TIME SCALES IN DIFFUSION POLICY LEARNING: ACTION-GRADIENT ATTENUATION AND POISSON-BASED POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous-time diffusion reinforcement learning involves two distinct time scales: the evolution of the environment and the dynamics used to sample actions. We show that a mismatch between these time scales changes the critic signal used to update the actor. Under stationary action resampling, the critic action-gradient signal vanishes linearly with the physical time step h, leading to increasingly weak actor updates as h tends to 0. We derive a finite-horizon corrector that connects this finite-step effect to the stationary-resampling and fast-action limits, and characterize the continuous-time policy-improvement direction through a performance-difference identity based on the Poisson equation. Normalizing the finite-step actor signal by h restores its magnitude, but recovers the limiting policy-improvement direction only when the relevant action modes share a common relaxation rate. An exactly solvable linear–quadratic (LQ) model further shows that a fixed scalar normalization can improve the first policy update yet cause the policy updates to diverge when applied repeatedly. Experiments with both exact and learned critics validate the predicted linear attenuation and show that normalization by h improves learning at small physical time steps, while performance remains sensitive to the choice of scalar coefficient.

1 INTRODUCTION

Reinforcement learning has achieved substantial success across diverse domains (Sutton et al., 1998; Tang et al., 2025). Many control problems, including robotic manipulation (Levine et al., 2016), autonomous driving (Chen et al., 2019; Kiran et al., 2022), and financial trading (Wu & Li, 2024; Wang & Zhou, 2020), evolve continuously in physical time. Diffusion models (Hyvärinen & Dayan, 2005; Ho et al., 2020) have emerged as expressive policy classes that can represent complex, multimodal action distributions. Whereas a standard feedforward actor can produce an action with a single policy-network evaluation (Haarnoja et al., 2018), a diffusion actor constructs an action through iterative sampling. In continuous-time control, this introduces two time scales: the environment evolves over a physical decision interval h, while the action sampler evolves over a sampling horizon τ. Their interaction shapes both the executed transition and the critic signal available for policy improvement.

Most reinforcement learning algorithms are formulated on discrete-time Markov decision processes (MDPs), often with a fixed decision interval (Fujimoto et al., 2018; Schulman et al., 2015). As this interval becomes small, the action-dependent component of a conventional one-step Q-function shrinks relative to its state-value component, making differences between actions harder to learn (Tallec et al., 2019). Continuous-time methods can retain control-relevant information through infinitesimal quantities, including Hamilton–Jacobi characterizations and instantaneous advantage rates such as the little-q function (Kim et al., 2021; Jia & Zhou, 2023). For diffusion policies, the transition underlying critic evaluation also depends on how long the action sampler evolves. Evolving the sampler for a finite horizon τ generally induces a different action transition from drawing directly from its invariant distribution. Moreover, a large ratio τ/h does not guarantee substantial action mixing: the ratio may diverge even when τ tends to zero, while a positive but finite τ need not suffice for equilibration. The physical decision interval and the action-sampling horizon must therefore be considered separately.

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.

This distinction is especially important for continuous-time score-based policy learning, where the actor is guided by the action gradient of the critic. Q-Score Matching (QSM) links the policy score to this gradient (Psenka et al., 2024) and its continuous-time extension, Continuous Q-Score Matching (CQSM), learns the critic through a martingale characterization (Hua et al., 2025). More generally, any dependence of the critic gradient on the physical decision interval can change the effective policy update. Yet it remains unclear how the physical and action-sampling time scales jointly shape the critic signal, what policy direction improves value in the continuous-time limit, and whether a simple scalar normalization can recover that direction.

Our analysis addresses three questions: how the critic action gradient scales with the physical decision interval, which policy field improves value in the small-step limit, and when a scalar normalization can reproduce that field. We show that, under stationary action resampling, the critic action gradient decreases linearly with the physical decision interval. We then characterize the corresponding continuous-time policy-improvement direction through an action Poisson equation. Normalization by the physical decision interval restores the scale of the critic signal, but does not necessarily recover the limiting policy-improvement direction. Exact scalar calibration is possible only when the relevant action modes share a common relaxation rate. Our main contributions are as follows:
(1) **Action-gradient attenuation and action mixing.** We separate the physical decision interval from the action-sampling horizon and derive a finite-horizon corrector that connects finite action mixing with the stationary-resampling and fast-action limits. Under stationary action resampling, we show that the action-dependent critic signal decreases linearly with the physical decision interval, revealing an intrinsic small-step attenuation even when the quadratic control penalty is kept fixed. This result explains why actor updates based directly on the critic action gradient become increasingly weak as the physical decision interval approaches zero.
(2) **Policy improvement and scalar calibration.** We derive an exact performance-difference identity and characterize a Poisson policy-improvement field that guarantees value improvement under exact updates. Although normalization by the physical decision interval restores the magnitude of the vanishing critic gradient, a scalar normalization recovers the Poisson field exactly only when the nonzero advantage rate is supported on action-relaxation modes sharing the same state-independent relaxation rate. Thus, correcting the small-step magnitude alone does not, in general, recover the limiting policy-improvement direction.
(3) **Analytical and empirical validation.** In an exactly solvable linear–quadratic model, we show that the advantage generally contains multiple action modes with different relaxation rates. We construct a case in which a fixed scalar calibration gives an exact and strictly improving first Poisson update, but causes the policy iterates to diverge when the same calibration is reused across iterations. Experiments with both exact and learned critics confirm the predicted linear attenuation of the critic signal and show that normalization by the physical decision interval improves learning at small decision intervals, while performance remains sensitive to the choice of scalar coefficient.

