathcal{G}] \le C_{\mathcal{Z}}\Sigma^2_{\sigma_e,\delta}/B, \qquad z \in \mathcal{Z}, \tag{37}$$

with a constant independent of h, σ_e, B.

Proof of Corollary A.7. We use (25) to bound the two blocks of $W_\chi m(z;\Xi)$. For each scalar residual (24), we separate its conditional mean from its centered transition error. This permits us to add the variances of the centered terms even though neighboring cells share states. With the history $\mathcal{F}_i^A$ defined above, fix any deterministic $z = (\theta, \omega)$ and write

$$\begin{aligned} D_i(z) &= M^\theta_{i+1} + F^z_i, \\ M^\theta_{i+1} &= -\gamma\theta^\top\{f(X_{i+1}) - P_h^{A_i} f(X_i)\}, \\ F^z_i &= \theta^\top\{f(X_i) - \gamma P_h^{A_i} f(X_i)\} - h(r_i - \omega^\top C(X_i)U_i). \end{aligned}$$

Then

$$\mathbb{E}[M^\theta_{i+1} \mid \mathcal{F}_i^A] = 0, \qquad F^z_i = \mathbb{E}[D_i(z) \mid \mathcal{F}_i^A].$$

Substituting this decomposition into the sampled critic vector gives

$$W_\chi m(z;\Xi) = \frac{1}{\delta}\sum_{i<L}\gamma^i \begin{bmatrix} f(X_0) \\ \chi S_i \end{bmatrix}(M^\theta_{i+1} + F^z_i).$$

We therefore need four bounds: the value and slope weights applied to the centered errors and to the conditional means. The cell bounds in Lemma A.8, Itô isometry, and Cauchy–Schwarz give

$$\begin{aligned} \mathbb{E}[|M^\theta_{i+1}|^2 \mid \mathcal{F}_i^A] &\le Ch\|\theta\|^2 w(X_i)^2(1 + \|U_i\|^4), \\ |F^z_i|^2 &\le Ch^2(1 + \|z\|^2)w(X_i)^2(1 + \|U_i\|^4). \end{aligned}$$

For the second bound, use $f(x) - \gamma P_h^a f(x) = (1-\gamma)f(x) - \gamma\mathbb{E}_x^a\int_0^h \mathcal{L}^a f(X_s)ds$ and $1 - \gamma = O(h)$. Multiplying by $\|f(X_0)\|^2$ or $\|S_i\|^2$ and taking expectations, the Gaussian probe moments and (40) give

$$\begin{aligned} \mathbb{E}\|f(X_0)M^\theta_{i+1}\|^2 &\le Ch\|\theta\|^2, & \mathbb{E}\|f(X_0)F^z_i\|^2 &\le Ch^2(1 + \|z\|^2), \\ \mathbb{E}\|S_iM^\theta_{i+1}\|^2 &\le Ch\sigma_e^{-2}\|\theta\|^2, & \mathbb{E}\|S_iF^z_i\|^2 &\le Ch^2\sigma_e^{-2}(1 + \|z\|^2). \end{aligned}$$

For either $I_i = f(X_0)$ or $I_i = S_i$, and $i < j$,

$$\mathbb{E}\langle I_iM^\theta_{i+1}, I_jM^\theta_{j+1}\rangle = \mathbb{E}\left\langle I_iM^\theta_{i+1}, I_j\mathbb{E}[M^\theta_{j+1} \mid \mathcal{F}_j^A]\right\rangle = 0.$$

Thus these centered contributions are orthogonal even though the states are dependent. For the slope equation, each has second moment of at most order h/σ_e^2; adding $L = \delta/h$ such contributions and dividing by δ^2 gives

$$\mathbb{E}\left\|\frac{1}{\delta}\sum_{i<L}\gamma^i S_iM^\theta_{i+1}\right\|^2 = \frac{1}{\delta^2}\sum_{i<L}\gamma^{2i}\mathbb{E}\|S_iM^\theta_{i+1}\|^2 \le \frac{C\|\theta\|^2}{\delta\sigma_e^2}. \tag{38}$$

The conditional means F_i^z need not be orthogonal. Their squared size is of order h^2, however, so even the Cauchy–Schwarz bound for a sum of L dependent terms stays bounded as h decreases:

$$\mathbb{E}\left\|\frac{1}{\delta}\sum_{i<L}\gamma^i S_i F_i^z\right\|^2 \le \frac{L}{\delta^2}\sum_{i<L}\gamma^{2i}\mathbb{E}\|S_i F_i^z\|^2 \le C\sigma_e^{-2}(1+\|z\|^2). \tag{39}$$

The value block uses $f(X_0)$ instead of the Gaussian score. The same argument applies without the factor σ_e^{-2}:

$$\mathbb{E}\left\|\frac{f(X_0)}{\delta}\sum_{i<L}\gamma^i M_{i+1}^\theta\right\|^2 = \frac{1}{\delta^2}\sum_{i<L}\gamma^{2i}\mathbb{E}\|f(X_0)M_{i+1}^\theta\|^2 \le \frac{C\|\theta\|^2}{\delta},$$

$$\mathbb{E}\left\|\frac{f(X_0)}{\delta}\sum_{i<L}\gamma^i F_i^z\right\|^2 \le \frac{L}{\delta^2}\sum_{i<L}\gamma^{2i}\mathbb{E}\|f(X_0)F_i^z\|^2 \le C(1+\|z\|^2).$$

The matrix W_χ combines the two blocks through their squared norms. Using the four bounds above, $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$, $\sigma_e \le 1$, $\delta \le \delta_{\max}$, and the fixed weight χ, we obtain

$$\begin{aligned}\mathbb{E}\|W_\chi m(z;\Xi)\|^2 &= \mathbb{E}\|m_V(z;\Xi)\|^2 + \chi^2\mathbb{E}\|m_Q(z;\Xi)\|^2\\ &\le \frac{C}{\delta\sigma_e^2}(1+\|z\|^2).\end{aligned}$$

Corollary A.6, which applies the block bounds of Theorem A.5, gives $\|z^\star\| \le C$. Hence $1+\|z\|^2 \le C(1+\|z-z^\star\|^2)$. Since centering reduces the second moment, this proves (35) for every z.

Conditional independence of complete trajectories divides the centered second moment by B, proving (36). The uniform root bound and compactness of $\mathcal{Z}$ then give (37). Independence is used across trajectories; within each window it is the displayed conditional centering that eliminates the cross-cell noise terms. □

A.2 Transition estimates and their proofs

We now prove the transition estimates used in the stability and sampling-noise arguments; these estimates follow from our assumptions.

Lemma A.8 (State moments and cell increments)**.** *Assume Assumptions A.1–A.2. Fix an allowed actor ϕ, sufficiently small $0 < h \le 1$, $0 < \sigma_e \le 1$, and an integer L with $Lh = \delta \in [\delta_{\min}, \delta_{\max}]$. For an exploratory window with $X_0 \sim d^0_{\phi,h}$,*

$$\mathbb{E}_{d^0_{\phi,h}} w^8 + \int w^8\, d\rho_\phi + \max_{0\le i\le L}\mathbb{E}w(X_i)^8 \le C.$$

For a window started at any $x \in \mathbb{R}^{d_s}$,

$$\mathbb{E}[w(X_i)^8 \mid X_0 = x] \le Cw(x)^8, \qquad i = 0,\dots,L. \tag{40}$$

For any scalar $v \in C^2_w$, $x \in \mathbb{R}^{d_s}$, and $u \in \mathbb{R}^{d_a}$, holding $a = \mu_\phi(x) + u$ fixed over one cell gives

$$\mathbb{E}_x^a\int_0^h \left\{|\mathcal{L}^a v(X_s)|^2 + \|\sigma(X_s,a)^\top\nabla v(X_s)\|^2\right\} ds \le Chw(x)^2(1+\|u\|^4)\|v\|^2_{C^2_w} \tag{41}$$

and

$$\begin{aligned}|\mathbb{E}_x^a[v(X_h) - v(x)]| &\le Chw(x)(1+\|u\|^2)\|v\|_{C^2_w},\\ \mathbb{E}_x^a|v(X_h) - v(x)|^2 &\le Chw(x)^2(1+\|u\|^4)\|v\|^2_{C^2_w}.\end{aligned} \tag{42}$$

The constants are uniform in the displayed parameters and in i.

Proof of Lemma A.8. We first bound the moments under the window-start law and the continuous-policy occupation measure. By Assumption A.1 and $w(x)^8 \le C(1+\|x\|^{16})$,

$$\mathbb{E}_{d^0_{\phi,h}} w^8 = (1-\gamma)\sum_{j\ge 0}\gamma^j\nu_0(P^0_{\phi,h})^j w^8 \le C,$$

$$\int w^8\, d\rho_\phi = \int_0^\infty e^{-\beta t}\mathbb{E}_{\nu_0}[w(X_t^\phi)^8]dt \le C.$$

We next bound the moments throughout an exploratory window, with $X_0 = x \in \mathbb{R}^{d_s}$, $0 < h \le 1$, and $Lh = \delta \le \delta_{\max}$. Since $\mathbb{E}\|U_i\|^{16} \le C\sigma_e^{16} \le C$, Assumptions A.1–A.2, Itô's formula, and Gronwall's inequality give

$$P^{\sigma_e}_{\phi,h} w^8(x) \le (1 + Ch) w(x)^8.$$

Therefore,

$$\begin{aligned}\mathbb{E}[w(X_i)^8 \mid X_0 = x] &= (P^{\sigma_e}_{\phi,h})^i w^8(x) \le (1+Ch)^i w(x)^8 \\ &\le e^{C\delta_{\max}} w(x)^8 \le C w(x)^8, \qquad i = 0, \dots, L.\end{aligned}$$

Averaging over $X_0 \sim d^0_{\phi,h}$ gives

$$\max_{0 \le i \le L} \mathbb{E} w(X_i)^8 \le C \mathbb{E}_{d^0_{\phi,h}} w^8 \le C.$$

We now bound the mean and second moment of $v(X_h) - v(x)$ over one cell, with $a = \mu_\phi(x) + u$ held fixed and $\|v\|_{C^2_w} < \infty$. By Assumptions A.1–A.2,

$$\begin{aligned}&\sup_{0 \le s \le h} \mathbb{E}^a_x (1 + \|X_s\| + \|a\|)^4 \le C w(x)^2 (1 + \|u\|^4), \\ &|\mathcal{L}^a v(y)| + \|\sigma(y,a)^\top \nabla v(y)\| \le C(1 + \|y\| + \|a\|)^2 \|v\|_{C^2_w}.\end{aligned}$$

Integrating these bounds gives (41). For the increments, Itô's formula gives

$$v(X_h) - v(x) = \int_0^h \mathcal{L}^a v(X_s) ds + \int_0^h \nabla v(X_s)^\top \sigma(X_s, a) dW_s.$$

Taking expectations and second moments, Cauchy–Schwarz and Itô isometry yield (42).

□

Lemma A.9 (Action sensitivity and exploration)**.** *Under the assumptions and parameter ranges of Lemma A.8, for every $x \in \mathbb{R}^{d_s}$, $u \in \mathbb{R}^{d_a}$, feature component f_r, and $j = 0, \dots, L$,*

$$\begin{aligned}\|\partial_a P^a_h f(x)|_{a = \mu_\phi(x) + u}\| &\le Ch w(x)(1 + \|u\|^2), \\ |(P^{\sigma_e}_{\phi,h} - P^0_{\phi,h})(P^0_{\phi,h})^j f_r(x)| &\le C\sigma_e^2 h w(x).\end{aligned} \tag{43}$$

The constants are uniform in the displayed parameters and in j, r.

Proof of Lemma A.9. We prove the two estimates in (43) by controlling the state and action derivatives of expected feature values. For a component f_r of f, set

$$v_j(x) = (P^0_{\phi,h})^j f_r(x) = \mathbb{E}_x[f_r(X_{jh})], \qquad j = 0, \dots, L.$$

For the rollout with $A_t = \mu_\phi(X_{ih})$ on $ih \le t < (i+1)h$, we have $\partial_x X_0 = I$ and $\partial_x^2 X_0 = 0$. Taking Euclidean norms of the derivative arrays, the cell variational equations, Itô's formula, and Young's inequality give, for $p \ge 2$,

$$\begin{aligned}\mathbb{E}\|\partial_x X_{(i+1)h}\|^p &\le (1 + C_p h) \mathbb{E}\|\partial_x X_{ih}\|^p, \\ \mathbb{E}\|\partial_x^2 X_{(i+1)h}\|^p &\le (1 + C_p h) \mathbb{E}\|\partial_x^2 X_{ih}\|^p + C_p h \mathbb{E}\|\partial_x X_{ih}\|^{2p}.\end{aligned}$$

Applying the first estimate also at exponent $2p$ and iterating gives

$$\max_{0 \le i \le L} \left(\mathbb{E}\|\partial_x X_{ih}\|^p + \mathbb{E}\|\partial_x^2 X_{ih}\|^p\right) \le C_p e^{C_p Lh} \le C_p, \qquad Lh \le \delta_{\max}.$$

These bounds imply the corresponding operator-norm bounds. Differentiating v_j gives

$$\begin{aligned}\partial_x v_j(x) &= \mathbb{E}[\partial_x f_r(X_{jh}) \partial_x X_{jh}], \\ \partial_x^2 v_j(x) &= \mathbb{E}[\partial_x^2 f_r(X_{jh})[\partial_x X_{jh}, \partial_x X_{jh}] + \partial_x f_r(X_{jh}) \partial_x^2 X_{jh}].\end{aligned}$$

The feature assumption and the fixed-duration state estimate give

$$\mathbb{E}_x w(X_{jh}) \le C w(x), \qquad \|\partial_x f_r(y)\| \le C\sqrt{w(y)}, \qquad \|\partial_x^2 f_r(y)\| \le C.$$

Hölder's inequality therefore yields

$$|v_j(x)| \le Cw(x), \qquad \|\partial_x v_j(x)\| + \|\partial_x^2 v_j(x)\| \le C\sqrt{w(x)}.$$

We next bound the action derivatives with x fixed and $a = \mu_\phi(x) + u$ held constant:

$$dX_t = b(X_t, a)dt + \sigma(X_t, a)dW_t, \qquad X_0 = x, \qquad 0 \le t \le h \le 1.$$

For action coordinates k, ℓ, write $Y_t^k = \partial_{a_k} X_t$ and $Z_t^{k\ell} = \partial_{a_k}\partial_{a_\ell} X_t$. Subscripts on b, σ below denote partial derivatives evaluated at (X_t, a). Differentiating the SDE gives

$$\begin{aligned} dY_t^k &= (b_x Y_t^k + b_{a_k})dt + (\sigma_x Y_t^k + \sigma_{a_k})dW_t, \\ dZ_t^{k\ell} &= \big(b_x Z_t^{k\ell} + b_{xx}[Y_t^k, Y_t^\ell] + b_{xa_\ell} Y_t^k + b_{xa_k} Y_t^\ell + b_{a_k a_\ell}\big)dt \\ &\quad + \big(\sigma_x Z_t^{k\ell} + \sigma_{xx}[Y_t^k, Y_t^\ell] + \sigma_{xa_\ell} Y_t^k + \sigma_{xa_k} Y_t^\ell + \sigma_{a_k a_\ell}\big)dW_t, \qquad Y_0^k = Z_0^{k\ell} = 0. \end{aligned}$$

For $p \ge 2$, Minkowski's inequality and the Burkholder–Davis–Gundy inequality give

$$\left\|\int_0^t F_s ds\right\|_{L^p} \le \int_0^t \|F_s\|_{L^p} ds, \qquad \left\|\int_0^t G_s dW_s\right\|_{L^p} \le C_p \left(\int_0^t \|G_s\|_{L^p}^2 ds\right)^{1/2}.$$

The coefficient derivatives are bounded, so for $t \le 1$,

$$\|Y_t^k\|_{L^p}^2 \le C_p \left(\int_0^t (1 + \|Y_s^k\|_{L^p}) ds\right)^2 + C_p \int_0^t (1 + \|Y_s^k\|_{L^p})^2 ds \le C_p t + C_p \int_0^t \|Y_s^k\|_{L^p}^2 ds.$$

Gronwall's inequality yields

$$\|Y_t^k\|_{L^p}^2 \le C_p t e^{C_p t} \le C_p t.$$

For $c = b$ or $c = \sigma$, Hölder's inequality and this estimate at exponent $2p$ give

$$\begin{aligned} &\|c_{xx}[Y_s^k, Y_s^\ell] + c_{xa_\ell} Y_s^k + c_{xa_k} Y_s^\ell + c_{a_k a_\ell}\|_{L^p} \\ &\quad \le C\big(\|Y_s^k\|_{L^{2p}} \|Y_s^\ell\|_{L^{2p}} + \|Y_s^k\|_{L^p} + \|Y_s^\ell\|_{L^p} + 1\big) \le C_p(s + \sqrt{s} + 1) \le C_p. \end{aligned}$$

Applying the integral inequalities to the equation for $Z^{k\ell}$ therefore gives

$$\|Z_t^{k\ell}\|_{L^p}^2 \le C_p t + C_p \int_0^t \|Z_s^{k\ell}\|_{L^p}^2 ds \le C_p t,$$

where the last inequality follows from Gronwall. Taking expectations in the two variational equations eliminates the stochastic integrals, and hence

$$\begin{aligned} \|\mathbb{E} Y_t^k\| &\le C \int_0^t (1 + \|Y_s^k\|_{L^2}) ds \le C \int_0^t (1 + \sqrt{s}) ds \le Ct, \\ \|\mathbb{E} Z_t^{k\ell}\| &\le C \int_0^t \big(1 + \|Z_s^{k\ell}\|_{L^2} + \|Y_s^k\|_{L^2} \|Y_s^\ell\|_{L^2} + \|Y_s^k\|_{L^2} + \|Y_s^\ell\|_{L^2}\big) ds \\ &\le C \int_0^t (1 + \sqrt{s} + s) ds \le Ct. \end{aligned}$$

The original SDE and linear coefficient growth similarly give

$$\|X_t - x\|_{L^p}^2 \le C_p \int_0^t (1 + \|X_s\|_{L^p} + \|a\|)^2 ds \le C_p t(1 + \|x\| + \|a\|)^2 + C_p \int_0^t \|X_s - x\|_{L^p}^2 ds.$$

Thus Gronwall and the coordinate derivative bounds imply

$$\begin{aligned} \|\partial_a X_h\|_{L^p} + \|\partial_a^2 X_h\|_{L^p} &\le C_p \sqrt{h}, \\ \|\mathbb{E}\partial_a X_h\| + \|\mathbb{E}\partial_a^2 X_h\| &\le Ch, \\ \|X_h - x\|_{L^p} &\le C_p \sqrt{h}(1 + \|x\| + \|a\|). \end{aligned}$$

We now use these estimates to bound the action derivatives of $P_h^a v_j(x)$. Adding and subtracting $\partial_x v_j(x)$ gives

$$\begin{aligned} \partial_a P_h^a v_j(x) &= \partial_x v_j(x) \mathbb{E}[\partial_a X_h] + \mathbb{E}[(\partial_x v_j(X_h) - \partial_x v_j(x))\partial_a X_h], \\ \partial_a^2 P_h^a v_j(x) &= \partial_x v_j(x) \mathbb{E}[\partial_a^2 X_h] + \mathbb{E}[(\partial_x v_j(X_h) - \partial_x v_j(x))\partial_a^2 X_h] + \mathbb{E}[\partial_x^2 v_j(X_h)[\partial_a X_h, \partial_a X_h]]. \end{aligned}$$

The Hessian bound and the fundamental theorem of calculus give

$$\partial_x v_j(y) - \partial_x v_j(x) = \int_0^1 \partial_x^2 v_j((1-t)x + ty)[y-x]dt,$$

$$\|\partial_x v_j(y) - \partial_x v_j(x)\| \le C\|y-x\| \int_0^1 (1 + (1-t)\|x\| + t\|y\|)dt \le C(1 + \|x\| + \|y\|)\|y-x\|.$$

Combining these formulas with Hölder's inequality yields

$$\|\partial_a P_h^a v_j(x)\| + \|\partial_a^2 P_h^a v_j(x)\| \le Chw(x)(1 + \|u\|^2), \qquad a = \mu_\phi(x) + u. \tag{44}$$

We finally average over $U \sim N(0, \sigma_e^2 I)$. Since $\mathbb{E}U = 0$, Taylor's formula gives

$$(P_{\phi,h}^{\sigma_e} - P_{\phi,h}^0)v_j(x) = \mathbb{E} \int_0^1 (1-s)\ \partial_a^2 P_h^a v_j(x)\big|_{a=\mu_\phi(x)+sU}\,[U,U]ds.$$

By (44),

$$|(P_{\phi,h}^{\sigma_e} - P_{\phi,h}^0)v_j(x)| \le Chw(x)\mathbb{E}[(1 + \|U\|^2)\|U\|^2] \le C\sigma_e^2 hw(x).$$

This proves the second estimate in (43); the first follows from (44) with $j = 0$, componentwise. □

B JOINT ACTOR CONVERGENCE AND CRITIC TRACKING

This appendix proves the joint convergence guarantee in Theorem 4.2. We first relate critic accuracy to policy-gradient accuracy, then combine this relation with the critic stability established in Appendix A to control actor stationarity and critic tracking along the same run.

B.1 FROM CRITIC ACCURACY TO POLICY-GRADIENT ACCURACY

For a fixed policy, the critic seeks coefficients at which its expected update vanishes. Corollary A.6 shows that there is a unique such vector, denoted by $z_\phi^\star$:

$$\bar{m}_\phi(z_\phi^\star) = 0.$$

We now ask how accurately this critic represents the policy's value and local action derivative. Assumption A.4(ii) provides reference coefficients $\widetilde{z}_\phi = (\widetilde{\theta}_\phi, \widetilde{\omega}_\phi)$ whose value and slope errors are at most e_V and e_Q. The theorem below bounds the distance from the fitted critic to these reference coefficients. The following corollary then translates this coefficient error into an error in the policy-gradient direction.

Theorem B.1 (Accuracy of the fitted critic)**.** *Under Assumptions A.1–A.4, for sufficiently small $h, \sigma_e > 0$, set*

$$\mathcal{E}_{\mathrm{ct}} = e_V + e_Q + \sigma_e^2 + h. \tag{45}$$

The four terms account for value approximation, slope approximation, exploration bias, and observation error, respectively. Then

$$\|z_\phi^\star - \widetilde{z}_\phi\| \le C_r \mathcal{E}_{\mathrm{ct}}. \tag{46}$$

For fixed h, σ_e, the map $\phi \mapsto z_\phi^\star$ is Lipschitz with a constant independent of h, σ_e. If $C_r\mathcal{E}_{\mathrm{ct}} \le r_{\mathcal{Z}}/2$, then $z_\phi^\star$ belongs to $\mathcal{Z}$ with boundary distance at least $r_{\mathcal{Z}}/2$. All constants are uniform over the allowed ϕ, h, σ_e.

Proof. Fix ϕ, h, σ_e. We first prove the coefficient-error bound (46), then verify the root's interiority and Lipschitz dependence on the actor. The coefficient error is controlled by the mean critic equation at the reference coefficients:

$$A_\phi(\widetilde{z}_\phi - z_\phi^\star) = \bar{m}_\phi(\widetilde{z}_\phi), \qquad \|\widetilde{z}_\phi - z_\phi^\star\| \le C\|W_\chi \bar{m}_\phi(\widetilde{z}_\phi)\|.$$

The inequality is the inverse bound in Corollary A.6. The block weight $W_\chi = \mathrm{diag}(I_{p_V}, \chi I_{p_Q})$ acts on the value and slope moments defined in (25), so

$$\|W_\chi \bar{m}_\phi(\widetilde{z}_\phi)\|^2 = \|\bar{m}_{\phi,V}(\widetilde{z}_\phi)\|^2 + \chi^2\|\bar{m}_{\phi,Q}(\widetilde{z}_\phi)\|^2. \tag{47}$$

It therefore suffices to estimate the two mean blocks separately; χ is fixed independently of h, σ_e. To see which residual estimates are needed, apply the conditional-expectation identity (26) before scaling the slope block:

$$\begin{bmatrix}\bar{m}_{\phi,V}(\widetilde{z}_\phi)\\ \bar{m}_{\phi,Q}(\widetilde{z}_\phi)\end{bmatrix} = \frac{1}{\delta}\sum_{i<L}\gamma^i \mathbb{E}\left[\begin{bmatrix}f(X_0)\\ S_i\end{bmatrix}\mathbb{E}[D_i(\widetilde{z}_\phi)\mid \mathcal{F}_i^A]\right], \quad S_i = \sigma_e^{-2}C(X_i)U_i. \tag{48}$$

Both blocks are built from the same conditional cell residual, with weights $f(X_0)$ and S_i. We bound these expectations and then combine them using (47).

Decompose the same residual for both critic equations. The scalar $D_i(z)$ is defined in (24). At the reference critic, the Markov property gives its conditional mean explicitly:

$$\mathbb{E}[D_i(\widetilde{z}_\phi)\mid \mathcal{F}_i^A] = V_{\widetilde{\theta}_\phi}(X_i) - \gamma P_h^{A_i} V_{\widetilde{\theta}_\phi}(X_i) - hr_i + h\widetilde{\omega}_\phi^\top C(X_i)U_i.$$

Write the reference value and slope errors as

$$\widetilde{v}(x) = V_{\widetilde{\theta}_\phi}(x) - V^\phi(x), \qquad e_g(x) = C(x)^\top \widetilde{\omega}_\phi - g^\phi(x).$$

By Assumption A.4(ii), $\|\widetilde{v}\|_{C_w^2} \le e_V$ and $\max_{i<L}\mathbb{E}\|e_g(X_i)\|^2 \le e_Q^2$. Two further errors arise even at the true value and slope. Define the one-cell observation error by

$$Q_h^\phi(x,a) := V^\phi(x) - \gamma P_h^a V^\phi(x) - h(\beta - \mathcal{L}^a)V^\phi(x), \tag{49}$$

and the nonlinear part of the action response by

$$\mathcal{R}^\phi(x,u) := q^\phi(x, \mu_\phi(x)+u) - g^\phi(x)^\top u. \tag{50}$$

The first measures the error of the generator approximation over a cell; the second measures the error of linearizing q^ϕ around the policy action. Substituting $V_{\widetilde{\theta}_\phi} = V^\phi + \widetilde{v}$ and $(\beta - \mathcal{L}^a)V^\phi = r(\cdot,a) - q^\phi(\cdot,a)$ into this conditional mean gives

$$\begin{aligned}\mathbb{E}[D_i(\widetilde{z}_\phi)\mid \mathcal{F}_i^A] = &\underbrace{\widetilde{v}(X_i) - \gamma P_h^{A_i}\widetilde{v}(X_i)}_{\text{value approximation}} + \underbrace{he_g(X_i)^\top U_i}_{\text{slope approximation}}\\ &- \underbrace{h\mathcal{R}^\phi(X_i,U_i)}_{\text{nonlinear action response}} + \underbrace{Q_h^\phi(X_i,A_i)}_{\text{observation error}}.\end{aligned} \tag{51}$$

We insert these four terms into (48) and estimate their contributions in that order for each block. The needed one-cell and Gaussian estimates are proved in Lemmas B.10 and B.11 at the end of this appendix. Throughout, the feature growth bounds and Lemma A.8 give

$$\|f(x)\| + \|C(x)\| \le Cw(x), \qquad \max_{0\le i\le L}\mathbb{E}w(X_i)^8 \le C.$$

These bounds justify all the weighted expectations below.

The value equation. For the value-approximation term, conditional expectation replaces $P_h^{A_i}\widetilde{v}(X_i)$ by $\widetilde{v}(X_{i+1})$ inside the expectation. The discounted sum then telescopes:

$$\begin{aligned}&\frac{1}{\delta}\sum_{i<L}\gamma^i\mathbb{E}[f(X_0)\{\widetilde{v}(X_i) - \gamma P_h^{A_i}\widetilde{v}(X_i)\}]\\ &\quad = \frac{1}{\delta}\mathbb{E}[f(X_0)\{\widetilde{v}(X_0) - \gamma^L\widetilde{v}(X_L)\}],\\ &\frac{1}{\delta}\left\|\mathbb{E}[f(X_0)\{\widetilde{v}(X_0) - \gamma^L\widetilde{v}(X_L)\}]\right\| \le \frac{Ce_V}{\delta}\mathbb{E}[w(X_0)\{w(X_0)+w(X_L)\}] \le Ce_V.\end{aligned}$$

The last step uses $\delta \ge \delta_{\min}$. The slope-approximation term vanishes because $f(X_0)$ and $e_g(X_i)$ are known before U_i is drawn:

$$\mathbb{E}[f(X_0)e_g(X_i)^\top U_i \mid \mathcal{F}_i^-] = f(X_0)e_g(X_i)^\top \mathbb{E}[U_i \mid \mathcal{F}_i^-] = 0.$$

For the nonlinear action term, Lemma B.11 gives $\mathbb{E}[|\mathcal{R}^\phi(X_i,U_i)| \mid \mathcal{F}_i^-] \le C\sigma_e^2 w(X_i)$. For the observation term, Lemma B.10 gives $|Q_h^\phi(X_i,A_i)| \le Ch^2 w(X_i)^2(1+\|U_i\|^4)$. Consequently,

$$\|\mathbb{E}[f(X_0)\mathcal{R}^\phi(X_i,U_i)]\| \le C\sigma_e^2\mathbb{E}[\|f(X_0)\|w(X_i)] \le C\sigma_e^2,$$
$$\|\mathbb{E}[f(X_0)Q_h^\phi(X_i,A_i)]\| \le Ch^2\mathbb{E}[\|f(X_0)\|w(X_i)^2(1+\|U_i\|^4)] \le Ch^2.$$

Insert these four estimates into the value block of (48). Since $Lh=\delta$,

$$\|\bar{m}_{\phi,V}(\widetilde{z}_\phi)\| \le Ce_V + \frac{C}{\delta}\sum_{i<L}\gamma^i(h\sigma_e^2+h^2) \le C(e_V+\sigma_e^2+h). \tag{52}$$

The slope equation. We now use the weight S_i in (48). For the value-approximation term, conditional Gaussian integration by parts (33) removes the inverse exploration factor:

$$\mathbb{E}[S_i\{\widetilde{v}(X_i)-\gamma P_h^{A_i}\widetilde{v}(X_i)\} \mid \mathcal{F}_i^-] = -\gamma C(X_i)\mathbb{E}_U[\nabla_a P_h^a \widetilde{v}(X_i)|_{a=\mu_\phi(X_i)+U}].$$

The first bound in Lemma B.10 uses only $\|\widetilde{v}\|_{C_w^2} \le e_V$ and yields

$$\left\|\mathbb{E}[S_i\{\widetilde{v}(X_i)-\gamma P_h^{A_i}\widetilde{v}(X_i)\}]\right\| \le Che_V\mathbb{E}[\|C(X_i)\|(1+\|X_i\|+\|A_i\|)] \le Che_V.$$

For slope approximation, the Gaussian covariance gives the exact conditional identity, followed by Cauchy–Schwarz:

$$\mathbb{E}[S_i e_g(X_i)^\top U_i \mid \mathcal{F}_i^-] = C(X_i)e_g(X_i),$$
$$\|\mathbb{E}[S_i h e_g(X_i)^\top U_i]\| \le h(\mathbb{E}\|C(X_i)\|^2)^{1/2}(\mathbb{E}\|e_g(X_i)\|^2)^{1/2} \le Che_Q.$$

For the nonlinear action term, the second bound in Lemma B.11 gives

$$\mathbb{E}[S_i\mathcal{R}^\phi(X_i,U_i) \mid \mathcal{F}_i^-] = C(X_i)\mathbb{E}_U[\sigma_e^{-2}U\mathcal{R}^\phi(X_i,U)],$$
$$\|\mathbb{E}[S_i h\mathcal{R}^\phi(X_i,U_i)]\| \le Ch\sigma_e^2\mathbb{E}[\|C(X_i)\|w(X_i)] \le Ch\sigma_e^2.$$

Finally, the observation term uses Gaussian integration by parts and the action-derivative bound on Q_h^ϕ in Lemma B.10:

$$\mathbb{E}[S_i Q_h^\phi(X_i,A_i) \mid \mathcal{F}_i^-] = C(X_i)\mathbb{E}_U[\nabla_a Q_h^\phi(X_i,a)|_{a=\mu_\phi(X_i)+U}],$$
$$\|\mathbb{E}[S_i Q_h^\phi(X_i,A_i)]\| \le Ch^2\mathbb{E}[\|C(X_i)\|w(X_i)^2(1+\|U_i\|^4)] \le Ch^2.$$

The two integration-by-parts arguments bound the norm of the expected score-weighted error; taking absolute values before averaging would lose the cancellation. Summing the four contributions in (48) now gives

$$\|\bar{m}_{\phi,Q}(\widetilde{z}_\phi)\| \le \frac{C}{\delta}\sum_{i<L}\gamma^i(he_V+he_Q+h\sigma_e^2+h^2) \le C(e_V+e_Q+\sigma_e^2+h).$$

We now combine this slope bound with the value bound (52). The block identity (47) gives

$$\begin{aligned}\|W_\chi\bar{m}_\phi(\widetilde{z}_\phi)\| &\le \|\bar{m}_{\phi,V}(\widetilde{z}_\phi)\| + \chi\|\bar{m}_{\phi,Q}(\widetilde{z}_\phi)\| \\ &\le C(e_V+\sigma_e^2+h) + \chi C(e_V+e_Q+\sigma_e^2+h) \le C\mathcal{E}_{\mathrm{ct}}.\end{aligned} \tag{53}$$

The inverse bound at the start of the proof therefore gives $\|z_\phi^\star - \widetilde{z}_\phi\| \le C_r\mathcal{E}_{\mathrm{ct}}$, which is (46).

Interiority and movement of the fitted critic. The reference coefficients are at least $r_{\mathcal{Z}}$ from the boundary. If $C_r\mathcal{E}_{\mathrm{ct}} \le r_{\mathcal{Z}}/2$, the coefficient-error bound places $z_\phi^\star$ inside $\mathcal{Z}$ and gives

$$\mathrm{dist}(z_\phi^\star,\partial\mathcal{Z}) \ge \mathrm{dist}(\widetilde{z}_\phi,\partial\mathcal{Z}) - \|z_\phi^\star-\widetilde{z}_\phi\| \ge r_{\mathcal{Z}}/2.$$

For Lipschitz dependence, keep h,σ_e fixed and subtract the two root equations $A_\phi z_\phi^\star = b_\phi$ and $A_{\phi'}z_{\phi'}^\star = b_{\phi'}$:

$$A_\phi(z_\phi^\star - z_{\phi'}^\star) = (b_\phi - b_{\phi'}) - (A_\phi - A_{\phi'})z_{\phi'}^\star.$$

The uniform inverse and root bounds in Corollary A.6, together with (29), yield

$$\|z_\phi^\star - z_{\phi'}^\star\| \le C\{\|b_\phi - b_{\phi'}\| + \|A_\phi - A_{\phi'}\|\,\|z_{\phi'}^\star\|\} \le L_z\|\phi-\phi'\|.$$

Here L_z is independent of h,σ_e, as required. □

AdvG uses the critic's slope $C^\top\omega$ to choose the actor's update direction. Its expected direction is

$$\bar{H}_{\phi,h}(z) = \frac{h}{1-\gamma}\mathbb{E}_{d^0_{\phi,h}}[\partial_\phi \mu_\phi^\top C^\top \omega], \qquad z = (\theta, \omega).$$

The next result separates the effect of an unfinished critic fit from the approximation, exploration, and observation errors already bounded in Theorem B.1.

Corollary B.2 (Accuracy of the policy-gradient direction)**.** *Under the hypotheses of Theorem B.1, for every deterministic $z = (\theta, \omega)$,*

$$\|\bar{H}_{\phi,h}(z) - \nabla J(\phi)\| \le C_H \|z - z^\star_\phi\| + C_E \mathcal{E}_{\mathrm{ct}}. \tag{54}$$

The constants are uniform over the allowed ϕ, h, σ_e.

Proof. Fix ϕ, h, σ_e and write $d = d^0_{\phi,h}$. To prove (54), we separate the error in the expected actor direction into four terms:

$$\begin{aligned}\bar{H}_{\phi,h}(z) - \nabla J(\phi) = &\underbrace{\bar{H}_{\phi,h}(z) - \bar{H}_{\phi,h}(z^\star_\phi)}_{\text{critic tracking}} \\ &+ \underbrace{\bar{H}_{\phi,h}(z^\star_\phi) - \bar{H}_{\phi,h}(\widetilde{z}_\phi)}_{\text{fitted-critic bias}} \\ &+ \underbrace{\frac{h}{1-\gamma}\mathbb{E}_d[\partial_\phi \mu_\phi^\top (C^\top \widetilde{\omega}_\phi - g^\phi)]}_{\text{slope approximation}} \\ &+ \underbrace{\frac{h}{1-\gamma}\mathbb{E}_d[\partial_\phi \mu_\phi^\top g^\phi] - \nabla J(\phi)}_{\text{state sampling}}.\end{aligned} \tag{55}$$

The first term vanishes at the fitted critic. The next two compare that fitted critic with the reference slope and then with the true slope. The last term remains even with the true slope because the algorithm samples states at observation times.

The three critic-dependent terms. The actor direction is linear in the slope coefficients. For any $z_j = (\theta_j, \omega_j)$, the feature and policy-derivative moments and the bounded ratio $h/(1-\gamma)$ give

$$\begin{aligned}\|\bar{H}_{\phi,h}(z_1) - \bar{H}_{\phi,h}(z_2)\| &\le \frac{h}{1-\gamma}\mathbb{E}_d[\|\partial_\phi \mu_\phi\|\,\|C\|]\,\|\omega_1 - \omega_2\| \\ &\le C\|z_1 - z_2\|.\end{aligned}$$

Apply this bound to the first two terms of (55), using Theorem B.1 for the second. For the third, use Cauchy–Schwarz and the slope-approximation bound (28) at $X_0 \sim d$. These give, in order,

$$\begin{aligned}\|\bar{H}_{\phi,h}(z) - \bar{H}_{\phi,h}(z^\star_\phi)\| &\le C\|z - z^\star_\phi\|, \\ \|\bar{H}_{\phi,h}(z^\star_\phi) - \bar{H}_{\phi,h}(\widetilde{z}_\phi)\| &\le C\|z^\star_\phi - \widetilde{z}_\phi\| \le C\mathcal{E}_{\mathrm{ct}}, \\ \left\|\frac{h}{1-\gamma}\mathbb{E}_d[\partial_\phi \mu_\phi^\top (C^\top \widetilde{\omega}_\phi - g^\phi)]\right\| &\le \frac{h}{1-\gamma}(\mathbb{E}_d\|\partial_\phi \mu_\phi\|^2)^{1/2}(\mathbb{E}_d\|C^\top \widetilde{\omega}_\phi - g^\phi\|^2)^{1/2} \le Ce_Q.\end{aligned}$$

The state-sampling term. To compare the true-slope average with $\nabla J(\phi)$, differentiate the evaluation equation in Assumption A.3:

$$(\beta - \mathcal{L}^{\mu_\phi})\nabla_\phi V^\phi = \partial_\phi \mu_\phi^\top g^\phi, \qquad \nu_0 \nabla_\phi V^\phi = \nabla J(\phi).$$

The second identity follows by differentiating $J(\phi) = \nu_0 V^\phi$ under the initial-state expectation. For continuous policy feedback, discounted Dynkin's formula and the vanishing terminal term give

$$\nabla J(\phi) = \int \partial_\phi \mu_\phi^\top g^\phi \, d\rho_\phi,$$

which also proves (6). For the algorithm's sampled states, we instead use the frozen-action transition $P^0_{\phi,h}$. Apply the C^4_w observation-error bound in Lemma B.10 componentwise to $\nabla_\phi V^\phi$, with $a = \mu_\phi(x)$. The preceding evaluation identity then gives

$$\left\|\nabla_\phi V^\phi(x) - \gamma P^0_{\phi,h}\nabla_\phi V^\phi(x) - h\partial_\phi\mu_\phi(x)^\top g^\phi(x)\right\| \le Ch^2 w(x)^2.$$

The sampling law (23) satisfies $d(I - \gamma P^0_{\phi,h}) = (1-\gamma)\nu_0$. Hence the last term in (55) is exactly

$$\begin{aligned}&\frac{h}{1-\gamma}\mathbb{E}_d[\partial_\phi\mu_\phi^\top g^\phi] - \nabla J(\phi)\\ &\qquad = -\frac{1}{1-\gamma}\mathbb{E}_d[\nabla_\phi V^\phi - \gamma P^0_{\phi,h}\nabla_\phi V^\phi - h\partial_\phi\mu_\phi^\top g^\phi].\end{aligned}$$

The one-cell bound, $\mathbb{E}_d w^2 \le C$, and $1-\gamma \asymp h$ now yield

$$\left\|\frac{h}{1-\gamma}\mathbb{E}_d[\partial_\phi\mu_\phi^\top g^\phi] - \nabla J(\phi)\right\| \le \frac{Ch^2}{1-\gamma} \le Ch. \tag{56}$$

Combining the four bounds in (55), we obtain

$$\|\bar H_{\phi,h}(z) - \nabla J(\phi)\| \le C\|z - z^\star_\phi\| + C\mathcal{E}_{\rm ct} + Ce_Q + Ch \le C_H\|z - z^\star_\phi\| + C_E\mathcal{E}_{\rm ct}.$$

This proves (54) because $e_Q + h \le \mathcal{E}_{\rm ct}$. □

B.2 Joint convergence bound and physical-time rates

Appendix A establishes critic stability, and Section B.1 bounds the policy-gradient error at a fixed policy. We now prove that AdvG can learn the policy and critic together. This requires controlling two interacting errors: changing the policy moves the coefficients that solve the expected critic equation, while an inaccurate critic perturbs the actor's update direction. Theorem B.4 bounds both the critic's tracking error and the actor's stationarity error over the same run. Choosing the sampling and step sizes then gives the physical-time rate in Theorem 4.2.

The proof combines contraction of the critic error with improvement in expected return. The actor's improvement bound includes a positive multiple of its squared step. Keeping this term lets us control the movement of the critic's fitting target, so a sufficiently small fixed ratio of actor to critic step sizes is enough. Both updates may use the same batch: the argument requires independence across fresh trajectories, not between the actor and critic updates.

Write $\widehat H_k = B^{-1}\sum_{b=1}^B H_{k,b}$ for the average actor direction from the fresh batch at iteration k. Both updates use the pre-update parameters and the same trajectories:

$$z_{k+1} = \Pi_{\mathcal{Z}}(z_k - \eta W_\chi \widehat m_k), \qquad \phi_{k+1} = \Pi_\Phi(\phi_k + \alpha\widehat H_k). \tag{57}$$

The critic average is defined in Algorithm 1; both endpoints of each cell use V_{θ_k}. Here Π_S is Euclidean projection onto a closed convex set S. For fixed $\lambda > 0$, the projected gradient is

$$\mathcal{G}_\lambda(\phi) = \lambda^{-1}\{\Pi_\Phi(\phi + \lambda\nabla J(\phi)) - \phi\}.$$

Projection optimality gives $\mathcal{G}_\lambda(\phi) = 0$ exactly when $\langle\nabla J(\phi), u - \phi\rangle \le 0$ for every $u \in \Phi$. Thus its norm measures constrained stationarity. AdvG steps in an estimated direction, so we need to relate its actual step to this exact-gradient measure. The following inequality does so while retaining part of the improvement as a squared-step term. That term will control how far the actor moves the critic's target.

Lemma B.3. *Suppose J has an L_J-Lipschitz gradient on Φ. For $\phi \in \Phi$ and any direction $\widehat H$, set $d = \widehat H - \nabla J(\phi)$ and $\Delta = \Pi_\Phi(\phi + \alpha\widehat H) - \phi$. If $0 < \alpha \le \min\{\lambda, (2L_J)^{-1}\}$, then*

$$J(\phi + \Delta) - J(\phi) \ge \frac{\alpha}{8}\|\mathcal{G}_\lambda(\phi)\|^2 + \frac{\|\Delta\|^2}{4\alpha} - \frac{5\alpha}{4}\|d\|^2. \tag{58}$$

Proof. First compare exact projected gradient steps at the algorithm's step size α and the fixed reference size λ. Put $u = \mathcal{G}_\alpha(\phi)$ and $v = \mathcal{G}_\lambda(\phi)$. Adding their projection inequalities gives

$$0 \ge \langle u - v, \alpha u - \lambda v\rangle \ge (\|u\| - \|v\|)(\alpha\|u\| - \lambda\|v\|).$$

Since $\alpha \le \lambda$, this excludes $\|v\| > \|u\|$. Next compare the actual step Δ with the exact step αu. Nonexpansiveness of projection gives

$$\|\Delta - \alpha u\| \le \alpha\|d\|, \qquad \alpha^2\|v\|^2 \le 2\|\Delta\|^2 + 2\alpha^2\|d\|^2.$$

The projection defining Δ also gives $\langle \widehat{H}, \Delta\rangle \ge \|\Delta\|^2/\alpha$. Using smoothness of J, followed by Young's inequality and $\alpha L_J \le 1/2$, we obtain

$$J(\phi+\Delta) - J(\phi) \ge (\alpha^{-1} - L_J/2)\|\Delta\|^2 - \langle d, \Delta\rangle \ge \frac{\|\Delta\|^2}{2\alpha} - \alpha\|d\|^2.$$

Split the positive term into two equal parts. The preceding comparison of v and Δ bounds one part below by $\alpha\|v\|^2/8 - \alpha\|d\|^2/4$; keeping the other as $\|\Delta\|^2/(4\alpha)$ proves (58). Thus one part of the return gain measures stationarity and the other controls the actor's movement. □

We need the following bounds for the convergence proof. Let $\Phi, \mathcal{Z}$ be compact convex sets, and let J have an L_J-Lipschitz gradient on a neighborhood of Φ. For each policy, let $z^\star_\phi$ be the coefficients that solve the expected critic equation $\bar{m}_\phi(z^\star_\phi) = 0$. We use the critic stability bounds (34), assume this fitted critic belongs to $\mathcal{Z}$ and is L_z-Lipschitz in ϕ, and use (54) in the form

$$\|\bar{H}_{\phi,h}(z) - \nabla J(\phi)\| \le C_H\|z - z^\star_\phi\| + C_E\mathcal{E}_{\mathrm{ct}}, \qquad \mathcal{E}_{\mathrm{ct}} = e_V + e_Q + \sigma_e^2 + h.$$

The first term bounds the error from an imperfectly fitted critic. The second bounds the error that can remain even at the fitted coefficients: it accounts for feature approximation, exploration around the policy action, and discrete observations. Keeping these errors separate lets us study critic tracking without assuming an exact critic at each iteration. For $z \in \mathcal{Z}$, the one-trajectory variances satisfy

$$\begin{aligned} \mathbb{E}\|W_\chi\{m(z;\Xi) - \bar{m}(z)\}\|^2 &\le C_m\Sigma^2_{\sigma_e,\delta}, \\ \mathbb{E}\|H_{\phi,h}(z;X_0) - \bar{H}_{\phi,h}(z)\|^2 &\le \sigma_a^2, \end{aligned} \qquad \Sigma^2_{\sigma_e,\delta} = (\delta\sigma_e^2)^{-1}.$$

All constants are uniform over the allowed policies and sufficiently small h, σ_e.

Under Assumptions A.1–A.4, Corollaries A.6–A.7, Theorem B.1, and Corollary B.2 supply these bounds whenever $C_r\mathcal{E}_{\mathrm{ct}} \le r_{\mathcal{Z}}/2$. For the actor variance, compactness of $\mathcal{Z}$ and the feature and state moments give

$$\mathbb{E}\|H_{\phi,h}(z;X_0)\|^2 \le C\left(\frac{h}{1-\gamma}\right)^2 \mathbb{E}[\|\partial_\phi\mu_\phi(X_0)\|^2\|C(X_0)\|^2] \le C.$$

The convergence argument only needs the fitted critic to belong to $\mathcal{Z}$; the boundary margin above is one sufficient condition. Section B.3 verifies the same bounds directly for an LQG family.

Theorem B.4 (Joint actor and critic convergence)**.** *Assume the critic stability, actor approximation, and variance bounds stated above. Initialize $(\phi_0, z_0) \in \Phi \times \mathcal{Z}$. Keep h, L, σ_e fixed within a run and $\delta = Lh \in [\delta_{\min}, \delta_{\max}] \subset (0,\infty)$. At each of K iterations, collect B fresh, conditionally independent complete trajectories at the current actor and apply (57). There are uniform $\varpi_0, \eta_0 > 0$ such that, for any fixed $\varpi \in (0, \varpi_0]$ and*

$$\alpha = \varpi\eta \le \lambda, \qquad 0 < \eta \le \eta_0, \tag{59}$$

$$\frac{1}{K}\sum_{k<K}\mathbb{E}\left[\|\mathcal{G}_\lambda(\phi_k)\|^2 + \|z_k - z^\star_{\phi_k}\|^2\right] \le C_\varpi\left[\frac{1 + \|z_0 - z^\star_{\phi_0}\|^2}{K\eta} + \frac{\eta\Sigma^2_{\sigma_e,\delta}}{B} + \mathcal{E}_{\mathrm{ct}}^2 + \frac{\sigma_a^2}{B}\right]. \tag{60}$$

The constant depends only on ϖ, the fixed bounds above, and the return range on Φ; it is independent of $h, \sigma_e, L, \delta, K, B, \eta$ and the initial iterates.

Proof. Let $\mathcal{F}_k$ be the history before batch k, and write $\mathbb{E}_k = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. We track the distance from the current critic to the coefficients fitted at the current policy:

$$z^\star_k = z^\star_{\phi_k}, \qquad e_k = z_k - z^\star_k, \qquad D_k = \mathbb{E}_k\|\phi_{k+1} - \phi_k\|^2.$$

Here D_k is the conditional mean squared actor step. It controls how far the next policy can move the fitted critic. Conditional independence of trajectories and the preceding bounds give, with $\zeta_k = W_\chi\{\widehat{m}_k - \bar{m}_{\phi_k}(z_k)\}$,

$$\begin{aligned} \mathbb{E}_k\zeta_k = 0, \qquad \mathbb{E}_k\|\zeta_k\|^2 \le \frac{C_m\Sigma^2_{\sigma_e,\delta}}{B}, \\ \mathbb{E}_k\|\widehat{H}_k - \nabla J(\phi_k)\|^2 \le 2C_H^2\|e_k\|^2 + 2C_E^2\mathcal{E}_{\mathrm{ct}}^2 + \frac{\sigma_a^2}{B}. \end{aligned} \tag{61}$$

For the actor, decompose the error as $(\widehat{H}_k - \bar{H}_{\phi_k,h}(z_k)) + (\bar{H}_{\phi_k,h}(z_k) - \nabla J(\phi_k))$. The first term has conditional mean zero, so its cross term with the second vanishes. Its variance is reduced by B, whereas the second term is controlled by critic tracking and $\mathcal{E}_{\mathrm{ct}}$. This reasoning does not require independent actor and critic estimates.

Critic tracking and actor progress. First hold the fitting target at $z_k^\star$. Since it belongs to $\mathcal{Z}$, projection cannot increase the distance to it. The critic noise ζ_k has conditional mean zero, so expansion of the squared update gives

$$\mathbb{E}_k\|z_{k+1} - z_k^\star\|^2 \le \|e_k - \eta W_\chi \bar{m}_{\phi_k}(z_k)\|^2 + \eta^2\mathbb{E}_k\|\zeta_k\|^2.$$

The expected equation vanishes at $z_k^\star$. Its positivity and Lipschitz bounds in (34) therefore imply, for $\eta \le \mu_\chi/L_\chi^2$,

$$\mathbb{E}_k\|z_{k+1} - z_k^\star\|^2 \le (1 - \mu_\chi\eta)\|e_k\|^2 + \frac{C_m\eta^2\Sigma^2_{\sigma_e,\delta}}{B}. \tag{62}$$

This is contraction toward the old target. The next tracking error instead uses the new target $z_{k+1}^\star$, which moves with the actor. Apply Young's inequality to $e_{k+1} = (z_{k+1} - z_k^\star) - (z_{k+1}^\star - z_k^\star)$, using $\mathbb{E}_k\|z_{k+1}^\star - z_k^\star\|^2 \le L_z^2 D_k$. For $\eta \le \min\{1, (2\mu_\chi)^{-1}\}$,

$$\begin{aligned} \mathbb{E}_k\|e_{k+1}\|^2 &\le (1 + \mu_\chi\eta/4)\mathbb{E}_k\|z_{k+1} - z_k^\star\|^2 + (1 + 4/(\mu_\chi\eta))L_z^2 D_k \\ &\le (1 - \mu_\chi\eta/2)\|e_k\|^2 + \frac{2C_m\eta^2\Sigma^2_{\sigma_e,\delta}}{B} + \frac{(1 + 4/\mu_\chi)L_z^2}{\eta}D_k. \end{aligned} \tag{63}$$

The final term is the cost of moving the fitting target. Young's inequality bounds it even when both updates use the same batch. To control that cost, we now use the actor's return improvement. Lemma B.3 and (61) give

$$\mathbb{E}_k[J(\phi_{k+1}) - J(\phi_k)] \ge \frac{\alpha}{8}\|\mathcal{G}_\lambda(\phi_k)\|^2 + \frac{D_k}{4\alpha} - \alpha\left(\frac{5}{2}C_H^2\|e_k\|^2 + \frac{5}{2}C_E^2\mathcal{E}_{\mathrm{ct}}^2 + \frac{5\sigma_a^2}{4B}\right). \tag{64}$$

The two inequalities show both directions of the interaction. Critic error reduces the actor's progress by a multiple of $\alpha\|e_k\|^2$, while actor movement increases the critic error by a multiple of D_k/η. The critic contraction supplies $\mu_\chi\eta\|e_k\|^2/2$ to cover the first cost; the actor supplies $D_k/(4\alpha)$ to cover the second. Both comparisons require only that α/η be sufficiently small.

The step-size restrictions below use five fixed constants from the preceding bounds. The critic constants μ_χ and L_χ are the strong-monotonicity and Lipschitz constants in (34), supplied by Corollary A.6. The constant C_H multiplies the critic tracking error in (54), as proved in Corollary B.2. The constant L_z satisfies $\|z_\phi^\star - z_{\phi'}^\star\| \le L_z\|\phi - \phi'\|$, as established in the proof of Theorem B.1, and L_J is the Lipschitz constant of ∇J in Assumption A.3. For fixed critic weight χ and window bounds $0 < \delta_{\min} \le \delta \le \delta_{\max} < \infty$, these constants are uniform over the allowed actors and sufficiently small h, σ_e; they do not depend on K, B or the learning step sizes. We now choose, omitting restrictions with zero denominators,

$$\begin{aligned} \varpi_0 &\le \min\left\{1, \frac{\mu_\chi}{10C_H^2}, \frac{1}{8(1 + 4/\mu_\chi)L_z^2}\right\}, \\ \eta_0 &\le \min\left\{1, \frac{1}{2\mu_\chi}, \frac{\mu_\chi}{L_\chi^2}, \frac{1}{2L_J}\right\}. \end{aligned}$$

These choices ensure the lemma's step restriction and

$$\frac{\mu_\chi\eta}{2} - \frac{5}{2}\alpha C_H^2 \ge \frac{\mu_\chi\eta}{4}, \qquad \frac{1}{4\alpha} - \frac{(1 + 4/\mu_\chi)L_z^2}{\eta} \ge \frac{1}{8\alpha}.$$

These inequalities leave a positive contribution for both critic tracking and actor movement after the two costs are absorbed. To combine them, use a quantity that increases when the return improves or the critic error decreases: $\Psi_k = J(\phi_k) - \|e_k\|^2$. Subtract the bound on $\mathbb{E}_k\|e_{k+1}\|^2 - \|e_k\|^2$ from the return-improvement bound (64), and discard the remaining nonnegative squared-step term. This gives

$$\mathbb{E}_k[\Psi_{k+1} - \Psi_k] \geq \eta \min\{\varpi/8, \mu_\chi/4\}\{\|\mathcal{G}_\lambda(\phi_k)\|^2 + \|e_k\|^2\} - \alpha\left(\frac{5}{2}C_E^2\mathcal{E}_{\mathrm{ct}}^2 + \frac{5\sigma_a^2}{4B}\right) - \frac{2C_m\eta^2\Sigma_{\sigma_e,\delta}^2}{B}. \tag{65}$$

Let $\Delta_J = \max_\Phi J - \min_\Phi J$ be the return range. It is finite because J is continuous on the fixed compact actor set. Since $\Psi_K - \Psi_0 \leq \Delta_J + \|e_0\|^2$, taking expectations and summing (65) gives

$$\frac{1}{K}\sum_{k<K}\mathbb{E}\big[\|\mathcal{G}_\lambda(\phi_k)\|^2 + \|e_k\|^2\big] \leq \frac{1}{\min\{\varpi/8, \mu_\chi/4\}}\left[\frac{\Delta_J + \|e_0\|^2}{K\eta} + \frac{2C_m\eta\Sigma_{\sigma_e,\delta}^2}{B} + \frac{5}{2}\varpi C_E^2\mathcal{E}_{\mathrm{ct}}^2 + \frac{5\varpi\sigma_a^2}{4B}\right].$$

Because $\varpi \leq 1$, an explicit constant in (60) is

$$C_\varpi = \frac{\max\{1, \Delta_J, 2C_m, \frac{5}{2}C_E^2, \frac{5}{4}\}}{\min\{\varpi/8, \mu_\chi/4\}}. \tag{66}$$

Indeed, $\max\{1, \Delta_J\}(1 + \|e_0\|^2)$ bounds the initial term's numerator, and the other entries dominate the three error coefficients. Moreover, $\min\{\varpi/8, \mu_\chi/4\}^{-1} \leq 8/\varpi + 4/\mu_\chi$, so this choice satisfies $C_\varpi \leq C(1 + \varpi^{-1})$ with C independent of the sampling parameters. This proves (60). □

The four terms in (60) have distinct meanings. The first decreases as $K\eta$ grows; the second is the critic's sampling noise after averaging a batch; the third is the systematic error in the actor direction even at a fitted critic; and the fourth is the actor's sampling noise. In particular, tracking the fitted critic does not by itself remove exploration, observation, or feature error. Section B.1 is what relates that fit to the true policy gradient.

Remark B.5. If R is independent and uniform on $\{0, \ldots, K-1\}$, the left side of (60) is exactly the expected joint error at (ϕ_R, z_R). Thus the theorem also gives a guarantee for a randomly selected iterate.

To express the convergence bound in physical time, we count the acquisition cost of a complete trajectory. Each trajectory contains the target-policy prefix from Algorithm 1 and one exploratory window. Its geometric prefix length has mean $\gamma/(1-\gamma)$, so the expected physical duration is

$$\tau_h = \frac{h\gamma}{1-\gamma} + \delta = \Theta(\beta^{-1} + \delta). \tag{67}$$

Both parts count toward the budget. With fixed discount and window bounds, τ_h is bounded above and away from zero for sufficiently small h, so K batches of B trajectories cost $KB\tau_h = \Theta(KB)$ in expected physical time. Throughout, δ remains positive as h decreases; finer observations increase the number of cells in a window.

Corollary B.6 (Convergence in physical time)**.** *Under Theorem B.4's hypotheses, assume exact features and fixed β, δ. For a sufficiently large expected physical-time budget T, choose, up to integer rounding,*

$$B \asymp T^{2/5}, \quad K \asymp T^{3/5}, \quad \eta \asymp T^{-1/5}, \quad \sigma_e \asymp T^{-1/10}, \quad h \asymp T^{-1/5}, \tag{68}$$

with $KB\tau_h \leq T$ and $\alpha = \varpi\eta$ for a fixed admissible ϖ. Then the joint squared error in (60) is $O(T^{-2/5})$.

Proof. Using a fixed fraction of the budget gives $K = \Theta(T/B)$ because each complete trajectory has expected duration bounded above and away from zero. With exact features,

$$\mathcal{E}_{\mathrm{ct}}^2 = (\sigma_e^2 + h)^2 \leq 2(\sigma_e^4 + h^2), \qquad \Sigma_{\sigma_e,\delta}^2 = O(\sigma_e^{-2}).$$

Substituting into (60), and using boundedness of the initial critic error, reduces the rate calculation to

$$C\left\{\frac{B}{T\eta}+\frac{\eta}{B\sigma_e^2}+\frac{1}{B}+\sigma_e^4+h^2\right\}.$$

Under (68),

$$\frac{B}{T\eta}\asymp\frac{\eta}{B\sigma_e^2}\asymp\frac{1}{B}\asymp\sigma_e^4\asymp h^2\asymp T^{-2/5}.$$

Larger batches reduce sampling noise but leave fewer updates within the budget. Smaller action perturbations reduce exploration bias while increasing the critic variance. The choices of B,η,σ_e balance these contributions, and $h\asymp T^{-1/5}$ brings the observation error h^2 to the same order.

For completeness, the schedule can satisfy the budget with integer numbers of trajectories and cells. For fixed $c_h,c_B>0$, take $L=\lceil\delta T^{1/5}/c_h\rceil$, $h=\delta/L$, $B=\lceil c_B T^{2/5}\rceil$, and

$$K=\left\lfloor\frac{T}{B\tau_h}\right\rfloor,\qquad KB\tau_h\le T,\qquad KB\tau_h=\Theta(T). \tag{69}$$

The last relation holds for large T because $B\tau_h=O(T^{2/5})=o(T)$, so rounding loses a vanishing fraction of the budget. Choose the fixed step constants to satisfy (59). Every term in the displayed error bound is then $O(T^{-2/5})$. □

Corollary B.7 (Convergence with feature approximation)**.** *Under the hypotheses of Theorem B.4, fix $\beta,\delta>0$. Assume the approximation errors are sufficiently small that the fitted roots remain in $\mathcal{Z}$, uniformly over the actors and sufficiently large T. For fixed $c_B,c_h>0$ and sufficiently small $c_\eta>0$, choose*

$$\begin{aligned}&\sigma_e=T^{-1/10},\qquad B=\lceil c_B T^{2/5}\rceil,\quad \eta=c_\eta T^{-1/5},\\ &L=\lceil\delta T^{1/5}/c_h\rceil,\quad h=\delta/L,\qquad K=\left\lfloor\frac{T}{B\tau_h}\right\rfloor,\end{aligned} \tag{70}$$

where τ_h is given by (67), and take $\alpha=\varpi\eta$ for a fixed admissible $\varpi>0$. Then the expected acquisition cost is $KB\tau_h\le T$, and

$$\frac{1}{K}\sum_{k<K}\mathbb{E}\left[\|\mathcal{G}_\lambda(\phi_k)\|^2+\|z_k-z^\star_{\phi_k}\|^2\right]\le C\left[T^{-2/5}+(e_V+e_Q)^2\right]. \tag{71}$$

The constant is independent of T. The schedule has the same orders as (68), including when e_V,e_Q are nonzero.

Proof. The schedule gives $h\le c_h T^{-1/5}$ and $\sigma_e^2=T^{-1/5}$, so

$$\begin{aligned}&\mathcal{E}_{\mathrm{ct}}=e_V+e_Q+\sigma_e^2+h\le e_V+e_Q+(1+c_h)T^{-1/5},\\ &\mathcal{E}_{\mathrm{ct}}^2\le C\left[(e_V+e_Q)^2+T^{-2/5}\right].\end{aligned}$$

In the setting of Assumptions A.1–A.4, sufficiently small fixed errors, for example $C_r(e_V+e_Q)\le r_{\mathcal{Z}}/4$, give

$$C_r\mathcal{E}_{\mathrm{ct}}\le r_{\mathcal{Z}}/4+C_r(1+c_h)T^{-1/5}\le r_{\mathcal{Z}}/2$$

for large T. Theorem B.1 therefore ensures the root requirement. The exploration scale is admissible for large T, and sufficiently small c_η ensures (59).

Since τ_h is bounded above and away from zero and $B\asymp T^{2/5}$, integer rounding gives $K\asymp T^{3/5}$ and $KB\tau_h=\Theta(T)$. The initial critic error and σ_a^2 are uniformly bounded. Using (60),

$$\begin{aligned}\frac{1}{K}\sum_{k<K}\mathbb{E}\left[\|\mathcal{G}_\lambda(\phi_k)\|^2+\|z_k-z^\star_{\phi_k}\|^2\right]&\le C\left[\frac{B}{T\eta}+\frac{\eta}{B\sigma_e^2}+\frac{1}{B}+(e_V+e_Q)^2+\sigma_e^4+h^2\right]\\ &\le C\left[T^{-2/5}+(e_V+e_Q)^2\right],\end{aligned}$$

which proves (71). □

Remark B.8. The physical-time budget and the number of observations describe the same run at different resolutions. Each recorded transition lasts h, so its expected transition count is

$$N = KB\left(\frac{\gamma}{1-\gamma} + L\right) = \frac{KB\tau_h}{h} = \Theta(T/h).$$

For the mesh $h \asymp T^{-1/5}$ in Corollaries B.6–B.7,

$$N = \Theta(T^{6/5}), \qquad T^{-2/5} = \Theta(N^{-1/3}).$$

Thus the joint error is $O(N^{-1/3})$ with exact features and at most $C[N^{-1/3} + (e_V + e_Q)^2]$ with feature approximation. This is a conversion under the specified observation mesh, rather than an additional independent budget bound.

B.3 An LQG example with a moving actor

We give a linear–quadratic Gaussian family for which the hypotheses of Theorem B.4 hold uniformly as the actor changes. Restricting the feedback gain to a stable interval controls the state moments; quadratic value features and a linear slope feature then represent the quantities needed by the actor exactly.

Fix $\beta, a_0, \sigma, Q, R, s_0 > 0, b_0 \neq 0$, and a compact gain interval $\Phi = [\kappa_-, \kappa_+]$ with nonempty interior such that $\alpha_\kappa = a_0 - b_0\kappa \geq \rho > 0$. Set $\phi = \kappa$ and

$$\begin{aligned} dX_t &= (-a_0 X_t + b_0 A_t)\,dt + \sigma\,dW_t, \qquad X_0 \sim N(0, s_0), \\ r(x,a) &= -\tfrac{1}{2}(Qx^2 + Ra^2), \qquad \mu_\kappa(x) = \kappa x. \end{aligned} \tag{72}$$

Resets are independent; s_0 is their variance and α_κ is the continuous-feedback mean-reversion rate. Use $f(x) = (1, x^2)^\top$ and $C(x) = x$, so $V_\theta(x) = \theta_0 + \theta_2 x^2$ and $q_{\omega,\kappa}(x,a) = \omega x(a - \kappa x)$. The evaluation equation gives

$$p_\kappa = \frac{Q + R\kappa^2}{\beta + 2\alpha_\kappa}, \qquad \omega_\kappa = -(R\kappa + b_0 p_\kappa), \qquad \Lambda_\kappa = \frac{s_0 + \sigma^2/\beta}{\beta + 2\alpha_\kappa}, \tag{73}$$

where $\Lambda_\kappa = \int x^2 \rho_\kappa(dx)$, and

$$\begin{aligned} V^\kappa(x) &= -\tfrac{1}{2}p_\kappa x^2 - \frac{\sigma^2 p_\kappa}{2\beta}, & q^\kappa(x, \kappa x + u) &= \omega_\kappa x u - \tfrac{R}{2}u^2, \\ J(\kappa) &= -\tfrac{1}{2}p_\kappa(s_0 + \sigma^2/\beta), & J'(\kappa) &= \omega_\kappa \Lambda_\kappa. \end{aligned} \tag{74}$$

The features therefore represent V^κ and the derivative $\nabla_a q^\kappa(x, \kappa x) = \omega_\kappa x$ exactly. The quadratic action term does not affect this derivative at the policy action and need not be represented by the critic.

Proposition B.9 (Verification of the joint convergence hypotheses)**.** *For the LQG family* (72) *and the features above, there exist $h_0, \sigma_{e,0} > 0$ and a fixed compact convex critic set $\mathcal{Z}$ such that all hypotheses of Theorem B.4 hold uniformly for $\kappa \in \Phi$, $0 < h \leq h_0$, $0 < \sigma_e \leq \sigma_{e,0}$, and $\delta = Lh \in [\delta_{\min}, \delta_{\max}]$.*

Proof. We verify state moments and feature coverage, then the fitted critic and actor bounds. The exact transition over one observation interval is $X^+ = M_{\kappa,h}X + d_h U + \zeta$, where $U \sim N(0, \sigma_e^2)$ and $\zeta \sim N(0, v_h)$ are independent, with

$$d_h = \frac{b_0(1 - e^{-a_0 h})}{a_0}, \qquad v_h = \frac{\sigma^2(1 - e^{-2a_0 h})}{2a_0}, \qquad M_{\kappa,h} = e^{-a_0 h} + \kappa d_h. \tag{75}$$

For small h, stability gives $0 < M_{\kappa,h} \leq 1 - \rho h/2$. Conditional on a target-policy prefix of n transitions, its endpoint is centered Gaussian with variance

$$c_n = M_{\kappa,h}^{2n} s_0 + (1 - M_{\kappa,h}^{2n})\frac{v_h}{1 - M_{\kappa,h}^2}, \qquad 0 < c \leq c_n \leq C. \tag{76}$$

During exploration, the same variance recursion uses $v_h + \sigma_e^2 d_h^2$ instead of v_h; its variances remain uniformly bounded above and away from zero. Hence all required Gaussian moments are uniform,

including after mixing over the prefix length. The continuous-policy process has mean-reversion rate $\alpha_\kappa \geq \rho$, so its moments are uniformly bounded as well. For a centered Gaussian state with variance c,

$$\mathbb{E}(u_0 + u_2 X^2)^2 = (u_0 + u_2 c)^2 + 2u_2^2 c^2.$$

Together with $\mathbb{E}X^2 = c > 0$, this proves coverage of both feature families. The linear dynamics, policy, and quadratic reward satisfy Assumptions A.1–A.2. Corollaries A.6 and A.7 therefore give uniform critic contraction and the required sampling-noise bound.

The Gaussian transition also makes the fitted critic explicit:

$$\begin{aligned} \theta_2^\star &= -\frac{h(Q + R\kappa^2)}{2(1 - \gamma M_{\kappa,h}^2)}, \\ \theta_0^\star &= \frac{\gamma\theta_2^\star(v_h + \sigma_e^2 d_h^2) - hR\sigma_e^2/2}{1 - \gamma}, \\ \omega^\star &= -R\kappa + 2\gamma M_{\kappa,h}\frac{d_h}{h}\theta_2^\star. \end{aligned} \tag{77}$$

Substitution into the residual, using centered Gaussian moments, gives

$$\mathbb{E}[D_i(z^\star) \mid \mathcal{F}_i^-] = 0, \qquad \mathbb{E}[S_i D_i(z^\star) \mid \mathcal{F}_i^-] = 0.$$

Multiplying the first identity by $f(X_0)$ and summing both identities with weights γ^i/δ verifies the expected critic equation. Uniqueness follows from Corollary A.6.

The normalized denominators $(1 - \gamma M_{\kappa,h}^2)/h$ and $(1 - \gamma)/h$ converge uniformly to $\beta + 2\alpha_\kappa$ and β, respectively. Both stay bounded away from zero. Since $d_h/h \to b_0$ and $\partial_\kappa M_{\kappa,h} = d_h = O(h)$, the coefficients and their gain derivatives are uniformly bounded. Expanding the formulas at $h = 0$ gives

$$\|z^\star - \widetilde{z}_\kappa\| \leq C(h + \sigma_e^2), \qquad |\omega^\star - \omega_\kappa| \leq Ch, \qquad \widetilde{z}_\kappa = \left(-\frac{\sigma^2 p_\kappa}{2\beta}, -\frac{p_\kappa}{2}, \omega_\kappa\right).$$

Choose $\mathcal{Z}$ as the ball of radius $1 + \sup_{\kappa\in\Phi}\|\widetilde{z}_\kappa\|$. For sufficiently small $h_0, \sigma_{e,0}$, every fitted critic is at least $1/2$ from its boundary. This verifies membership and Lipschitz policy dependence of the fitted critic.

Finally, the discounted second moment of the prefix states is

$$\Lambda_{\kappa,h} = h\sum_{n\geq 0}\gamma^n c_n = \frac{h}{1 - \gamma M_{\kappa,h}^2}\left(s_0 + \frac{\gamma v_h}{1 - \gamma}\right) = \Lambda_\kappa + O(h).$$

Since $\bar{H}_{\kappa,h}(z) = \omega\Lambda_{\kappa,h}$ and $J'(\kappa) = \omega_\kappa\Lambda_\kappa$, we obtain

$$|\bar{H}_{\kappa,h}(z) - J'(\kappa)| \leq C\|z - z_\kappa^\star\| + Ch. \tag{78}$$

This implies the required actor-error bound. Also, J is smooth on the stable gain interval, and $H_{\kappa,h}(z; X_0) = h\omega X_0^2/(1 - \gamma)$ has bounded variance on $\mathcal{Z}$ by the fourth-moment bound. Every hypothesis of Theorem B.4 is now verified. □

The conclusion concerns constrained stationarity on the stable interval, which need not contain the globally optimal LQG gain. The fixed-policy variance experiment and its acquisition protocol are given in Section C.4.

B.4 Technical estimates for critic and policy-gradient accuracy

The proofs of Theorem B.1 and Corollary B.2 use two local estimates. The first controls the effect of a fixed action over one observation interval, including the error in its generator approximation. The second controls the error from replacing the true advantage by its linear response to a Gaussian action perturbation. Both estimates depend only on Assumptions A.1–A.3.

Lemma B.10 (Fixed-action sensitivity and observation error)**.** *Assume Assumptions A.1–A.3. For every allowed ϕ, $x \in \mathbb{R}^{d_s}$, $u \in \mathbb{R}^{d_a}$, $a = \mu_\phi(x) + u$, and sufficiently small $0 < h \le 1$, the following bounds hold uniformly. For every scalar $v \in C^2_w$,*

$$\|\nabla_a P^a_h v(x)\| \le Ch(1 + \|x\| + \|a\|)\|v\|_{C^2_w}. \tag{79}$$

For every scalar $v \in C^4_w$,

$$|v(x) - \gamma P^a_h v(x) - h(\beta - \mathcal{L}^a)v(x)| \le Ch^2 w(x)^2(1 + \|u\|^4)\|v\|_{C^4_w}. \tag{80}$$

For the true-value observation error Q^ϕ_h defined in (49),

$$\|\nabla_a Q^\phi_h(x, \mu_\phi(x) + u)\| \le Ch^2 w(x)^2(1 + \|u\|^4). \tag{81}$$

The constants are independent of ϕ, x, u, h and v.

The C^2_w bound applies to the value-approximation error in the critic. The C^4_w bounds apply to the true value and, coordinate by coordinate, to its policy derivative. Thus these estimates use exactly the different regularities imposed on the approximation error and the true value.

Proof. We first prove (79). Let X_t start at x and evolve with the action a held fixed. The stochastic-flow estimates in the proof of Lemma A.9 give

$$\|\mathbb{E}^a_x[\partial_a X_h]\| \le Ch, \qquad \|\partial_a X_h\|_{L^2} \le C\sqrt{h}, \qquad \|X_h - x\|_{L^2} \le C\sqrt{h}(1 + \|x\| + \|a\|).$$

The mean action sensitivity is of order h, although its L^2 norm is only of order $\sqrt{h}$. To retain this cancellation, we subtract the gradient at the initial state before estimating the fluctuating part:

$$\begin{aligned}
\nabla_a P^a_h v(x) &= (\mathbb{E}^a_x[\partial_a X_h])^\top \nabla v(x) + \mathbb{E}^a_x[(\partial_a X_h)^\top \{\nabla v(X_h) - \nabla v(x)\}], \\
\|\nabla_a P^a_h v(x)\| &\le \|v\|_{C^2_w}\{Ch(1 + \|x\|) + \|\partial_a X_h\|_{L^2}\|X_h - x\|_{L^2}\} \\
&\le Ch(1 + \|x\| + \|a\|)\|v\|_{C^2_w}.
\end{aligned}$$

Here the bounded Hessian of v controls the gradient difference. This proves (79) using only C^2_w regularity.

We next prove (80). With a fixed, two applications of discounted Dynkin's formula give

$$v(x) - \gamma P^a_h v(x) - h(\beta - \mathcal{L}^a)v(x) = -\int_0^h \int_0^s e^{-\beta t} P^a_t (\mathcal{L}^a - \beta)^2 v(x)\, dt\, ds.$$

The coefficient bounds and $v \in C^4_w$ imply

$$|(\mathcal{L}^a - \beta)^2 v(y)| \le C\|v\|_{C^4_w}(1 + \|y\| + \|a\|)^4.$$

The fixed-action moment estimate in Lemma A.8, together with the linear growth of μ_ϕ, yields

$$\sup_{0 \le t \le h} \mathbb{E}^a_x(1 + \|X_t\| + \|a\|)^4 \le C(1 + \|x\| + \|a\|)^4 \le Cw(x)^2(1 + \|u\|^4).$$

Integrating over $0 \le t \le s \le h$ now proves (80).

It remains to prove (81). We estimate the action derivative of the semigroup before taking the observation error. This avoids differentiating the double-Dynkin integrand and uses no fifth state derivative. All state derivatives below hold a fixed. For $k = 0, 1, 2$, the coefficient bounds and $V^\phi \in C^4_w$ give

$$\begin{aligned}
\|\partial^k_x(\mathcal{L}^a V^\phi)(x)\| &\le C(1 + \|x\| + \|a\|)^2, \\
\|\partial^k_x\{(\nabla_a \mathcal{L}^a)V^\phi\}(x)\| &\le C(1 + \|x\| + \|a\|).
\end{aligned}$$

The variational equations for the state flow, whose coefficient derivatives are bounded, also give

$$\sup_{0 \le t \le 1}\{\|\partial_x X_t\|_{L^p} + \|\partial^2_x X_t\|_{L^p}\} \le C_p, \qquad p \ge 2.$$

Differentiating Dynkin's formula through this flow therefore yields, for $k = 1, 2$ and $0 \le t \le 1$,

$$\begin{aligned}
\partial^k_x(P^a_t V^\phi - V^\phi)(x) &= \int_0^t \partial^k_x P^a_s(\mathcal{L}^a V^\phi)(x)\, ds, \\
\|\partial^k_x(P^a_t V^\phi - V^\phi)(x)\| &\le Ct(1 + \|x\| + \|a\|)^2.
\end{aligned}$$

These are the two state derivatives needed by the action derivative of the generator. In fact, for a smooth scalar function F,

$$(\partial_{a_j}\mathcal{L}^a)F = (\partial_{a_j}b)^\top \nabla F + \tfrac{1}{2}\operatorname{tr}[\partial_{a_j}(\sigma\sigma^\top)\nabla_x^2 F].$$

Since $\partial_a b$ is bounded and $\|\partial_a(\sigma\sigma^\top)(x,a)\| \le C(1+\|x\|+\|a\|)$, the preceding derivative bounds imply

$$\|(\nabla_a\mathcal{L}^a)(P_t^a V^\phi - V^\phi)(x)\| \le Ct(1+\|x\|+\|a\|)^3,$$
$$\|\mathcal{L}^a\{(\nabla_a\mathcal{L}^a)V^\phi\}(x)\| \le C(1+\|x\|+\|a\|)^3.$$

The first bound controls the change in the value inside the semigroup; the second controls the change in the action derivative of the generator along the state trajectory.

For two fixed actions a, a', Itô's formula gives

$$P_h^{a'}V^\phi - P_h^a V^\phi = \int_0^h P_s^{a'}(\mathcal{L}^{a'} - \mathcal{L}^a)P_{h-s}^a V^\phi \, ds.$$

Taking directional difference quotients as $a' \to a$ is justified by the polynomial bounds above and gives Duhamel's formula. Subtracting the leading term $h(\nabla_a\mathcal{L}^a)V^\phi$ separates precisely the two changes just estimated:

$$\begin{aligned}\nabla_a P_h^a V^\phi - h(\nabla_a\mathcal{L}^a)V^\phi &= \int_0^h P_s^a(\nabla_a\mathcal{L}^a)(P_{h-s}^a V^\phi - V^\phi)\, ds \\ &\quad + \int_0^h (P_s^a - I)(\nabla_a\mathcal{L}^a)V^\phi \, ds.\end{aligned}$$

Apply the first bound directly to the first integral. For the second, Dynkin's formula and the second bound give

$$\|(P_s^a - I)(\nabla_a\mathcal{L}^a)V^\phi(x)\| \le C\int_0^s \mathbb{E}_x^a(1+\|X_t\|+\|a\|)^3 dt \le Cs(1+\|x\|+\|a\|)^3.$$

The same fixed-action moment estimate in the first integral then shows that

$$\begin{aligned}\|\nabla_a P_h^a V^\phi(x) - h(\nabla_a\mathcal{L}^a)V^\phi(x)\| &\le C(1+\|x\|+\|a\|)^3 \int_0^h \{(h-s)+s\}\, ds \\ &= Ch^2(1+\|x\|+\|a\|)^3.\end{aligned}$$

Finally, differentiating the definition of Q_h^ϕ gives

$$\nabla_a Q_h^\phi(x,a) = h(1-\gamma)(\nabla_a\mathcal{L}^a)V^\phi(x) - \gamma\{\nabla_a P_h^a V^\phi(x) - h(\nabla_a\mathcal{L}^a)V^\phi(x)\}.$$

Using $1-\gamma \le \beta h$ and $a = \mu_\phi(x) + u$, we obtain

$$\|\nabla_a Q_h^\phi(x,a)\| \le Ch^2(1+\|x\|+\|a\|)^3 \le Ch^2 w(x)^2(1+\|u\|^4).$$

This proves (81). □

Lemma B.11 (Gaussian averages of the nonlinear advantage error)**.** *Assume Assumptions A.1–A.3. For the nonlinear remainder $\mathcal{R}^\phi$ defined in (50) and $U \sim N(0,\sigma_e^2 I)$ with $0 < \sigma_e \le 1$,*

$$\mathbb{E}_U|\mathcal{R}^\phi(x,U)| \le C\sigma_e^2 w(x), \qquad \|\mathbb{E}_U[\sigma_e^{-2}U\mathcal{R}^\phi(x,U)]\| \le C\sigma_e^2 w(x).$$

The constants are uniform over the allowed ϕ, x, σ_e.

Proof. We prove both estimates by separating the quadratic action term from the higher-order remainder. The identity $q^\phi = r + \mathcal{L}^a V^\phi - \beta V^\phi$ and the stated regularity give

$$\|\partial_a^k q^\phi(x,a)\| \le C(1+\|x\|+\|a\|), \qquad k = 2,3.$$

Since $q^\phi(x,\mu_\phi(x)) = 0$ and $g^\phi(x) = \nabla_a q^\phi(x,\mu_\phi(x))$, Taylor's formula gives

$$\begin{aligned}\mathcal{R}^\phi(x,u) &= \tfrac{1}{2}u^\top\partial_a^2 q^\phi(x,\mu_\phi(x))u + R_3^\phi(x,u), \\ R_3^\phi(x,u) &= \tfrac{1}{2}\int_0^1 (1-t)^2\partial_a^3 q^\phi(x,\mu_\phi(x)+tu)[u,u,u]dt, \\ |R_3^\phi(x,u)| &\le Cw(x)(1+\|u\|^2)\|u\|^3.\end{aligned} \tag{82}$$

Writing $U = \sigma_e Z$ with $Z \sim N(0, I)$ makes the needed Gaussian moment estimates explicit:

$$\mathbb{E}_U\|U\|^p = \sigma_e^p \mathbb{E}\|Z\|^p \le C_p \sigma_e^p, \qquad p = 2, 3, 4, 5, 6.$$

The quadratic term and the remainder in (82) therefore satisfy

$$\begin{aligned}\mathbb{E}_U|\mathcal{R}^\phi(x, U)| &\le Cw(x)\mathbb{E}_U[\|U\|^2 + \|U\|^3 + \|U\|^5] \\ &\le Cw(x)(\sigma_e^2 + \sigma_e^3 + \sigma_e^5) \le C\sigma_e^2 w(x).\end{aligned}$$

For the second estimate, the Gaussian score introduces a factor $\sigma_e^{-2}U$, but its product with the quadratic term is odd:

$$\mathbb{E}_U[U\{U^\top \partial_a^2 q^\phi(x, \mu_\phi(x))U\}] = 0.$$

Only the cubic remainder contributes to this mean. Consequently,

$$\begin{aligned}\|\mathbb{E}_U[\sigma_e^{-2}U\mathcal{R}^\phi(x, U)]\| &= \|\mathbb{E}_U[\sigma_e^{-2}UR_3^\phi(x, U)]\| \\ &\le Cw(x)\sigma_e^{-2}\mathbb{E}_U[\|U\|^4 + \|U\|^6] \\ &\le Cw(x)\sigma_e^{-2}(\sigma_e^4 + \sigma_e^6) \le C\sigma_e^2 w(x).\end{aligned}$$

This proves both estimates. The quadratic cancellation is what allows the score-weighted mean to have order σ_e^2. □

C How local pairing improves advantage-derivative estimation

In Appendix B, we have establishes convergence of the joint actor–critic updates in our AdvG algorithm. Now we examine a core mechanism behind AdvG, namely, the advantage-gradient estimate. As shown in equation (6), the actor needs $g^\phi(x) = \nabla_a q^\phi(x, \mu_\phi(x))$, which our linear critic represents by $C(x)^\top \omega$, ω being the learnable parameter. A natural question about cell-pairing design is: why should each action perturbation be paired with its own cell residual when learning ω?

In this section, we compare the advantage updates given by AdvG and CT-DDPG on the same trajectory trajectories, the same fixed policy and jointly learned value function. CT-DDPG uses a canonical SGD-type update, which multiplies each action feature by residuals from every cell, including noise observed before that action was chosen. In contrast, AdvG uses only the corresponding cell residual. Under the sufficient comparison conditions below, this difference produces an order-$L = \delta/h$ difference between the MSEs of the slope estimators produced by AdvG and CT-DDPG. We also use an exact Gaussian example to verify this difference via calculation, and use an LQG experiment to illustrate it with a quadratic value function.

C.1 The two slope equations

We fix the actor ϕ and use the sampling notation and window range of Appendix A, with $L \ge 2$. The n complete trajectories are independent. We retain the critic (7), the residuals (8), and the AdvG moment m^{p} in (9).

To express the full-window gradient easily, we introduce a notation for the accumulated action feature

$$P_i = h\sum_{j<i} \gamma^j C(X_j)U_j, \qquad P_L = \nabla_\omega D^V(z).$$

The common value equation and the two slope equations are

$$\begin{aligned}&m_V(z) = \delta^{-1} f(X_0)D^V(z), \\ &m_Q^{\mathrm{w}}(z) = \frac{P_L D^V(z)}{\delta h \sigma_e^2}, \qquad m_Q^{\mathrm{p}}(z) = \frac{1}{\delta}\sum_{i<L} \gamma^i S_i D_i(z) \qquad m^j(z) = \begin{bmatrix} m_V(z) \\ m_Q^j(z)\end{bmatrix}.\end{aligned} \tag{83}$$

Here $S_i = \sigma_e^{-2}C(X_i)U_i$ is the slope weight from Appendix A, and m^{w} denotes the moment induced by CT-DDPG's update rule, where w stands for "full window fitting".

Since U_i's are isotropic exploratory actions,

$$\mathbb{E}[f(X_0)(C(X_i)U_i)^\top] = 0, \qquad \mathbb{E}[(C(X_i)U_i)(C(X_k)U_k)^\top] = 0 \quad (i < k).$$

Consequently, the expected coefficient matrices have the form

$$A^j := \partial_z \mathbb{E} m^j(z) = \begin{bmatrix} A_V & 0 \\ B_j & Q_j \end{bmatrix}, \qquad j \in \{\mathrm{w}, \mathrm{p}\},$$
$$Q_{\mathrm{w}} = G_{2,h} := \frac{h}{\delta} \sum_{i<L} \gamma^{2i} \mathbb{E}[C(X_i)C(X_i)^\top], \qquad Q_{\mathrm{p}} = \frac{h}{\delta} \sum_{i<L} \gamma^{i} \mathbb{E}[C(X_i)C(X_i)^\top]. \tag{84}$$

The value block A_V is the one in (30). The common value equation determines the same θ^* for both methods; the conditions below also give them a common slope root.

C.2 A FINITE-SAMPLE COMPARISON WITH A LEARNED VALUE FUNCTION

We use an exact Gaussian regression for the cell residuals to isolate the variance caused by window aggregation. At the correct finite-mesh coefficients, each residual is centered Gaussian noise with variance proportional to the cell duration. This is a stronger, directly checkable comparison setting, not a consequence of the general diffusion assumptions.

For random vectors and matrices, write $\|Y\|_q = (\mathbb{E}\|Y\|^q)^{1/q}$, using the Frobenius norm where indicated.

Assumption C.1 (Conditions for the variance comparison)**.** *The feature dimensions are fixed, and all constants below are uniform in h, σ_e and the models under comparison.*

(i) There is a uniformly bounded coefficient vector $z^* = (\theta^*, \omega^*)$ satisfying condition (iii). A known $\lambda_0 > 0$ satisfies $s_{\min}(A^j) \geq \lambda_0$ for both methods, and $G_{2,h} \preceq g_+ I$.

(ii) The features have bounded fourth moments and cover every slope direction:

$$\mathbb{E}\|f(X_0)\|^4 + \sup_{i<L} \mathbb{E}\|C(X_i)\|_F^4 \leq K_4, \qquad \mathbb{E}[C(X_i)C(X_i)^\top] \succeq \kappa_Q I \quad (i < L).$$

For one trajectory Ξ, write $m^j(z;\Xi) = \widehat{A}_j(\Xi)z - \widehat{b}_j(\Xi)$. The coefficient and residual fluctuations satisfy

$$\mathbb{E}\|\widehat{A}_j - A^j\|_F^4 + \mathbb{E}\|m^j(z^*;\Xi)\|^4 \leq \begin{cases} K_4/(h^2\sigma_e^4), & j = \mathrm{w}, \\ K_4/(\delta^2\sigma_e^4), & j = \mathrm{p}. \end{cases} \tag{85}$$

(iii) For a nonrandom variance v_h and fixed $\nu_-, \nu_+ > 0$,

$$D_i(z^*) \mid \mathcal{F}_i^A \sim N(0, v_h), \qquad \nu_- h \leq v_h \leq \nu_+ h, \qquad i < L. \tag{86}$$

The three conditions provide identifiability, control of empirical fluctuations, and nonvanishing cell noise, respectively. For example, all three hold for the scalar diffusion

$$dX_t = (-\lambda X_t + \omega^\circ A_t)dt + \sigma dW_t, \qquad r(x,a) = (\beta + \lambda)x,$$

at $\mu = 0$, with $f(x) = x$, $C(x) = 1$, and reset law $N(0, \sigma^2/(2\lambda))$, for fixed $\lambda, \sigma, \beta > 0$, bounded $|\omega^\circ|$, and sufficiently small h.

Let N_i denote the residual accumulated before cell i:

$$N_i = \sum_{j<i} \gamma^j D_j(z^*), \qquad N_0 = 0, \qquad N_L = D^V(z^*).$$

The two stages of each cell, drawing U_i and then observing the transition, give

$$\mathbb{E}[C(X_i)U_i N_i \mid \mathcal{F}_i^-] = 0, \qquad \mathbb{E}[P_{i+1} D_i(z^*) \mid \mathcal{F}_i^A] = 0.$$

Also, $\mathbb{E}[D_i(z^*) \mid \mathcal{F}_i^A] = 0$ implies $\mathbb{E}[f(X_0)D_i(z^*)] = \mathbb{E}[S_i D_i(z^*)] = 0$. Hence

$$\mathbb{E}[P_L N_L] = \sum_{i<L} \mathbb{E}[h\gamma^i C(X_i)U_i N_i + \gamma^i P_{i+1} D_i(z^*)] = 0, \qquad \mathbb{E} m^{\mathrm{w}}(z^*) = \mathbb{E} m^{\mathrm{p}}(z^*) = 0.$$

Thus z^* is the common population root, and N_L is its window residual.

For each method, average the observed coefficients over the n trajectories to obtain $\bar{A}_j, \bar{b}_j$, and use the same inverse cutoff:

$$\widehat{z}_j = \begin{cases} \bar{A}_j^{-1}\bar{b}_j, & s_{\min}(\bar{A}_j) \geq \lambda_0/2, \\ 0, & \text{otherwise.} \end{cases} \tag{87}$$

The cutoff uses the known lower conditioning bound. The bounded population root controls the error when the empirical matrix is rejected.

Before bounding the slope error, we account for the jointly estimated value. Solving the two block equations for the linearized error from one trajectory gives

$$\begin{aligned} \zeta_{V,j} &= -A_V^{-1} m_V(z^*), \\ \zeta_{\omega,j} &= -Q_j^{-1}\{m_Q^j(z^*) - B_j A_V^{-1} m_V(z^*)\}. \end{aligned}$$

For the full-window fit, this becomes

$$\zeta_{\omega,\mathrm{w}} = -G_{2,h}^{-1} \frac{(P_L - Kf(X_0))N_L}{\delta h \sigma_e^2}, \qquad K = h\sigma_e^2 B_{\mathrm{w}} A_V^{-1}.$$

The subtraction by $Kf(X_0)$ is therefore the correction from learning the value. The next lemma shows which part of the noise remains after this correction.

Lemma C.2. *Under Assumption C.1,* $\mathbb{E}[(P_L - Kf(X_0))N_L] = 0$ *and*

$$\mathrm{Cov}((P_L - Kf(X_0))N_L) \succeq h^2\sigma_e^2 \sum_{i<L} \gamma^{2i}\mathbb{E}[N_i^2 C(X_i)C(X_i)^\top] \succeq c\delta^2 h\sigma_e^2 I. \tag{88}$$

Proof. We first separate the noise generated at the two stages of each cell:

$$\begin{aligned} &(P_{i+1} - Kf(X_0))N_{i+1} - (P_i - Kf(X_0))N_i \\ &\quad = h\gamma^i C(X_i)U_i N_i + \gamma^i (P_{i+1} - Kf(X_0))D_i(z^*). \end{aligned}$$

The first term is centered given $\mathcal{F}_i^-$ and is $\mathcal{F}_i^A$-measurable; the second is centered given $\mathcal{F}_i^A$. These are successive martingale increments, so

$$\begin{aligned} &\mathbb{E}[(P_L - Kf(X_0))N_L] = 0, \\ &\mathrm{Cov}((P_L - Kf(X_0))N_L) \\ &\quad = h^2\sigma_e^2 \sum_{i<L} \gamma^{2i}\mathbb{E}[N_i^2 C(X_i)C(X_i)^\top] \\ &\qquad + v_h \sum_{i<L} \gamma^{2i}\mathbb{E}[(P_{i+1} - Kf(X_0))(P_{i+1} - Kf(X_0))^\top]. \end{aligned}$$

Dropping the second positive semidefinite sum proves the first inequality.

It remains to bound the first sum from below. By (86), the successive residuals are independent $N(0, v_h)$ variables. Therefore

$$N_i \sim N(0, s_i^2), \qquad s_i^2 = v_h \sum_{j<i} \gamma^{2j} \geq \nu_- e^{-2\beta\delta_{\max}} ih.$$

For a unit vector u, put $Z_i = \|C(X_i)^\top u\|^2$. Condition (ii) gives $\mathbb{E}Z_i \geq \kappa_Q$ and $\mathbb{E}Z_i^2 \leq K_4$. Choose $\varepsilon > 0$ so that $\sqrt{K_4 \Pr(|G| < \varepsilon)} \leq \kappa_Q/2$ for $G \sim N(0,1)$. For $i \geq 1$, Cauchy–Schwarz then gives

$$\begin{aligned} \mathbb{E}[N_i^2 Z_i] &\geq \varepsilon^2 s_i^2 \{\mathbb{E}Z_i - \mathbb{E}[Z_i \mathbf{1}_{\{|N_i| < \varepsilon s_i\}}]\} \\ &\geq \varepsilon^2 s_i^2 \{\kappa_Q - \sqrt{K_4 \Pr(|G| < \varepsilon)}\} \geq \tfrac{1}{2}\kappa_Q \varepsilon^2 s_i^2 \geq cih. \end{aligned}$$

This bound allows dependence between N_i and $C(X_i)$. Since $L \geq 2$ and $\delta = Lh$,

$$h^2\sigma_e^2 \sum_{i<L} \gamma^{2i}\mathbb{E}[N_i^2 Z_i] \geq ch^3\sigma_e^2 \sum_{i<L} i = \frac{c}{2} h^3 \sigma_e^2 L(L-1) \geq c'\delta^2 h\sigma_e^2.$$

The bound in every unit direction proves (88). □

Theorem C.3 (Finite-sample slope comparison)**.** *Under Assumption C.1, there are constants $c, C, C_0 > 0$ such that, for $n \geq C_0/(h\sigma_e^2)$, the joint value–slope fits (87) satisfy*

$$\mathrm{Cov}(\widehat{\omega}_{\mathrm{w}}) \succeq \frac{c}{nh\sigma_e^2} I, \qquad \mathbb{E}[(\widehat{\omega}_{\mathrm{p}} - \omega^*)(\widehat{\omega}_{\mathrm{p}} - \omega^*)^\top] \preceq \frac{C}{n\delta\sigma_e^2} I. \tag{89}$$

Consequently,

$$\mathbb{E}\|\widehat{\omega}_{\mathrm{w}} - \omega^*\|^2 \geq c' \frac{\delta}{h} \mathbb{E}\|\widehat{\omega}_{\mathrm{p}} - \omega^*\|^2.$$

With the same trajectories and a jointly learned value function, AdvG reduces the slope mean-square error about the common population root by at least an order-$L = \delta/h$ factor relative to its CT-DDPG counterpart.

Proof. We bound the linearized error first, and then the error of the empirical inverse. For $\zeta_j = -(A^j)^{-1} m^j(z^*)$, Lemma C.2 gives

$$\mathrm{Cov}(\zeta_{\omega,\mathrm{w}}) \succeq \frac{c}{h\sigma_e^2} G_{2,h}^{-2} \succeq \frac{c}{g_+^2 h\sigma_e^2} I.$$

For AdvG, (85) gives $\mathbb{E}\|\zeta_{\omega,\mathrm{p}}\|^2 \leq \lambda_0^{-2} (\mathbb{E}\|m^{\mathrm{p}}(z^*)\|^4)^{1/2}$. Thus

$$\mathrm{Cov}(\zeta_{\omega,\mathrm{w}}) \succeq \frac{c}{h\sigma_e^2} I, \qquad \mathbb{E}\|\zeta_{\omega,\mathrm{p}}\|^2 \leq \frac{C}{\delta\sigma_e^2}. \tag{90}$$

For the empirical inverse of either method j, suppress its index and write

$$u_n = \frac{1}{n} \sum_{k=1}^{n} m(z^*; \Xi_k), \qquad \Delta A = \bar{A} - A, \qquad H = \{s_{\min}(\bar{A}) \geq \lambda_0/2\}.$$

Since $\Pr(H^c) \leq \Pr(\|\Delta A\| > \lambda_0/2) \leq C\mathbb{E}\|\Delta A\|_F^4$, the trajectories are independent, and $\widehat{A}(\Xi_k) - A$ and $m(z^*; \Xi_k)$ have mean zero, together with (85) and the preceding probability bound, we have

$$\mathbb{E}\|u_n\|^4 + \mathbb{E}\|\Delta A\|_F^4 + \Pr(H^c) \leq \begin{cases} C/(n^2 h^2 \sigma_e^4), & j = \mathrm{w}, \\ C/(n^2 \delta^2 \sigma_e^4), & j = \mathrm{p}. \end{cases}$$

The inverse identity gives the exact error decomposition

$$\widehat{z} - z^* = -A^{-1} u_n + E_n, \qquad E_n = \begin{cases} \bar{A}^{-1} \Delta A A^{-1} u_n, & H, \\ -z^* + A^{-1} u_n, & H^c. \end{cases}$$

Using $\|\bar{A}^{-1}\| \leq 2/\lambda_0$ on H and boundedness of z^*,

$$\begin{aligned} \mathbb{E}\|E_n\|^2 &\leq C\big\{(\mathbb{E}\|\Delta A\|_F^4\, \mathbb{E}\|u_n\|^4)^{1/2} + \Pr(H^c) + (\mathbb{E}\|u_n\|^4\, \Pr(H^c))^{1/2}\big\} \\ &\leq \begin{cases} C/(n^2 h^2 \sigma_e^4), & j = \mathrm{w}, \\ C/(n^2 \delta^2 \sigma_e^4), & j = \mathrm{p}. \end{cases} \end{aligned} \tag{91}$$

No independence between ΔA and u_n is used.

For every unit slope direction v, the two bounds above yield

$$\sqrt{\mathrm{Var}(v^\top \widehat{\omega}_{\mathrm{w}})} \geq \sqrt{\frac{c}{nh\sigma_e^2}} - \frac{C}{nh\sigma_e^2},$$

$$\|v^\top(\widehat{\omega}_{\mathrm{p}} - \omega^*)\|_2 \leq \sqrt{\frac{C}{n\delta\sigma_e^2}} + \frac{C}{n\delta\sigma_e^2}.$$

Since $h \leq \delta$, the condition $n \geq C_0/(h\sigma_e^2)$ absorbs both remainders. Squaring proves (89), and taking traces with $p_Q = \dim \omega$ gives

$$\mathbb{E}\|\widehat{\omega}_{\mathrm{w}} - \omega^*\|^2 \geq \frac{c p_Q}{nh\sigma_e^2} = \frac{c}{C} \frac{\delta}{h} \frac{C p_Q}{n\delta\sigma_e^2} \geq \frac{c}{C} \frac{\delta}{h} \mathbb{E}\|\widehat{\omega}_{\mathrm{p}} - \omega^*\|^2.$$

□

The same bounds apply to the represented action derivative: multiply the covariance matrices in (89) by $C(x)^\top$ and $C(x)$. For estimation of $g^\phi(x)$ itself, the paired error also includes the population approximation error:

$$\mathbb{E}\|C(x)^\top \widehat{\omega}_{\mathrm{p}} - g^\phi(x)\|^2 \le \frac{C\|C(x)\|_F^2}{n\delta\sigma_e^2} + 2\|C(x)^\top \omega^* - g^\phi(x)\|^2.$$

Section B.1 bounds the corresponding approximation errors averaged over the sampled states and the discounted occupation measure. The comparison here isolates the statistical cost of the choice of slope equation.

C.3 An exact Gaussian illustration

The aggregation penalty is explicit even when no value function needs to be learned. Observe $n > 2$ windows of $L \ge 2$ cells with

$$Y_{ki} = \vartheta h U_{ki} + \sigma\sqrt{h}\, Z_{ki}, \qquad U_{ki} \sim N(0, \sigma_e^2), \quad Z_{ki} \sim N(0,1), \qquad 1 \le k \le n, \quad 1 \le i \le L,$$

where $h, \sigma_e, \sigma > 0$, $\vartheta \in \mathbb{R}$ is unknown, and all probes and noises are independent. Each Y_{ki} is the state increment produced by action U_{ki}. On these same observations, the two fits are

$$\widehat{\vartheta}_{\mathrm{p}} = \frac{\sum_{ki} U_{ki} Y_{ki}}{h \sum_{ki} U_{ki}^2}, \qquad \widehat{\vartheta}_{\mathrm{w}} = \frac{\sum_k (\sum_i U_{ki})(\sum_i Y_{ki})}{h \sum_k (\sum_i U_{ki})^2}.$$

Proposition C.4. *The two estimates are unbiased and satisfy*

$$\mathbb{E}(\widehat{\vartheta}_{\mathrm{p}} - \vartheta)^2 = \frac{\sigma^2}{h\sigma_e^2(nL-2)}, \qquad \mathbb{E}(\widehat{\vartheta}_{\mathrm{w}} - \vartheta)^2 = \frac{\sigma^2}{h\sigma_e^2(n-2)}. \tag{92}$$

Their mean-square error ratio is $(nL-2)/(n-2)$.

Proof. Write $U = (U_{ki})$ for all observed probes. The centered errors are

$$\widehat{\vartheta}_{\mathrm{p}} - \vartheta = \frac{\sigma}{\sqrt{h}} \frac{\sum_{ki} U_{ki} Z_{ki}}{\sum_{ki} U_{ki}^2}, \qquad \widehat{\vartheta}_{\mathrm{w}} - \vartheta = \frac{\sigma}{\sqrt{h}} \frac{\sum_k (\sum_i U_{ki})(\sum_i Z_{ki})}{\sum_k (\sum_i U_{ki})^2}.$$

Conditional independence of the Z_{ki} gives

$$\mathbb{E}[\widehat{\vartheta}_j - \vartheta \mid U] = 0 \quad (j \in \{\mathrm{p}, \mathrm{w}\}),$$
$$\mathrm{Var}(\widehat{\vartheta}_{\mathrm{p}} \mid U) = \frac{\sigma^2}{h \sum_{ki} U_{ki}^2}, \qquad \mathrm{Var}(\widehat{\vartheta}_{\mathrm{w}} \mid U) = \frac{L\sigma^2}{h \sum_k (\sum_i U_{ki})^2}.$$

Finally,

$$\frac{\sum_{ki} U_{ki}^2}{\sigma_e^2} \sim \chi^2_{nL}, \qquad \frac{\sum_k (\sum_i U_{ki})^2}{L\sigma_e^2} \sim \chi^2_n, \qquad \mathbb{E}[(\chi^2_d)^{-1}] = \frac{1}{d-2}.$$

Taking expectations of the conditional means and variances proves unbiasedness and (92). □

The same calculation quantifies the information lost if intermediate observations are discarded. Conditional on U, retaining all increments or only their sum within each window gives Fisher information

$$I_{\mathrm{trajectory}}(U) = \frac{h}{\sigma^2} \sum_{ki} U_{ki}^2, \qquad I_{\mathrm{endpoint}}(U) = \frac{h}{L\sigma^2} \sum_k \Big(\sum_i U_{ki} \Big)^2.$$

With $\bar{U}_k = L^{-1} \sum_i U_{ki}$, their difference is

$$I_{\mathrm{trajectory}}(U) - I_{\mathrm{endpoint}}(U) = \frac{h}{\sigma^2} \sum_{ki} (U_{ki} - \bar{U}_k)^2. \tag{93}$$

Intermediate observations reveal the response to changes in the action. The gap vanishes for actions held constant throughout a window; with fresh independent probes, the risk ratio grows as L. When all cell observations remain available, the window fit loses precision through its estimating equation. When only window sums are retained, (93) describes an actual loss of information.

C.4 Fixed-policy LQG variance experiment

We next compare the two fits in the scalar LQG model of Section B.3, where the value is learned using quadratic features. Figure 4 shows the empirical slope variances on the same trajectories for each method.

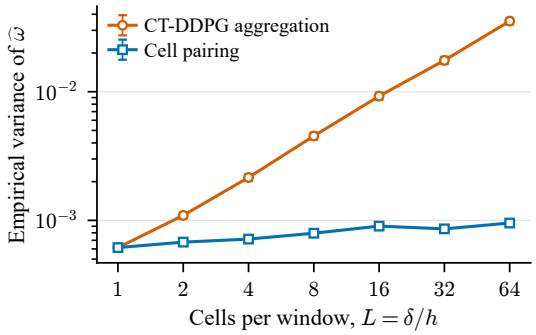


Figure 4: Scalar LQG slope variance at fixed $\mu(x) = -x/2$, $\delta = 1$, and $\sigma_e = 0.5$: 2,000 repetitions, each fitting both critics to the same $n = 2{,}048$ complete trajectories. Bars show ± 1.96 Monte Carlo standard errors; lines join empirical variances. The window budget $n\delta$ is fixed across L; total cost includes geometric prefixes (Section C.4).

Figure 4 uses the model of Section B.3 at gain $\kappa = -1/2$, with $a_0 = b_0 = \sigma = \beta = Q = R = s_0 = \delta = 1$ and $\sigma_e = 1/2$. For each $L \in \{1, 2, 4, 8, 16, 32, 64\}$, we generate 2,000 independent data sets of $n = 2{,}048$ complete trajectories using the exact transition (75). Each geometric prefix endpoint is sampled from its exact conditional Gaussian law (76), while its physical duration is charged. The expected total cost is $n\{1 + h/(e^h - 1)\}$, ranging from 3,239.9 at $L = 1$ to 4,080.0 at $L = 64$, and is identical for the two methods at each L.

Both methods jointly fit $(\theta_0, \theta_2, \omega)$ using the normalized equations (83), with current value at both endpoints. In this known simulation model, each cutoff is half the smallest singular value of that method's exact population matrix; rejected fits would return zero. No fit was rejected. At $L = 1$ the normalized systems and fitted coefficients coincide, providing an implementation check. The master random seed is 2026091701; all repetitions enter the results.

The conditional identities in Proposition B.9 give both fits the population root (77). In this quadratic model, however, the cell residual need not be conditionally centered:

$$\mathbb{E}[D_i(z^*) \mid \mathcal{F}_i^A] = \left(\frac{hR}{2} - \gamma\theta_2^* d_h^2\right)(U_i^2 - \sigma_e^2).$$

Thus the experiment lies outside the exact Gaussian regression in Assumption C.1(iii). We report empirical slope variance about each method's sample mean, separately from its bias relative to this finite-mesh solution. At $L = 64$, the full-window and AdvG variances are 0.03539 and 0.0009535, a ratio of 37.1 with Monte Carlo standard error 1.7. The experiment illustrates the benefit of local pairing with a learned quadratic value function, beyond the sufficient setting of Theorem C.3.

D Minimax rates for gradient estimation and policy optimization

In Appendix C we demonstrated how local pairing improves advantage-gradient estimation. Through the deterministic policy-gradient formula (6), this local estimate guides policy improvement. In this section we therefore turn to policy-gradient accuracy and the return of the resulting policy, two central quantities in reinforcement learning. Our aim is to determine the best accuracy achievable from a fixed amount of physical interaction and whether AdvG's paired construction attains it; such minimax characterizations remain scarce for reinforcement learning in continuous time and space. Section F.3 discusses the related statistical literature.

We will prove that, with exact value and action-derivative features, a batch estimator built from AdvG's paired critic achieves policy-gradient mean-squared error of order $T^{-2/3}$, where T counts all expected trajectory time used for critic fitting and state sampling. Moreover, under an additional projected gradient-dominance condition, repeated gradient estimates yield expected return loss of the same order $T^{-2/3}$, relative to the best actor in the prescribed family. We then establish matching

lower bounds on selected fixed diffusion classes, showing that the $O(T^{-2/3})$ rates are statistically sharp: even when learners may choose actions adaptively and observe continuous state paths, neither rate can be uniformly improved. The results provably connect the estimation benefit of local pairing to optimal statistical accuracy for policy learning under a physical interaction budget.

D.1 Model class and observation budget

A minimax statement compares the worst-case error of all estimators over a fixed class of unknown models. The class and the information given to the learner must stay fixed as T grows. In particular, a feature dictionary chosen separately for each unknown model could itself reveal the gradient. We therefore fix the dimensions, discount $\beta > 0$, and the known data

$$\mathcal{S} = (\nu_0, \mu, \phi, f, C).$$

These are the reset distribution, actor family and evaluation parameter, value features $f : \mathbb{R}^{d_s} \to \mathbb{R}^{p_V}$, and slope features $C : \mathbb{R}^{d_s} \to \mathbb{R}^{p_Q \times d_a}$. The drift, diffusion coefficient, and reward are unknown; rewards are observed along with transitions. The target is $\nabla J(\phi)$ for the continuous-feedback policy.

Use the sampling notation of Appendix A: $\gamma = e^{-\beta h}$, $\delta = Lh \in [\delta_{\min}, \delta_{\max}]$, and the starting-state law $d^0_{\phi,h}$ in (23).

The class conditions have three purposes: control the state and local transition errors, ensure that the value and its policy derivative are regular, and make the critic coefficients identifiable from sampled states. We use the fixed quadratic growth convention of Appendix A.

Definition D.1 (A diffusion class with fixed critic features)**.** Let $\mathcal{K}$ be known common bounds, independent of T, including positive coverage constants and fixed ranges of h, σ_e, δ. The class $\mathfrak{M}(\mathcal{S}, \mathcal{K}; e_V, e_Q)$ consists of controlled diffusions with the known data $\mathcal{S}$ satisfying:

(i) *Smooth dynamics, rewards, and features.* Assumptions A.1 and A.2 hold with bounds in $\mathcal{K}$. They impose smooth coefficients and rewards, controlled feature growth, and uniform state moments under the target and exploratory policies.

(ii) *Regular value and policy derivative.* On a fixed neighborhood of the evaluation parameter, the true value satisfies $(\beta - \mathcal{L}^{\mu_\phi})V^\phi = r(\cdot, \mu_\phi)$, is C^1 in ϕ, and obeys the uniform bound

$$\|V^\phi\|_{C^4_w} + \|\nabla_\phi V^\phi\|_{C^4_w} \le K.$$

These are the value-regularity conditions of Assumption A.3; no condition on how a critic root moves during policy optimization is needed here.

(iii) *Feature coverage and approximation.* The coverage bounds (27) hold uniformly. There are reference coefficients $\widetilde{z} = (\widetilde{\theta}, \widetilde{\omega})$, independent of h, σ_e, with $\|\widetilde{z}\| \le K$, satisfying the value and local-slope approximation bounds (28) with errors e_V, e_Q. No projection or interiority condition is imposed.

Here and below, K denotes a constant determined only by the fixed class data. We say the features are exact if $e_V = e_Q = 0$.

Exact features represent $V^\phi(x)$ and $g^\phi(x) = \nabla_a q^\phi(x, \mu_\phi(x))$; they do not require $q^\phi(x, a)$ to be affine in a. The unknown action response may therefore remain nonlinear even when both quantities used by the actor are represented exactly. The family used for the lower bound will illustrate this distinction with only two value features and one slope feature. Condition (ii) also implies the policy-gradient identity and controls the sampled state average through (56).

Data and time budget. Our estimators use the independent trajectories of Appendix A. The lower bounds allow more general data collection: the learner may observe states and rewards continuously and use the history to choose actions, resets, and when to stop. These choices may be randomized. Given the same history, the learner uses the same decision and sampling rules under every model, including for its random draws and the initial and reset states. If episode e lasts τ_e, the expected total trajectory time must satisfy

$$\mathcal{C} = \sum_e \tau_e, \qquad \sup_M \mathbb{E}_M \mathcal{C} \le T,$$

Algorithm 2 Fixed-policy gradient from the AdvG critic

Require: Fixed actor ϕ, known $\mathcal{S}$, expected budget T; h, σ_e, L, weight χ and $\lambda_0 = \mu_\chi$ from Corollary A.6, and solve residual tolerance $\eta_{\text{solve}} \geq 0$.

Compute $\gamma = e^{-\beta h}$, $\delta = Lh$, and n from (94); require $n \geq 1$.
Collect n independent critic trajectories as in Appendix A, starting from (23).
Form the affine coefficients A_{Ξ_b}, b_{Ξ_b} of (25); set $\overline{A} = n^{-1}\sum_b W_\chi A_{\Xi_b}$ and $\overline{b} = n^{-1}\sum_b W_\chi b_{\Xi_b}$.
if $s_{\min}(\overline{A}) \geq \lambda_0/2$ **then**
    Find $\widehat{z} = (\widehat{\theta}, \widehat{\omega})$ with $\|\overline{A}\widehat{z} - \overline{b}\| \leq \eta_{\text{solve}}$.
else
    $\widehat{z} \leftarrow 0$.
end if
Independently collect n target-policy prefixes with endpoints $Z_1, \ldots, Z_n$.
Set $\widehat{L} = \dfrac{h}{n(1-\gamma)} \sum_{b=1}^{n} \mathcal{B}_\phi(Z_b)$.
return $\widehat{g} = \widehat{L}\widehat{\omega}$.

where the supremum is over the model class in the stated bound. All acquired trajectory time, including prefixes, is counted.

D.2 Gradient-estimation upper bound

To estimate the state average in (6), fit the slope coefficients from paired trajectories and use an independent batch of states. Write

$$\mathcal{B}_\phi(x) = \partial_\phi \mu_\phi(x)^\top C(x)^\top.$$

For the trajectory equation $m(z;\Xi)$ in (25), suppress ϕ and denote its affine coefficients by

$$A_\Xi = \partial_z m(z;\Xi), \qquad b_\Xi = -m(0;\Xi), \qquad m(z;\Xi) = A_\Xi z - b_\Xi.$$

Thus $A = \mathbb{E}A_\Xi$, $b = \mathbb{E}b_\Xi$, and $z^* = (\theta^*, \omega^*) = A^{-1}b$ are the mean coefficients and fitted root already studied in Appendix A. Take the fixed weight χ and set $\lambda_0 = \mu_\chi$ from Corollary A.6; write R_* for its uniform bound on $\|z^*\|$. Definition D.1 supplies the regularity and coverage used in that corollary.

In Algorithm 2, we collect n critic trajectories and n independent target-policy prefixes, whose endpoints are $Z_1, \ldots, Z_n$. A prefix has expected duration $h\gamma/(1-\gamma) \leq 1/\beta$; hence the cost per critic trajectory together with one additional prefix is

$$\tau_h^{\text{grad}} = \delta + \frac{2h\gamma}{1-\gamma}, \qquad n = \lfloor T/\tau_h^{\text{grad}} \rfloor. \tag{94}$$

Thus $\delta \leq \tau_h^{\text{grad}} \leq \delta + 2/\beta$ and $n = \Theta(T)$ for sufficiently large T, uniformly in the allowed mesh and window ranges. The expected total cost is $n\tau_h^{\text{grad}} \leq T$, including all prefixes. It corresponds to $2n$ resets and $n\tau_h^{\text{grad}}/h$ expected transitions.

We form the estimators $\overline{A} = n^{-1}\sum_b W_\chi A_{\Xi_b}$ and $\overline{b} = n^{-1}\sum_b W_\chi b_{\Xi_b}$. We accept a solve only if the least singular value of $\overline{A}$ is at least $\lambda_0/2$, and return zero otherwise. This prevents rare, nearly singular sampled matrices from producing arbitrarily large coefficients. On acceptance, the numerical residual is at most η_{solve}.

Lemma D.2. *Under Definition D.1, for sufficiently small $h, \sigma_e > 0$, Algorithm 2 satisfies*

$$\mathbb{E}\|\widehat{z} - z^*\|^2 \leq K\{(n\sigma_e^2)^{-1} + \eta_{\text{solve}}^2\}. \tag{95}$$

This bound includes the error when the sampled matrix is rejected.

Proof. At the mean root, trajectory independence and (36), with $\delta \geq \delta_{\min}$, give

$$\overline{m}_* := \overline{A}z^* - \overline{b} = \frac{1}{n}\sum_{b=1}^{n} W_\chi m(z^*;\Xi_b), \qquad \mathbb{E}\overline{m}_* = 0, \qquad \mathbb{E}\|\overline{m}_*\|^2 \leq \frac{K}{n\sigma_e^2}.$$

For the matrix variance, let $p_V = \dim\theta$ and $p_Q = \dim\omega$ be the numbers of value and slope coefficients. Write $p = p_V + p_Q$, and let e_j be the jth coordinate vector in $\mathbb{R}^p$. Applying (35) at 0 and e_j gives

$$(A_\Xi - A)e_j = \{m(e_j;\Xi) - \bar{m}(e_j)\} - \{m(0;\Xi) - \bar{m}(0)\},$$
$$\mathbb{E}\|\overline{A} - W_\chi A\|_F^2 = \frac{1}{n}\sum_{j=1}^{p}\mathbb{E}\|W_\chi(A_\Xi - A)e_j\|^2 \le \frac{K}{n\sigma_e^2}.$$

Here (35) applies because z^* is uniformly bounded. Let $E = \{s_{\min}(\overline{A}) \ge \lambda_0/2\}$ be the acceptance event. The solve tolerance and the inverse bound in Corollary A.6 yield

$$\|\widehat{z} - z^*\|\mathbf{1}_E \le \frac{2}{\lambda_0}(\|\overline{m}_*\| + \eta_{\text{solve}})\mathbf{1}_E,$$
$$\mathbb{P}(E^c) \le \mathbb{P}\big(\|\overline{A} - W_\chi A\| \ge \lambda_0/2\big) \le \frac{4}{\lambda_0^2}\mathbb{E}\|\overline{A} - W_\chi A\|_F^2.$$

On rejection $\widehat{z} = 0$, so

$$\mathbb{E}\|\widehat{z} - z^*\|^2 \le \frac{8}{\lambda_0^2}(\mathbb{E}\|\overline{m}_*\|^2 + \eta_{\text{solve}}^2) + R_*^2\mathbb{P}(E^c)$$
$$\le K\{(n\sigma_e^2)^{-1} + \eta_{\text{solve}}^2\}.$$

□

The independent state batch separates the fitted-slope error from the error in the policy-gradient average.

Theorem D.3 (Fixed-policy gradient risk)**.** *Uniformly over the class in Definition D.1, for sufficiently small $h, \sigma_e > 0$ and sufficiently large expected physical budget T, Algorithm 2 satisfies*

$$\mathbb{E}\|\widehat{g} - \nabla J(\phi)\|^2 \le K\left\{\frac{1}{T\sigma_e^2} + \big(e_V + e_Q + \sigma_e^2 + h\big)^2 + \eta_{\text{solve}}^2\right\}. \tag{96}$$

The same bound holds conditionally on a previously selected actor when the class bounds are uniform over its actor region and all data used in this call are fresh.

Proof. Set $L_h = \mathbb{E}\widehat{L} = h(1-\gamma)^{-1}\mathbb{E}_{d^0_{\phi,h}}\mathcal{B}_\phi$, so $L_h\omega = \bar{H}_{\phi,h}(\theta,\omega)$. Use (53), (56), and the inverse bound in Corollary A.6. For the reference slope, apply Cauchy–Schwarz and (28) at $X_0 \sim d^0_{\phi,h}$. These give

$$\|z^* - \widetilde{z}\| \le \lambda_0^{-1}\|W_\chi\bar{m}(\widetilde{z})\| \le K(e_V + e_Q + \sigma_e^2 + h),$$
$$\|L_h\omega^* - \nabla J(\phi)\| \le \|L_h\|\,\|\omega^* - \widetilde{\omega}\| + \|L_h\widetilde{\omega} - \nabla J(\phi)\|$$
$$\le K\|z^* - \widetilde{z}\| + K(e_Q + h) \le K(e_V + e_Q + \sigma_e^2 + h).$$

These estimates use precisely the value regularity and approximation in Definition D.1; the projection and moving-actor conditions of Section B.1 are not used.

The independent prefixes and the feature moments give

$$\mathbb{E}\|\widehat{L} - L_h\|_F^2 \le K/n, \qquad \mathbb{E}\|\widehat{L}\|_F^2 \le K.$$

Separate the critic error, state-average error, and bias:

$$\widehat{g} - \nabla J(\phi) = \widehat{L}(\widehat{\omega} - \omega^*) + (\widehat{L} - L_h)\omega^* + \{L_h\omega^* - \nabla J(\phi)\}.$$

Independence of the two batches and Lemma D.2 therefore yield

$$\mathbb{E}\|\widehat{g} - \nabla J(\phi)\|^2 \le 3\mathbb{E}\|\widehat{L}\|_F^2\,\mathbb{E}\|\widehat{\omega} - \omega^*\|^2 + 3\|\omega^*\|^2\mathbb{E}\|\widehat{L} - L_h\|_F^2 + 3\|L_h\omega^* - \nabla J(\phi)\|^2$$
$$\le K\{(n\sigma_e^2)^{-1} + \eta_{\text{solve}}^2 + n^{-1} + (e_V + e_Q + \sigma_e^2 + h)^2\}.$$

Use $n = \Theta(T)$ and $\sigma_e \le 1$ to obtain (96). The conditional claim follows by fixing the previously selected actor and using fresh trajectories in both batches. □

With exact features, the essential terms in this bound are

$$\underbrace{\frac{1}{T\sigma_e^2}}_{\text{sampling variance}} + \underbrace{\sigma_e^4}_{\text{squared action bias}} + \underbrace{h^2}_{\text{squared observation error}}.$$

The first two terms balance when $\sigma_e^6 = T^{-1}$. Observation frequency can make the third term smaller, but does not remove this variance–bias tradeoff in physical time.

Corollary D.4 (The $T^{-2/3}$ gradient rate)**.** *For the exact-feature class, take $\sigma_e = T^{-1/6}$, $h \lesssim T^{-1/3}$, a window duration in $[\delta_{\min}, \delta_{\max}]$, and $\eta_{\text{solve}} \lesssim T^{-1/3}$. Then*

$$\sup_{M \in \mathfrak{M}(\mathcal{S},\mathcal{K};0,0)} \mathbb{E}_M \|\widehat{g} - \nabla J_M(\phi)\|^2 \le K T^{-2/3}. \tag{97}$$

Here J_M is the return in model M; the class bounds and constants in the parameter choices are fixed independently of T.

Proof. Theorem D.3 with $e_V = e_Q = 0$ gives

$$\mathbb{E}\|\widehat{g} - \nabla J(\phi)\|^2 \le K \left\{ \frac{1}{T\sigma_e^2} + \sigma_e^4 + h^2 + \eta_{\text{solve}}^2 \right\}.$$

Each term is $O(T^{-2/3})$ under the stated choices. □

Remark D.5. Choosing $h \asymp T^{-1/3}$ in Corollary D.4 uses $\Theta(T^{4/3})$ expected transitions, including prefixes. This is a sufficient observation schedule for this estimator. For $p = p_V + p_Q$, forming the empirical equation takes $O(nLp^2)$ arithmetic and a dense factorization takes $O(p^3)$; computation cost is separate from physical acquisition time.

D.3 Return optimization under gradient dominance

The gradient estimator supplies an improvement direction at any fixed actor. To turn this accuracy into a return guarantee, we need a condition connecting a feasible gradient step to the remaining return gap.

Use Definition D.1 with exact features, uniformly over a known compact convex actor set Φ with computable projection. Let $\mathfrak{C}_{\text{gd}}$ denote the models that also satisfy the return conditions below, with all known data and bounds fixed. Assume J has uniformly bounded range on Φ and an L_J-Lipschitz gradient on a neighborhood of Φ. The benchmark is $J^* = \max_{\phi \in \Phi} J(\phi)$.

To connect a feasible gradient step with this benchmark, assume that there are known common $\lambda, \mu > 0$ such that

$$\begin{aligned} \mathcal{G}_\lambda(\phi) &= \frac{\Pi_\Phi(\phi + \lambda \nabla J(\phi)) - \phi}{\lambda}, \\ \|\mathcal{G}_\lambda(\phi)\|^2 &\ge 2\mu\{J^* - J(\phi)\}, \qquad \phi \in \Phi. \end{aligned} \tag{98}$$

Here Π_Φ is Euclidean projection. The projected gradient $\mathcal{G}_\lambda$ measures the feasible movement from an exact policy-gradient step. Condition (98), called gradient dominance, says that this movement cannot be small while the actor's return is far below J^*. It rules out suboptimal stationary actors and lets an ascent estimate control return loss. This is an additional condition on the optimization problem, not a consequence of smooth diffusion dynamics. The lower-bound family below satisfies it with constants independent of the unknown model.

For a constrained actor, the projection in this condition matters. Strong concavity alone need not suffice: on $[0, 1]$, the return $J(x) = -x - x^2$ has squared projected gradient x^2/λ^2 near its boundary maximizer, whereas the return gap is $x + x^2$. We therefore verify (98) explicitly for the example used in the lower bound.

Increasing the batch size. At each epoch, estimate the current actor's gradient with fresh data using Algorithm 2, then take one projected ascent step. Later steps reduce the effect of earlier estimation errors. We therefore increase the batch size geometrically rather than requiring the final accuracy at every epoch. Choose a fixed allowed window duration δ and

$$0 < \alpha \le \min\{\lambda, (2L_J)^{-1}, \mu^{-1}\}, \qquad \rho = 1 - \alpha\mu/4, \qquad b = \rho^{-3/4} > 1.$$

The proof below identifies ρ as a return-gap contraction factor. At epoch k, use

$$\begin{aligned} M_k &= \lceil M_0 b^k \rceil, & \sigma_{e,k} &= M_k^{-1/6}, \\ h_k &= \frac{\delta}{\lceil \delta M_k^{1/3} \rceil}, & \eta_k^{\mathrm{solve}} &= M_k^{-1/3}. \end{aligned} \tag{99}$$

Here M_k is the number of trajectories in each of the two independent batches. The choices of exploration, mesh, and solve tolerance are those of Corollary D.4; take M_0 large enough for its uniform small-parameter conditions. Exact solves are allowed.

Use the expected cost $M_k \tau_{h_k}^{\mathrm{grad}}$ from (94), which already includes both groups of prefixes. Complete only epochs affordable under T and return the last actor. The number K of completed epochs is deterministic given T and the schedule, although the observed prefix durations are random.

Algorithm 3 Policy improvement with growing AdvG critic batches

Require: Known model-class data and projection Π_Φ; initial $\phi_0 \in \Phi$; expected budget T; α, δ, M_0, b and schedule (99); uniform weight χ and $\lambda_0 = \mu_\chi$ from Corollary A.6.

$k \leftarrow 0$; $C_{\mathrm{used}} \leftarrow 0$.
loop
  Compute $M_k, \sigma_{e,k}, h_k, \eta_k^{\mathrm{solve}}$ from (99); set $L_k = \delta/h_k$.
  if $C_{\mathrm{used}} + M_k \tau_{h_k}^{\mathrm{grad}} > T$ **then**
    break
  end if
  At ϕ_k, call Algorithm 2 with mesh h_k, exploration $\sigma_{e,k}$, L_k cells, tolerance η_k^{solve}, and expected budget $M_k \tau_{h_k}^{\mathrm{grad}}$. Use M_k fresh critic trajectories and M_k fresh independent prefixes.
  $\phi_{k+1} \leftarrow \Pi_\Phi(\phi_k + \alpha \widehat{g}_k)$.
  $C_{\mathrm{used}} \leftarrow C_{\mathrm{used}} + M_k \tau_{h_k}^{\mathrm{grad}}$; $k \leftarrow k + 1$.
end loop
return ϕ_k.

Theorem D.6 (Policy-return upper bound)**.** *For the class $\mathfrak{C}_{\mathrm{gd}}$ and schedule above, Algorithm 3 returns an actor ϕ_K with*

$$\sup_{P \in \mathfrak{C}_{\mathrm{gd}}} \mathbb{E}_P[J_P^* - J_P(\phi_K)] \le C T^{-2/3} \tag{100}$$

for sufficiently large expected physical budgets T. The constant depends only on the fixed class data and algorithm parameters, not on T.

Proof. We first show how gradient-estimation error affects the return after one epoch. We then sum these errors and compare the last batch with the total budget.

One ascent step with an estimated gradient. Let $\mathscr{H}_k$ contain the data collected before epoch k. Conditional on this history, ϕ_k is fixed and both batches are fresh. Apply the conditional bound in Theorem D.3 with exact features, the schedule (99), and $\tau_{h_k}^{\mathrm{grad}} \asymp 1$:

$$\mathbb{E}[\|\widehat{g}_k - \nabla J(\phi_k)\|^2 \mid \mathscr{H}_k] \le C M_k^{-2/3}. \tag{101}$$

Apply Lemma B.3 to the projected actor step, and drop its nonnegative squared-step term. This gives

$$J(\phi_{k+1}) - J(\phi_k) \ge \frac{\alpha}{8} \|\mathcal{G}_\lambda(\phi_k)\|^2 - \frac{5\alpha}{4} \|\widehat{g}_k - \nabla J(\phi_k)\|^2.$$

By gradient dominance, the first term is at least $\alpha\mu\{J^* - J(\phi_k)\}/4$. Thus, writing $\Delta_k = \mathbb{E}[J^* - J(\phi_k)]$ and taking expectations,

$$\Delta_{k+1} \le \rho\Delta_k + C\alpha M_k^{-2/3}, \qquad \rho = 1 - \alpha\mu/4. \tag{102}$$

This inequality does not require an unbiased gradient estimate. The possible bias and the sampling fluctuations both enter through the squared error already bounded in (101).

Iterating (102) gives

$$\Delta_K \le \rho^K \Delta_0 + C\alpha \sum_{k=0}^{K-1} \rho^{K-1-k} M_k^{-2/3}.$$

The factor ρ^{K-1-k} shows how later ascent steps reduce the effect of error from epoch k. Since $M_0 b^k \le M_k \le (M_0+1)b^k$, comparison with the last batch yields

$$\sum_{k=0}^{K-1} \rho^{K-1-k} M_k^{-2/3} \le C M_{K-1}^{-2/3} \sum_{j=0}^{K-1} (\rho b^{2/3})^j \le \frac{C}{1-\sqrt{\rho}}\, M_{K-1}^{-2/3}.$$

The last inequality uses our choice $b = \rho^{-3/4}$, for which $\rho b^{2/3} = \sqrt{\rho} < 1$. The initial gap is uniformly bounded, and $\rho^K \le b^{-2K/3}$, so its contribution is also $O(M_{K-1}^{-2/3})$. Hence

$$\Delta_K \le C M_{K-1}^{-2/3}.$$

In this way the earlier, cheaper estimates do not increase the order of the final error. There is no need to pay for a final-accuracy batch at every epoch.

Relating the last batch to physical time. For large enough T, at least one epoch completes. The stopping rule and the per-trajectory cost bounds imply

$$\delta M_{K-1} \le T < \sum_{k=0}^{K} M_k \tau_{h_k}^{\mathrm{grad}} \le (\delta + 2/\beta)(M_0+1)\frac{b^{K+1}-1}{b-1} \le C M_{K-1}.$$

The last inequality uses $M_{K-1} \ge M_0 b^{K-1}$ and $M_0 \ge 1$. Therefore $M_{K-1} = \Theta(T)$, which proves (100). Every prefix and every rejected solve has already been included in this cost. □

With (99), $h_k \asymp M_k^{-1/3}$. The expected number of recorded transitions is consequently

$$N = \sum_{k<K} \frac{M_k \tau_{h_k}^{\mathrm{grad}}}{h_k} \asymp \sum_{k<K} M_k^{4/3} \asymp T^{4/3}.$$

This count describes a sufficient mesh for this algorithm; the minimax budget in the theorem remains physical time T.

D.4 Shared construction and matching lower bounds

We now construct two models whose gradients and optimal actions differ, but whose trajectories provide little information about that difference. A smooth concave action response supplies the return geometry. A small signed perturbation changes its derivative and moves its maximum. The same comparison of the two observation distributions will therefore prove both lower bounds.

D.4.1 A diffusion family with exact critic features

Consider the scalar diffusion

$$dX_t = \{-X_t + F(A_t)\}\,dt + dW_t, \qquad r(X_t, A_t) = X_t, \qquad F \in \mathcal{H}_3(\mathbf{L}), \tag{103}$$

where $\mathcal{H}_3(\mathbf{L}) = \{F \in C^3(\mathbb{R}) : \|F^{(j)}\|_\infty \le L_j,\ 0 \le j \le 3\}$ for fixed positive L_j. Only the action-dependent drift F is unknown. Take reset state zero and the constant actor $\mu_\phi(x) = \phi$. Fix

any evaluation action a_0 and use the actor interval $\Phi = [a_0 - 1, a_0 + 1]$. Solving the value equation gives

$$\begin{aligned} V_F^\phi(x) &= \frac{x}{\beta+1} + \frac{F(\phi)}{\beta(\beta+1)}, \\ q_F^\phi(x,a) &= \frac{F(a) - F(\phi)}{\beta+1}, \qquad g_F^\phi(x) = \frac{F'(\phi)}{\beta+1}. \end{aligned} \tag{104}$$

Consequently,

$$J_F(\phi) = \frac{F(\phi)}{\beta(\beta+1)}, \qquad \nabla J_F(a_0) = \frac{F'(a_0)}{\beta(\beta+1)}. \tag{105}$$

Thus policy-gradient estimation contains the problem of learning the local derivative $F'(a_0)$. The value and slope have exact features $f(x) = (1, x)^\top$ and $C(x) = 1$, even though the unknown function F ranges over an infinite-dimensional smoothness class. Exact critic representation therefore does not remove the statistical difficulty.

Proposition D.7. *Fix the data $\nu_0 = \delta_0$, $\mu_\phi(x) = \phi$, $f(x) = (1, x)^\top$, and $C(x) = 1$, where δ_0 is point mass at zero. For suitable fixed $\mathcal{K}$, every $F \in \mathcal{H}_3(\mathbf{L})$ in (103) belongs to $\mathfrak{M}(\mathcal{S}, \mathcal{K}; 0, 0)$, uniformly on the actor interval.*

Proof. The exact coefficients in (104) satisfy

$$\widetilde{\theta}_\phi = \begin{pmatrix} F(\phi)/[\beta(\beta+1)] \\ 1/(\beta+1) \end{pmatrix}, \qquad \widetilde{\omega}_\phi = \frac{F'(\phi)}{\beta+1}, \qquad \|\widetilde{z}_\phi\| + \|V_F^\phi\|_{C_w^4} + \|\nabla_\phi V_F^\phi\|_{C_w^4} \le K.$$

The coefficient, reward, actor, and feature regularity follows from $F \in \mathcal{H}_3(\mathbf{L})$. For state moments, variation of constants from the reset state zero gives, under any predictable action process,

$$X_t = \int_0^t e^{-(t-s)} F(A_s)\, ds + Z_t, \qquad Z_t = \int_0^t e^{-(t-s)}\, dW_s \sim N\!\left(0, \frac{1 - e^{-2t}}{2}\right),$$

$$\sup_{t,A} \mathbb{E}|X_t|^{16} \le 2^{15} \left(L_0^{16} + \sup_t \mathbb{E}|Z_t|^{16} \right) \le K.$$

For coverage, let G be the geometric prefix length in (23) and $M_{\phi,h} = \mathbb{E}[f(X_{Gh}) f(X_{Gh})^\top]$ under the target policy. Conditional variance and the moment bound imply

$$\begin{aligned} &\det M_{\phi,h} = \mathrm{Var}(X_{Gh}) \ge \mathbb{E}\frac{1 - e^{-2Gh}}{2} = \frac{\gamma(1 - e^{-2h})}{2(1 - \gamma e^{-2h})} \longrightarrow \frac{1}{\beta+2}, \\ &\operatorname{tr} M_{\phi,h} \le K, \qquad \lambda_{\min}(M_{\phi,h}) \ge \frac{\det M_{\phi,h}}{\operatorname{tr} M_{\phi,h}} \ge \kappa_V > 0, \\ &\frac{h}{\delta} \sum_{i<L} \gamma^i = \frac{h(1 - e^{-\beta\delta})}{\delta(1-\gamma)} \ge \frac{1 - e^{-\beta\delta}}{\beta\delta} \ge \kappa_Q > 0. \end{aligned}$$

These bounds are uniform for small h and $\delta \in [\delta_{\min}, \delta_{\max}]$. Thus the class constants are independent of F, with $e_V = e_Q = 0$. □

One perturbation for both targets. Choose a smooth even function ζ, equal to one on $[-1, 1]$, zero outside $[-2, 2]$, and taking values in $[0, 1]$. Set

$$F_0(v) = -\ell \int_0^v u\zeta(u)\, du, \qquad \psi(v) = v\zeta(2v).$$

The baseline F_0 equals $-\ell v^2/2$ on $[-1, 1]$ and is bounded outside that interval; $\psi(v) = v$ for $|v| \le 1/2$ and vanishes for $|v| \ge 1$. Choose fixed $\ell, \kappa > 0$ small enough that

$$\|F_0^{(j)}\|_\infty + \kappa \|\psi^{(j)}\|_\infty \le L_j, \qquad 0 \le j \le 3.$$

Such a choice exists because F_0 and its derivatives scale with ℓ. For $s \in \{+1, -1\}$, define

$$F_{s,r}(a) = F_0(a - a_0) + s\kappa r^3 \psi((a - a_0)/r), \qquad 0 < r \le r_0. \tag{106}$$

Fix $r_0 \le 1/4$ so that

$$\kappa r_0 \|\psi''\|_\infty \le \ell/2, \qquad \kappa r_0/\ell \le 1/2.$$

Only s, r are unknown. The center, features, smoothness bounds, and constants ℓ, κ, r_0 remain fixed as T grows.

Lemma D.8. *The family (106) belongs to $\mathcal{H}_3(\mathbf{L})$ and to one class $\mathfrak{C}_{\mathrm{gd}}$ with the common data above and bounds independent of s, r. Writing $J_{s,r} = J_{F_{s,r}}$, its global action maximizer and its policy gradient at a_0 are*

$$a_s = a_0 + \frac{s\kappa r^2}{\ell}, \qquad g_s := \nabla J_{s,r}(a_0) = \frac{s\kappa r^2}{\beta(\beta+1)}. \tag{107}$$

Proof. For $0 \le j \le 3$ and $r \le r_0 \le 1$,

$$\|F_{s,r}^{(j)}\|_\infty \le \|F_0^{(j)}\|_\infty + \kappa r^{3-j}\|\psi^{(j)}\|_\infty \le L_j.$$

Proposition D.7 therefore supplies the exact critic features, value regularity, state moments, and coverage. Formula (105) gives g_s. On Φ,

$$-\frac{3\ell}{2} \le F''_{s,r} \le -\frac{\ell}{2}, \qquad |a_s - a_0| \le r/2, \qquad F'_{s,r}(a_s) = -\ell(a_s - a_0) + s\kappa r^2 = 0.$$

Thus a_s is the unique maximizer on Φ. Outside Φ the perturbation vanishes and $F_{s,r} \le -\ell/2$, whereas $F_{s,r}(a_s) \ge F_{s,r}(a_0) = 0$. It is therefore the global action maximizer.

It remains to verify projected gradient dominance. Write $J_s = J_{s,r}$, set $m_J = \ell/[2\beta(\beta+1)]$, and take a common smoothness bound $L_J \ge 3\ell/[2\beta(\beta+1)]$ on a fixed neighborhood of Φ. For $0 < \lambda \le 1/L_J$, $x \mapsto x + \lambda J'_s(x)$ fixes a_s and has derivative between zero and $1 - \lambda m_J$ on Φ. Nonexpansiveness of projection and $J'_s(a_s) = 0$ give

$$|\lambda \mathcal{G}_\lambda(x)| \ge |x - a_s| - |\Pi_\Phi(x + \lambda J'_s(x)) - a_s| \ge \lambda m_J |x - a_s|,$$
$$J_s^* - J_s(x) \le \frac{L_J}{2}|x - a_s|^2 \le \frac{L_J}{2m_J^2}|\mathcal{G}_\lambda(x)|^2.$$

This proves (98) with $\mu = m_J^2/L_J$, uniformly in s, r. □

D.4.2 Information and estimation error

A bound on the difference between the two drifts controls how well any observation rule can distinguish them. We first establish this bound for arbitrary bounded F, G, then use it for (106).

Lemma D.9 (Information in a stopped trajectory)**.** *For the diffusion (103) and any observation rule in Section D.1, let P_F, P_G be the probability distributions of all data available to the learner, including its random draws, under bounded drift functions F, G. If total duration has finite expectation under both models, then*

$$D_{\mathrm{KL}}(P_F \| P_G) \le \frac{1}{2}\mathbb{E}_F \sum_e \int_0^{\tau_e} |F(A_{e,t}) - G(A_{e,t})|^2 dt \le \frac{1}{2}\|F - G\|_\infty^2 \mathbb{E}_F \mathcal{C}. \tag{108}$$

Here D_{KL} is relative entropy between the observation laws.

Proof. Put $\Delta = F - G$. To apply Girsanov's formula over a bounded time interval, first retain at most m episodes and m units of total observation time. The duration retained from episode $e \le m$ is

$$\tau_{e,m} = \min\left\{\tau_e, \left(m - \sum_{j<e}\tau_j\right)_+\right\}, \qquad \sum_{e \le m} \tau_{e,m} \le m.$$

Here $x_+ = \max\{x, 0\}$, and $\tau_e = 0$ if episode e is not reached. Let $P_{F,m}, P_{G,m}$ be the distributions of the retained data under F, G. The integrated squared drift difference satisfies

$$Q_m := \sum_{e \le m} \int_0^{\tau_{e,m}} \Delta(A_{e,t})^2 dt \le m\|\Delta\|_\infty^2, \qquad \mathbb{E}_G e^{Q_m/2} \le e^{m\|\Delta\|_\infty^2/2} < \infty.$$

The second bound is Novikov's condition. Only the drift differs between the two models; the learner's rules are fixed, and the reward equals the observed state. Girsanov's formula gives

$$\frac{dP_{F,m}}{dP_{G,m}} = \exp\left\{\sum_{e \le m} \int_0^{\tau_{e,m}} \Delta(A_{e,t})\, dW_{e,t}^G - \frac{1}{2}Q_m\right\}.$$

Under F, use $dW^G_{e,t} = dW^F_{e,t} + \Delta(A_{e,t})dt$ and Itô isometry:

$$\log \frac{dP_{F,m}}{dP_{G,m}} = M^F_m + \tfrac{1}{2}Q_m, \qquad M^F_m := \sum_{e\le m} \int_0^{\tau_{e,m}} \Delta(A_{e,t})\, dW^F_{e,t},$$
$$\mathbb{E}_F M^F_m = 0, \qquad \mathbb{E}_F|M^F_m|^2 = \mathbb{E}_F Q_m \le m\|\Delta\|^2_\infty,$$
$$D_{\mathrm{KL}}(P_{F,m}\|P_{G,m}) = \tfrac{1}{2}\mathbb{E}_F Q_m.$$

As $m \to \infty$, the retained data recover all observations. Let $\mathscr{H}_m$ be the sigma-field generated by the retained data, representing the information they contain, and let $\mathscr{H}_\infty = \sigma(\bigcup_m \mathscr{H}_m)$ represent all observed information. Also, $Q_m \uparrow Q := \sum_e \int_0^{\tau_e} \Delta(A_{e,t})^2 dt$. For bounded $\mathscr{H}_\infty$-measurable u, set $R = (P_F + P_G)/2$ and $u_m = \mathbb{E}_R[u \mid \mathscr{H}_m]$. Martingale convergence gives $u_m \to u$ under both laws and $|u_m| \le \|u\|_\infty$. The entropy variational formula and dominated convergence yield

$$\begin{aligned}\mathbb{E}_F u - \log \mathbb{E}_G e^u &= \lim_m \{\mathbb{E}_F u_m - \log \mathbb{E}_G e^{u_m}\} \\ &\le \lim_m D_{\mathrm{KL}}(P_{F,m}\|P_{G,m}) \le D_{\mathrm{KL}}(P_F\|P_G).\end{aligned}$$

Taking the supremum over u makes both inequalities equalities. Monotone convergence now gives

$$\begin{aligned}D_{\mathrm{KL}}(P_F\|P_G) &= \lim_m \tfrac{1}{2}\mathbb{E}_F Q_m = \tfrac{1}{2}\mathbb{E}_F Q \\ &\le \tfrac{1}{2}\|\Delta\|^2_\infty \mathbb{E}_F \mathcal{C} < \infty.\end{aligned}$$

This proves (108); retaining only grid observations cannot increase relative entropy. □

The shared information bound. Fix any observation rule with expected duration at most T under each of the two models (106). Let P_+, P_- be its data distributions, including any randomness used by its output. Since the baseline cancels, Lemma D.9 gives

$$\begin{aligned}\|F_{+,r} - F_{-,r}\|_\infty &= 2\kappa r^3 \|\psi\|_\infty, \\ D_{\mathrm{KL}}(P_+\|P_-) &\le 2\kappa^2 \|\psi\|^2_\infty T r^6.\end{aligned} \tag{109}$$

Choose

$$c_0 = \min\{r_0, (16\kappa^2\|\psi\|^2_\infty)^{-1/6}\}, \qquad r = c_0 T^{-1/6}, \qquad T \ge 1. \tag{110}$$

Then $r \le r_0$, the divergence is at most $1/8$, and Pinsker's inequality gives

$$\|P_+ - P_-\|_{\mathrm{TV}} := \sup_B |P_+(B) - P_-(B)| \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P_+\|P_-)} \le \frac{1}{4}.$$

Let v_+, v_- be two possible target vectors, with $d = \|v_+ - v_-\|$, and let $\widehat{v}$ be any estimate. Choose the nearer target, selecting $+$ on $B = \{\|\widehat{v} - v_+\| \le \|\widehat{v} - v_-\|\}$. A wrong choice forces an error of at least $d/2$. Consequently,

$$\begin{aligned}\max_{s=\pm 1} \mathbb{E}_s\|\widehat{v} - v_s\|^2 &\ge \frac{d^2}{8}\{P_+(B^c) + P_-(B)\} \\ &\ge \frac{d^2}{8}(1 - \|P_+ - P_-\|_{\mathrm{TV}}) \ge \frac{3d^2}{32}.\end{aligned} \tag{111}$$

The last inequality uses the preceding choice of r. We now apply this formula to the gradient and to the maximizing action.

D.4.3 The two minimax conclusions

Theorem D.10 (Gradient-estimation lower bound)**.** *For sufficiently large T, every acquisition rule with $\sup_{F\in\mathcal{H}_3(\mathbf{L})} \mathbb{E}_F\mathcal{C} \le T$ and every estimate $\widehat{g}$ of the policy gradient at a_0 satisfy*

$$\sup_{F\in\mathcal{H}_3(\mathbf{L})} \mathbb{E}_F|\widehat{g} - \nabla J_F(a_0)|^2 \ge cT^{-2/3}. \tag{112}$$

The learner may know the common OU drift term, diffusion coefficient, and reward, and may use the adaptive full-path observations above.

Proof. Use the models (106) with the radius in (110). Their target gradients in (107) are separated by $d = 2\kappa r^2/[\beta(\beta+1)]$. Applying (111) with $v_s = g_s$ gives

$$\max_{s=\pm 1} \mathbb{E}_s|\widehat{g} - g_s|^2 \geq \frac{3\kappa^2 r^4}{8\beta^2(\beta+1)^2} = \frac{3\kappa^2 c_0^4}{8\beta^2(\beta+1)^2}\, T^{-2/3}.$$

Both models belong to the fixed $\mathcal{H}_3(\mathbf{L})$ by Lemma D.8. Taking the supremum over that class proves (112) for every observation rule and estimator. □

Corollary D.11 (Minimax policy-gradient risk)**.** *Fix the known data of Proposition D.7, evaluation parameter $\phi = a_0$, and common bounds $\mathcal{K}$ large enough to include its entire OU family. For the exact-feature class $\mathfrak{M} = \mathfrak{M}(\mathcal{S}, \mathcal{K}; 0, 0)$,*

$$cT^{-2/3} \;\leq \inf_{\text{acquisition rule},\, \widehat{g}} \sup_{M \in \mathfrak{M}} \mathbb{E}_M \|\widehat{g} - \nabla J_M(a_0)\|^2 \;\leq\; CT^{-2/3}.$$

The infimum is over rules with expected total physical cost at most T in every model, allowing adaptive actions and continuous path observations as above.

Proof. Corollary D.4 supplies one admissible estimator with the uniform upper bound, using only grid observations. Proposition D.7 places the whole OU family inside this same fixed class. Restricting any estimator to that family gives the lower bound in Theorem D.10, even under the full-path observation scheme. □

Theorem D.12 (Minimax return loss)**.** *Fix the known data of Lemma D.8 and common class constants chosen so that $\mathfrak{C}_{\mathrm{gd}}$ contains its entire family. For sufficiently large expected physical budgets T,*

$$cT^{-2/3} \leq \inf_{\mathsf{A}:\, \sup_{P \in \mathfrak{C}_{\mathrm{gd}}} \mathbb{E}_P \mathcal{C} \leq T} \sup_{P \in \mathfrak{C}_{\mathrm{gd}}} \mathbb{E}_P[J_P^* - J_P(\widehat{\phi}_{\mathsf{A}})] \leq CT^{-2/3}. \tag{113}$$

Here A ranges over the adaptive full-path acquisition rules of Section D.1, including randomized outputs $\widehat{\phi}_{\mathsf{A}} \in \Phi$. The cost $\mathcal{C}$ includes all acquired physical time, including prefixes.

Proof. Theorem D.6 gives the upper bound. For the lower bound, use (106) and (110). Write $J_s = J_{s,r}$. The curvature and stationarity established in Lemma D.8 imply

$$J_s^* - J_s(\widehat{\phi}) \geq \frac{\ell}{4\beta(\beta+1)} |\widehat{\phi} - a_s|^2, \qquad \widehat{\phi} \in \Phi. \tag{114}$$

The maximizers in (107) are separated by $d = 2\kappa r^2/\ell$. Use (111) with $v_s = a_s$ and combine it with (114):

$$\max_{s=\pm 1} \mathbb{E}_s[J_s^* - J_s(\widehat{\phi})] \geq \frac{3\ell d^2}{128\beta(\beta+1)} = \frac{3\kappa^2 c_0^4}{32\ell\beta(\beta+1)}\, T^{-2/3}.$$

Both alternatives belong to $\mathfrak{C}_{\mathrm{gd}}$, with known data and bounds independent of T. Since the observation rule and returned actor were arbitrary, this proves (113). □

The common construction explains the two rates:

$$\|F_{+,r} - F_{-,r}\|_\infty \asymp r^3, \qquad |g_+ - g_-| \asymp |a_+ - a_-| \asymp r^2.$$

Time T supplies at most order Tr^6 information about the sign. At $r \asymp T^{-1/6}$, the resulting squared gradient error is of order r^4; curvature makes the return loss of that same order. The upper bounds hold on their respective classes, whereas each matching lower bound uses the fixed class containing this family.

D.5 CONTINUED LEARNING AFTER DEPLOYMENT

The preceding return bound concerns an actor fixed after training. A deployed controller might instead keep learning from new observations. For the same family, the lower bound survives this extra information: discounted return can be evaluated through an independent exponential observation time, whose mean is only $1/\beta$.

After training, allow any policy $\widehat{\pi}$ with real-valued actions predictable from the training data, subsequent state observations, and model-independent randomization. Deployment starts at zero on one fresh trajectory, with independent Brownian noise and no resets. Write $J_{s,r}(\widehat{\pi})$ for its expected discounted return conditional on the training data and policy selection; the outer expectation below averages that training stage.

Corollary D.13 (Return loss with continued learning)**.** *Every such procedure with* $\sup_{s=\pm 1,\, 0<r\le r_0} \mathbb{E}_{s,r}\mathcal{C} \le T$ *satisfies*

$$\sup_{s=\pm 1,\, 0<r\le r_0} \mathbb{E}_{s,r}[J^*_{s,r} - J_{s,r}(\widehat{\pi})] \ge cT^{-2/3} \tag{115}$$

for sufficiently large T. On this family, $J^*_{s,r} = F_{s,r}(a_s)/[\beta(\beta+1)]$ is optimal among all admissible controls, including history-dependent ones. The budget T counts training only; the deployed policy may continue learning.

Proof. We apply (111) to an action chosen during deployment. Fix r and write $F_s = F_{s,r}$, $J_s = J_{s,r}$, and $\mathbb{E}_s = \mathbb{E}_{s,r}$.

The variation-of-constants formula and state-moment bound in the proof of Proposition D.7 apply to any predictable actions. Taking expectations and using Fubini gives

$$\mathbb{E}_s J_s(\widehat{\pi}) = \frac{1}{\beta+1}\mathbb{E}_s \int_0^\infty e^{-\beta u} F_s(A_u)\, du \le \frac{F_s(a_s)}{\beta(\beta+1)}.$$

Boundedness of F_s justifies the interchange. The constant action a_s attains equality, so this is the optimum even among controls that use the whole observed history.

Let $\tau_\beta \sim \mathrm{Exp}(\beta)$ be independent of training and deployment and unrevealed to the policy. Its density $\beta e^{-\beta t}$ rewrites the return gap as

$$\mathbb{E}_s[J^*_s - J_s(\widehat{\pi})] = \frac{\mathbb{E}_s[F_s(a_s) - F_s(A_{\tau_\beta})]}{\beta(\beta+1)}. \tag{116}$$

The random time is only a proof device; the policy still runs over an infinite discounted horizon.

Information from training and deployment. Let P_s be the distribution of the training data and Q_s the distribution including the independent clock τ_β, the deployment observations through that time, and the controller's random draws. Apply Lemma D.9 to training and, conditional on the training data, to the single deployment trajectory. The relative-entropy chain rule and $\mathbb{E}\tau_\beta = 1/\beta$ give

$$\begin{aligned} D_{\mathrm{KL}}(Q_+\|Q_-) &\le \frac{1}{2}\|F_+ - F_-\|^2_\infty(\mathbb{E}_+\mathcal{C} + 1/\beta) \\ &\le 2\kappa^2\|\psi\|^2_\infty r^6(T + 1/\beta). \end{aligned}$$

Use the constant c_0 in (110) and take $r = c_0(T+1/\beta)^{-1/6}$. For sufficiently large T, $r \le r_0$ and $\|Q_+ - Q_-\|_{\mathrm{TV}} \le 1/4$.

Outside $\Phi = [a_0 - 1, a_0 + 1]$ the perturbation vanishes and $F_s(a) \le F_s(\Pi_\Phi a)$. On Φ, $F_s'' \le -\ell/2$, so for every real a,

$$F_s(a_s) - F_s(a) \ge \frac{\ell}{4}|\Pi_\Phi a - a_s|^2.$$

Apply (111) to $\widehat{a} = \Pi_\Phi A_{\tau_\beta}$ and $v_s = a_s$. The projection is used only in this error comparison; the controller may take any real action. With $d = 2\kappa r^2/\ell$,

$$\max_s \mathbb{E}_s|\widehat{a} - a_s|^2 \ge \frac{d^2}{8}(1 - \|Q_+ - Q_-\|_{\mathrm{TV}}) \ge \frac{3d^2}{32}.$$

Combining this with (116) proves

$$\max_s \mathbb{E}_s[J_s^* - J_s(\widehat{\pi})] \geq \frac{3\ell d^2}{128\beta(\beta+1)} \geq c(T+1/\beta)^{-2/3} \geq c' T^{-2/3}$$

for sufficiently large T, as required by (115). □

E NUMERICAL EXPERIMENTS

In this section we demonstrate the experiment compare AdvG with continuous- and discrete-time actor–critic methods on coupled stochastic LQ control and a noisy pendulum. Figures 5 and 6 show how policy performance develops during training and how it changes when the observation interval is reduced from $h = 0.02$ to $h = 0.005$. Within each task, both intervals receive the same physical training budget. The implementations use learned neural critics.

Evaluation. For an episode of duration H, we evaluate the discounted return

$$J_h(\mu) = \mathbb{E}_\mu\left[\sum_{i=0}^{H/h-1} \gamma^i R_i\right], \qquad \gamma = e^{-\beta h} = 0.8^h, \quad \beta = -\log(0.8),$$

where R_i is the reward accumulated during cell i. Larger returns are better. Evaluation uses deterministic actions, including SAC's mean action, while retaining process noise and random initial states. The horizontal axis counts pooled physical training time Nh across four environments. Each curve averages all prescribed training seeds. Final tables report mean ± sample standard deviation across trained policies on a separate evaluation bank, without selecting intermediate checkpoints.

E.1 COUPLED STOCHASTIC LQ CONTROL

Environment. The state has dimension $d_s = 4$, the action lies in $[-5, 5]^2$, and

$$dX_t = (AX_t + Ba_t)\,dt + 0.5\,dW_t, \qquad r(x,a) = -\tfrac{1}{2}(x^\top Q_c x + a^\top R_c a),$$

where

$$A = \begin{pmatrix} -.12 & 1 & 0 & .25 \\ -1 & -.20 & .35 & 0 \\ 0 & -.35 & -.15 & .9 \\ -.25 & 0 & -.9 & -.18 \end{pmatrix}, \qquad B = \begin{pmatrix} .3 & .1 \\ 1 & .35 \\ .15 & .4 \\ -.3 & 1 \end{pmatrix},$$

$$Q_c = \begin{pmatrix} 2 & .25 & .15 & 0 \\ .25 & .7 & 0 & .1 \\ .15 & 0 & 1.5 & -.2 \\ 0 & .1 & -.2 & .6 \end{pmatrix}, \qquad R_c = \begin{pmatrix} .15 & .025 \\ .025 & .20 \end{pmatrix}.$$

We use $H = 10$, zero terminal reward, and $X_0 \sim \mathcal{N}(0, \mathrm{diag}(1, .5, 1, .5))$. The simulator draws the exact Gaussian transition for a held action, with mean $e^{Ah}X_i + \int_0^h e^{Au}B\,du\,a_i$ and covariance $\frac{1}{4}\int_0^h e^{Au}e^{A^\top u}\,du$. Rewards use the left-endpoint quadrature $R_i = h\,r(X_i, a_i)$. Learners receive no dynamics matrices, Riccati solutions, analytic critic targets, or prescribed quadratic features.

Methods and training settings. We compare AdvG with released full-window CT-DDPG (Cheng et al., 2025), one-step DDPG (Lillicrap et al., 2016), TD3 (Fujimoto et al., 2018), and SAC (Haarnoja et al., 2018). All actor and critic networks have two width-32 ReLU hidden layers, and all methods use Adam with actor and critic learning rates 3×10^{-4}. Deterministic actors use a bounded $5\tanh(\cdot)$ output, and critics have linear outputs. AdvG initializes its advantage output head at zero. Deterministic methods use Gaussian exploration with standard deviation 0.5 in physical action units, clipping actions to $[-5, 5]^2$.

The physical training budget is 800 per policy: 40,000 transitions at $h = 0.02$ and 160,000 at $h = 0.005$. Replay capacities match these transition counts, and collection warmup lasts 2,000 or 8,000 transitions, respectively. Each method makes 9,500 learner calls with minibatches of size 128,

collecting four or sixteen new transitions per call at the two meshes. AdvG retains a normalized actor movement cap of 2×10^{-4}.

For the discrete-time baselines, target-network coefficients are 0.005. TD3 uses an actor delay of two and target smoothing noise with standard deviation 0.2, clipped at 0.5, in normalized action coordinates. SAC uses automatic entropy adjustment with target entropy -2 and initial temperature $0.2h$. Adam uses moment coefficients $(0.9, 0.999)$ and $\epsilon = 10^{-8}$, with no weight decay or learning-rate decay during these runs.

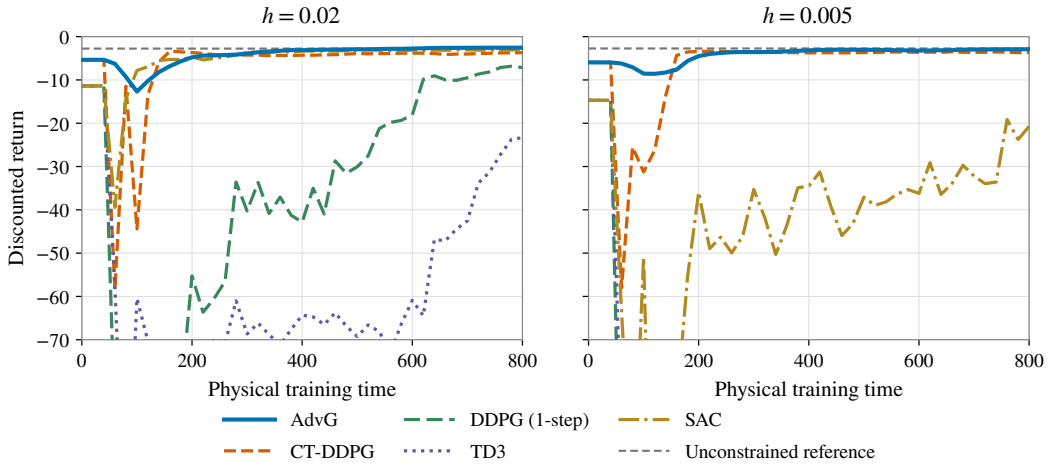


Figure 5: Coupled stochastic LQ learning curves at $h = 0.02$ and $h = 0.005$, with physical training budget 800. Curves use all 41 scheduled checkpoints without smoothing, averaging three training seeds and five evaluation episodes per seed. Gaps occur where curves fall below the displayed lower limit of -70. The dashed reference is the exact expected return of the corresponding unconstrained finite-horizon Riccati problem. Final results on a separate evaluation bank are reported in Table 1.

Method	$h = 0.02$	$h = 0.005$
AdvG	-3.166 ± 0.076	-3.390 ± 0.218
CT-DDPG	-4.149 ± 1.104	-4.474 ± 0.730
DDPG (one-step)	-7.489 ± 5.129	-135.631 ± 95.398
TD3	-26.350 ± 18.689	-155.982 ± 94.340
SAC	-3.445 ± 0.147	-20.653 ± 17.234

Table 1: LQ final discounted returns: mean $\pm$ sample SD across three training seeds, with fifty separate evaluation episodes per final policy. All methods receive physical training budget 800.

Results. Figure 5 shows that AdvG recovers from an initial decline and learns a strong policy at both observation intervals. Although all methods menage to learn in the $h = 0.02$ setting, at the finer mesh $h = 0.005$, one-step DDPG and TD3 remain below the displayed range for most of training. On the contrary, AdvG and CT-DDPG recover from early fluctuation and continue improving. AdvG attains the highest final mean among the methods in Table 1.

We also notice that compared to CT-DDPG, AdvG is able to continuously learn and steadily improve returns even in later stages of the learning, while the CT-DDPG returns stale. On the separate final evaluation bank, AdvG's mean discounted cost is 23.7% lower than CT-DDPG's at $h = 0.02$ and 24.2% lower at $h = 0.005$. Its final return changes from -3.166 to -3.390 under the fourfold mesh refinement, whereas SAC, one-step DDPG, and TD3 deteriorate substantially under their fixed settings. This experiment demonstrates the failure of discrete-time RL algorithms for continuous-time settings, thus emphasizing the need for reliably convergent and generally applicable continuous-time RL algorithms, which is a central motivation for proposing AdvG.

E.2 NOISY PENDULUM WITH ALL-RELU NETWORKS

Environment. The pendulum has torque $a_t \in [-2, 2]$ and dynamics

$$d\vartheta_t = \omega_t\, dt, \qquad d\omega_t = (15 \sin \vartheta_t + 3a_t - 0.25\omega_t)\, dt + 0.5\, dW_t.$$

Its running reward is $r(\vartheta, \omega, a) = -(\mathrm{wrap}(\vartheta)^2 + 0.1\omega^2 + 0.001a^2)$, where the reward angle is wrapped to $[-\pi, \pi)$. Physical angles and velocities are not clipped. Initial coordinates are independent and uniform on $[-\pi, \pi]$ and $[-1, 1]$, respectively. Episodes last $H = 4$ and have zero terminal continuation value. Each held-action interval contains two explicit Euler–Maruyama microsteps of size $h/2$, using the old state for both drift components and independent velocity-noise increments of standard deviation $0.5\sqrt{h/2}$. The cell reward sums $(h/2)r$ at the two microstep left endpoints.

Networks and hyperparameters. AdvG, CT-DDPG, SAC, one-step DDPG, and TD3 all use two width-128 ReLU hidden layers in every actor and critic. Deterministic actors retain the bounded $2\tanh(\cdot)$ output, critics have linear outputs, and SAC has an additional variance head. Adam actor and critic learning rates are 10^{-4} and 3×10^{-4}, minibatch size is 64, and gradient norms are capped at 10 per optimizer, jointly for twin critics. There is no input standardization. Deterministic exploration has standard deviation 0.3 in physical torque units with action clipping; warmup uses the same Gaussian perturbations for all methods, including SAC's mean policy. Target-network coefficients for SAC, DDPG, and TD3 are 0.005. SAC starts with temperature $0.2h$ and target entropy -1; TD3 uses the delay and smoothing settings given above.

Each policy receives physical training time 2000: 100,000 transitions at $h = 0.02$ or 400,000 at $h = 0.005$. Replay capacities are 20,000 and 80,000 transitions, and warmup lasts 2,000 and 8,000 transitions, respectively. A learner call follows every eight or thirty-two new transitions, giving 12,250 calls at either mesh. AdvG and CT-DDPG sample one window length uniformly from 1 to 10 or 1 to 40 per minibatch, so the maximum duration is 0.2; they make four value fits and one advantage fit per call, with terminal-value penalty 0.002.

For both AdvG and CT-DDPG, actor updates start at physical acquisition time 400. The normalized actor movement cap increases linearly from 5×10^{-4} at time 400 to 3×10^{-3} at time 1200 and remains fixed thereafter. SAC, DDPG, and TD3 use their usual actor schedules without this hold or movement cap. These baselines share the inherited training configuration.

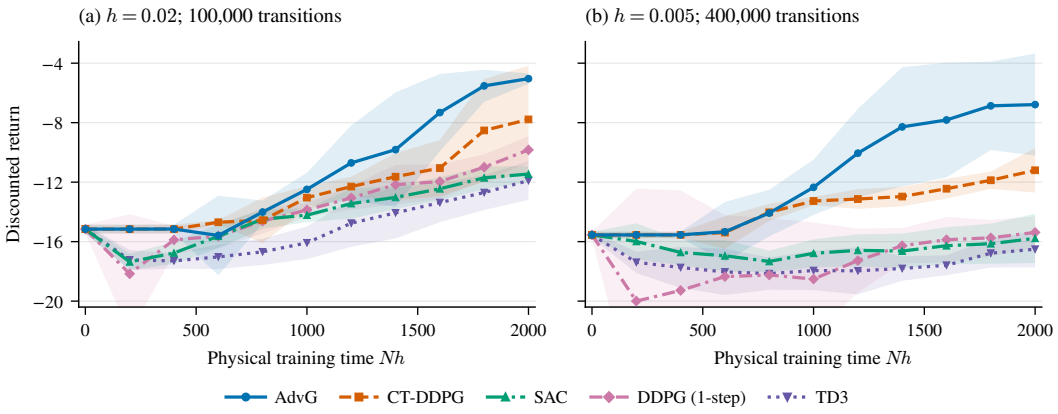


Figure 6: All-ReLU noisy-pendulum learning curves at $h = 0.02$ and $h = 0.005$, with the same physical acquisition and discounted-return axes. Curves average five training seeds and twenty evaluation episodes per seed and checkpoint; bands are pointwise 95% Student-t intervals across training seeds. No smoothing is applied. Final results in Table 2 use a separate 100-episode bank.

Results and interpretation. Figure 6 shows sustained improvement by AdvG at both observation intervals. At $h = 0.02$, AdvG attains the highest final mean return among the five methods and has a relatively narrow across-seed spread. The final discounted return is -7.66 ± 0.38, compared with -9.80 ± 3.01 for CT-DDPG, and the three discrete-time reference methods have lower means

	Discounted return J_h		Undiscounted score U_h	
Method	$h = 0.02$	$h = 0.005$	$h = 0.02$	$h = 0.005$
AdvG	-7.66 ± 0.38	-9.63 ± 3.48	-497.3 ± 36.8	-674.3 ± 312.7
CT-DDPG	-9.80 ± 3.01	-13.02 ± 1.06	-682.6 ± 261.4	-958.0 ± 84.5
SAC	-13.37 ± 0.34	-17.91 ± 1.12	-991.2 ± 35.6	-1387.0 ± 98.7
DDPG (one-step)	-11.54 ± 0.86	-17.07 ± 0.92	-837.0 ± 71.0	-1324.5 ± 77.2
TD3	-13.73 ± 0.57	-18.27 ± 0.84	-1032.0 ± 43.9	-1418.1 ± 68.2

Table 2: Pendulum final performance: mean $\pm$ sample SD across five training seeds, with 100 evaluation episodes per final policy and physical training budget 2000. The supplementary score is $U_h = \mathbb{E}[\sum_i R_i]/0.02$, using the same reference scale at both meshes. Larger values are better.

(Table 2). This result demonstrates effective nonlinear control with ReLU critics, without task-specific quadratic heads or additional simulator branches.

At $h = 0.005$, AdvG again attains the highest final mean under both return metrics. Its discounted return is -9.63 ± 3.48, compared with -13.02 ± 1.06 for CT-DDPG. On the common scheduled-evaluation bank, AdvG improves from initialization in all five training seeds at both meshes. The result therefore supports that AdvG successfully learns across the two different regimes of interval length h, with obvious advantage for settings finer mesh. In these settings, learning action-effects against noise is more difficult and AdvG's cell pairing design helps extract more information about the actions from the local behavior of advantage functions. The favorable AdvG means are consistent with a benefit from preserving the local action–residual association while retaining multi-step value learning.

F Related work and methodological connections

In this section, we discuss existing research that are related to our paper, better reflecting the novelty and contributions of our work. We review the literature underlying AdvG's design and theoretical analysis, beginning with continuous-time value and policy-gradient methods. These methods explain the role of the advantage rate in deterministic policy improvement and provide the foundation for AdvG's estimation of its action derivative. We then examine convergence results for policy evaluation, policy iteration, and actor–critic learning, emphasizing critic stability and the additional challenges of updating the actor and critic simultaneously. Finally, we discuss regret and gradient-estimation guarantees, distinguishing their error criteria, observation models, and acquisition budgets to clarify how our minimax results contribute to existing statistical theory.

F.1 Value and policy-gradient learning in RL

Deterministic policy-gradient methods improve a policy using the action derivative of a learned critic (Silver et al., 2014). This principle underlies DDPG and TD3 (Lillicrap et al., 2016; Fujimoto et al., 2018), with TD3 introducing twin critics, delayed actor updates, and target-policy smoothing to address function-approximation error. Extending this approach to continuous time requires an action-dependent learning signal that remains meaningful as the observation interval decreases.

For a fixed deterministic policy μ, let $Q_h^\mu(x, a)$ denote the expected discounted return from holding action a for time h and then following μ. Under the regularity used here,

$$Q_h^\mu(x, a) = V^\mu(x) + h\, q^\mu(x, a) + o(h).$$

The action-dependent part of Q_h^μ therefore vanishes with h, whereas the advantage rate q^μ describes the effect of changing the action per unit time. Advantage-based methods address this scaling problem in discretized control (Tallec et al., 2019), and continuous-time q-learning develops trajectory-based equations for learning the corresponding rate function (Jia & Zhou, 2023). Early continuous-time policy-gradient work considers deterministic dynamics (Munos, 2006); CT-DDPG (Cheng et al., 2025) establishes the deterministic policy-gradient identity for controlled diffusions, identifying the action derivative of the advantage rate as the quantity needed for policy improvement.

AdvG builds on this deterministic policy-gradient formulation and introduces a different way to learn the required derivative. CT-DDPG uses multi-step temporal differences over physical observation windows to stabilize value learning and fits its advantage network through an accumulated window residual. AdvG retains multi-step value estimation but learns the local derivative $C(x)^\top \omega$ by pairing each exploratory action deviation with the residual from the same observation interval. Forming these products before summation preserves the association between an action perturbation and its observed response. Appendix C quantifies the benefit of this construction relative to the released full-window CT-DDPG gradient, using fixed features and a common value equation.

Local regression provides motivation for the paired equation. With the actor and value coefficients held fixed, let $U_i = A_i - \mu_\phi(X_i)$ and use the residual $D_i(z)$ from (8). Then

$$\nabla_\omega \frac{1}{2} \sum_{i<L} \gamma^i \big(D_i(z)/h\big)^2 = \frac{1}{h} \sum_{i<L} \gamma^i C(X_i) U_i D_i(z) = \frac{\delta \sigma_e^2}{h}\, m_Q^{\mathrm{p}}(z).$$

Thus the paired update estimates the advantage derivative through local rate regression. Related residual-based constructions appear in the martingale test-function updates of Jia & Zhou (2023) and the generator-based regression of Nguyen (2026). AdvG makes this pairing a central element of an algorithm that combines multi-step value learning, local advantage-derivative estimation, and simultaneous deterministic actor updates. This construction enables the first finite-time analysis of joint actor and critic convergence in model-free deterministic policy-gradient learning for nonlinear controlled diffusions.

AdvG also learns a separate state-value function, a design used in the original SAC formulation (Haarnoja et al., 2018). This choice matters when constructing short-time residuals: replacing endpoint values by fresh independent action averages introduces sampling variances multiplied by factors of order h^{-2} after division by the observation interval. Direct value evaluation avoids this additional action-integration noise, while its approximation and estimation errors remain part of our analysis.

Other continuous-time methods combine value estimation with different policy-improvement rules. Martingale actor–critic and q-learning methods support entropy-regularized stochastic policies (Jia & Zhou, 2022b; 2023), while continuous-time performance-difference identities support policy-gradient, TRPO, and PPO updates (Zhao et al., 2023). Hamiltonian-flow methods use fitted critics for Gibbs or greedy improvement and include CT-SAC and CT-TD3 implementations (Nguyen, 2026). Optimal-PhiBE constructs PDE approximations from discrete transitions for policy iteration (Zhu et al., 2025). These approaches offer different combinations of critic estimation and policy improvement.

F.2 Convergence of actor–critic methods

We now review convergence guarantees for actor–critic methods, focusing on how critic estimation and policy improvement can be controlled jointly. In discrete-time MDPs, Xiong et al. (2022) establish nonasymptotic stationarity bounds for deterministic policy gradients with linear Q approximation, covering on-policy and corrected off-policy TD updates and retaining approximation errors. Chen & Zhao (2025) analyze simultaneous critic tracking and actor stationarity for single-timescale stochastic-policy updates with a linear value critic and Markovian sampling. Their continuous state and action spaces still belong to a discrete-time MDP. For continuous-time stochastic control, Zhou & Lu (2024) prove global linear convergence of a coupled actor–critic function-space flow. Their actor uses the model Hamiltonian, and the analysis requires action-independent uniformly elliptic diffusion, strong Hamiltonian concavity, regularity, and an additional condition relating objective gaps to distances from the optimal control. In a structured continuous-time optimal-execution problem, Wang et al. (2026) derive error bounds and establish linear convergence for an actor–critic scheme that alternates policy evaluation and improvement with critic recalibration. For neural actor–critic methods applied to Hamilton–Jacobi–Bellman equations, Cohen et al. (2025) analyze the infinite-width limit of the coupled training dynamics and give conditions under which its stationary points solve the stochastic control problem. These results establish important precedents for joint learning, while leaving distinct challenges in analyzing sampled deterministic actor–critic updates for unknown diffusions.

A necessary foundation for such an analysis is understanding how the critic learns from diffusion observations. For fixed-policy evaluation, Kobeissi & Bach (2023) establish convergence rates for model-free and model-based TD methods with vanishing observation intervals. They also derive positivity of the linear value-update matrix from the dynamics and feature geometry, providing a structural explanation for critic stability. Settai et al. (2025) prove exponential convergence of a differential-TD value-function flow under ellipticity and coercivity conditions on a bounded domain. Their guarantee concerns an idealized function-space evolution; finite-dimensional approximation and sampled updates introduce additional errors. Together, these works explain how continuous-time dynamics can support stable value learning, which is one component of the joint value and advantage-derivative estimation required by AdvG.

Another line of research combines critic fitting with policy improvement. Mou (2026) derive operator positivity and finite-sample oracle inequalities for off-policy fitted value and advantage learning under action-independent uniformly elliptic diffusion, action coverage, and regular convex approximation classes. Nguyen (2026) analyze Hamiltonian-flow and fitted-critic procedures with Gibbs or greedy improvement, including finite-sample bounds that propagate critic-fitting errors under sampling, coverage, and contraction assumptions. For entropy-regularized nonlinear control, Tang & Zhou (2024) obtain quantitative bounds for coupled value/q updates under action-independent diffusion and approximation, moment, and stability conditions. These results account for interactions between evaluation and improvement through their respective update rules. AdvG requires a further analysis of critic tracking when a separately parameterized deterministic actor takes gradient steps using the critic's current estimate.

The central difficulty is to establish stability of the coupled critic equations while their target changes with the actor. Some analyses impose positivity or dissipativity conditions on the mean learning dynamics (Chen & Zhao, 2025; Tang & Zhou, 2024), whereas Kobeissi & Bach (2023) and Mou (2026) derive the relevant positivity in their settings. Our key observation concerns the particular interaction between multi-step value estimation and local action–residual pairing. For each fixed actor, centered exploration makes the expected value update independent of the learned advantage-derivative coefficients, while pairing gives the derivative equation a positive-definite feature Gram matrix under feature coverage. The resulting dependence between the two critic equations is one-way: value errors affect derivative learning, but derivative errors do not feed back into the expected value update. Bounding the remaining coupling and choosing a suitable relative weight for the updates therefore yields joint critic stability uniformly as the observation interval decreases. Theorem 4.1 establishes this property for nonlinear controlled diffusions, including action-dependent noise, and shows that refining the observation mesh introduces no additional $1/h$ factor in the critic-noise bound when the physical window duration and exploration scale remain fixed. This structure provides the control needed to analyze a critic whose fitted target moves with the actor. Following this inspiration, Theorem 4.2 establishes a finite-time guarantee for simultaneous stochastic actor and critic updates with fixed critic features and fresh batches.

To summarize, our work shows how an algorithmic design choice can resolve a theoretical obstacle to joint learning and provides a foundation for understanding when continuous-time actor–critic methods can learn stably from trajectory observations.

Table 3 summarizes the scope of selected related results.

F.3 Minimax Properties in RL

The preceding convergence guarantees describe what particular learning procedures achieve. We now turn to the statistical limits of learning: what's the smallest worst-case error attainable from a given amount of interaction? Establishing such limits requires an estimator that attains an upper bound and a matching lower bound that applies to every admissible procedure. For controlled diffusions, the observation model is central to this comparison. Observing a trajectory more frequently increases the number of recorded transitions without increasing its physical duration, and the learner may adapt its actions and observation schedule to the data. Our minimax results account for these choices when studying policy-gradient estimation and policy-return accuracy.

Continuous-time learning already has various statistical guarantees, particularly for regret in structured control problems. Basei et al. (2022) study episodic linear–quadratic control with unknown drift coefficients. Under identification conditions on the initial and optimal feedback con-

Work	Learning setting	Convergence guarantee
Kobeissi & Bach (2023)	Fixed-policy diffusion evaluation with linear features and vanishing observation intervals.	Value TD convergence, with positivity derived for the mean update.
Mou (2026)	Off-policy fitted value/advantage learning with action-independent elliptic diffusion and regular function classes.	Finite-sample oracle inequalities for fitted value and advantage errors.
Chen & Zhao (2025)	Discrete-time stochastic actor–critic with linear value features and Markovian sampling.	Joint critic tracking and actor stationarity under single-timescale updates.
Zhou & Lu (2024)	Continuous-time actor–critic function-space flow using the model Hamiltonian.	Global linear convergence under ellipticity, strong Hamiltonian concavity, and additional regularity and error-bound conditions.
AdvG, Theorem 4.2	Sampled deterministic actor–critic with fixed features, fresh batches, and possibly action-dependent diffusion.	Joint critic tracking and projected actor stationarity, with explicit sampling, approximation, exploration, and discretization errors.

Table 3: Scope of selected convergence guarantees. Each result holds under the assumptions of the corresponding work. Flow time, update count, and physical acquisition time measure different resources, so the rates are not directly interchangeable.

trols, their least-squares learning and control procedure achieves high-probability regret of order $\log N \log\log N$ over N episodes at fixed confidence, with additional bounds for discrete observations and piecewise-constant controls. Guo et al. (2023) extend model-based learning to linear jump diffusions with general convex costs and strong convexity in the control. Combining feedback stability with identification and tail conditions yields high-probability regret of order $\sqrt{N \log N}$ at fixed confidence. Later model-free analyses address Gaussian policies in a specialized LQ setting (Huang et al., 2024) and Gibbs policies for entropy-regularized nonlinear control (Tang & Zhou, 2024). These contributions establish statistical guarantees under different dynamics and policy classes; their cumulative regret objectives also differ from the gradient-estimation and final-policy errors considered here.

For finite-state and finite-action continuous-time MDPs, Gao & Zhou (2025) obtain episodic regret bounds with matching square-root dependence up to logarithmic factors. In continuous state and action spaces, the available observations further determine the learning problem. Zhao et al. (2025b) study general function approximation with known diffusion and noisy measurements of drift and reward, whereas Zhao et al. (2025a) fit state distributions from discrete observations and analyze how regret depends on reward variability and measurement resolution. Croissant et al. (2024) study near-continuous interaction with controlled jump dynamics, separating learning error from diffusion-approximation error. These results motivate specifying both the observations and their acquisition cost before comparing statistical rates.

For policy-gradient estimation, the closest comparisons come from inference using logged discrete-time data. Kallus & Uehara (2020b) derive semiparametric efficiency bounds and an attaining estimator for differentiable stochastic policies under overlap and nuisance-estimation conditions. For deterministic policies, Kallus & Uehara (2020a) use kernelized doubly robust estimation and obtain $O(n^{-4/7})$ asymptotic mean-squared error for marginal policy-gradient estimation under their smoothness and nuisance-rate assumptions; minimax optimality is conjectured rather than proved. Our setting instead permits the learner to acquire trajectories from an unknown controlled diffusion. Both the critic and the state average appearing in the policy-gradient formula must be estimated within the same physical interaction budget.

Derivative estimation from noisy function evaluations provides another useful comparison. Under bounded observation variance and a cubic Taylor-remainder bound, Lam et al. (2020) establish the $n^{-2/3}$ minimax mean-square order, including a lower bound for arbitrary nonlinear estimators at fixed query locations. For optimization, Yu et al. (2024) establish matching upper and lower bounds of order $n^{-2/3}$ for simple regret under adaptive noisy function evaluations, assuming strong convexity and a Lipschitz Hessian. Here n counts function evaluations. The exponent therefore has a

precedent in numerical differentiation; However, establishing it for our RL problem requires constructing an estimator from diffusion trajectories and proving a lower bound that remains valid when actions, resets, and stopping decisions depend on previous observations.

Our upper bound makes the connection to an implementable learning procedure through AdvG's paired critic equations. Uniform critic stability controls the error of an empirical affine solve, while Gaussian integration by parts and transition estimates control the bias introduced by exploration and time discretization. An independent batch of target-policy trajectory prefixes estimates the discounted state average needed for the policy gradient. Charging all critic-fitting trajectories and state-sampling prefixes to the budget gives a bound in total expected physical time T. With exact value and local action-derivative features and exact empirical solves, the mean-squared error is bounded by a constant times

$$\frac{1}{T\sigma_e^2} + \sigma_e^4 + h^2.$$

The terms represent sampling variance, squared exploration bias, and squared discretization bias. Choosing $\sigma_e \asymp T^{-1/6}$ and $h \lesssim T^{-1/3}$ yields the $O(T^{-2/3})$ upper bound in Theorem 5.1 for every admissible fixed exact-feature class. This construction shows how the statistical accuracy can be attained while learning the quantities required by the policy-gradient formula.

The matching lower bound requires a diffusion family that preserves this exact-feature setting without revealing the unknown gradient through the supplied features. We construct such a family using an unknown smooth nonlinear action response in the drift, while keeping the value features, derivative features, reward function, and diffusion coefficient common across models. Exact representation of the value and local advantage derivative therefore remains compatible with an infinite-dimensional unknown action response. Within this family, smooth localized drift perturbations separate the target gradients while keeping the resulting trajectory distributions difficult to distinguish. A stopped-trajectory Girsanov argument bounds their relative entropy by expected total acquisition time, including multiple episodes with adaptive actions, resets, and stopping. Applying two-point testing then gives a $T^{-2/3}$ lower bound even when the learner observes continuous state paths. This establishes sharpness on a selected fixed diffusion class and shows that the lower bound is not an artifact of a particular observation mesh. As a remark, the upper-bound scope remains broader: the result does not assert that every admissible exact-feature class has the same minimax difficulty.

Our contribution connects the design of continuous-time RL algorithms with the fundamental limits of learning from trajectories. AdvG's local pairing principle supports both stable joint actor–critic learning and separate minimax-optimal procedures for gradient estimation and, under gradient dominance, policy optimization on the specified diffusion classes. Our paper features the union of a concrete algorithmic choice, provable convergence, and attainable statistical accuracy. By accounting for physical interaction throughout, our results provide a common basis for understanding what a learning method achieves and what the available observations permit, establishing a benchmark for the design and evaluation of future continuous-time RL methods.