In Section 2, we review related work. Section 3 introduces the control model and analyzes policy evaluation across physical and action-sampling time scales. Section 4 develops Poisson policy improvement and scalar calibration criteria. Section 5 provides an exact LQ analysis, and Section 6 presents numerical experiments. Finally, Section 7 summarizes the main findings. Proofs and additional derivations are provided in Appendices A, B, C, and D. Appendix E provides additional experimental details and additional results.

2 Related work

In this section we review works closely related to ours: continuous-time stochastic control and multiscale analysis, continuous-time reinforcement learning, and diffusion and score-based policy learning.

Continuous-time stochastic control and multiscale analysis. Continuous-time stochastic control provides the classical dynamic-programming and Hamilton–Jacobi–Bellman framework for controlled diffusions (Fleming & Soner, 2006). For systems with separated time scales, averaging and homogenization describe effective slow dynamics by averaging over fast variables, with Poisson equations providing the associated correctors (Pavliotis & Stuart, 2008; Borkar & Gaitsgory, 2007; Wu & Yin, 2022). These works mainly characterize effective dynamics in asymptotic regimes with separated time scales. In contrast, we explicitly keep the physical decision interval finite and study

how its interaction with the action-sampling time scale affects the critic signal and the resulting policy-improvement direction.

Continuous-time reinforcement learning. Continuous-time reinforcement learning has long been studied from a control-theoretic perspective. Early work often focused on deterministic systems or relied on structural assumptions that differ from modern model-free reinforcement learning (Doya, 2000; Munos, 2006). More recent work has developed stochastic, model-free, and actor–critic formulations for continuous-time settings (Zhao et al., 2025). Wang et al. (2020) formulate exploration in continuous time through entropy-regularized relaxed control. Other developments include mean-field analyses of policy-gradient optimization for entropy-regularized MDPs with continuous state and action spaces (Leahy et al., 2022), jump-diffusion extensions (Guo et al., 2023), and continuous-time counterparts of trust-region and proximal policy optimization methods (Zhao et al., 2023). Jia & Zhou (2022a) characterize policy evaluation through martingale conditions, while Jia & Zhou (2022b) develop continuous-time policy-gradient and actor–critic methods. Jia & Zhou (2023) further introduce the little-q function as an instantaneous advantage rate related to the Hamiltonian, thereby avoiding the collapse of the conventional discrete-time Q-function in the continuous-time limit. Our focus is complementary: rather than taking the infinitesimal formulation itself as the learning object, we analyze the finite-step critic signal used by the actor when physical evolution and action sampling operate on different time scales.

Diffusion and score-based policy learning. Diffusion models provide expressive policy classes that can represent complex and multimodal action distributions (Chi et al., 2025). Diffusion-QL incorporates value maximization into diffusion-policy training for offline reinforcement learning (Wang et al., 2023), while diffusion actor–critic methods extend diffusion policies to online reinforcement learning (Wang et al., 2024). Closer to our setting, Q-Score Matching (QSM) links the policy score to the action gradient of a learned Q-function (Psenka et al., 2024), and Hua et al. (2025) extend this principle to continuous-time state–action diffusions. In its experiments, QSM scales the critic-gradient target by a task-dependent coefficient, increasing it for harder tasks (Psenka et al., 2024). Our analysis examines how this target depends on the physical decision interval: under stationary action resampling, the finite-step critic gradient scales linearly with the physical decision interval, and correcting this scale alone does not necessarily recover the Poisson policy-improvement direction or ensure stable policy updates when the same scalar calibration is reused across iterations.

3 Time scales and policy evaluation

In this section, we specify the control model and transition operators, derive the small-step scaling of the stationary-resampling critic action gradient, and connect the stationary-resampling and fast-action limits through a finite-horizon corrector.

3.1 Formulation and Preliminaries

Notation. The norm $\|\cdot\|_\infty = \|\cdot\|_{C^0}$ is the supremum norm; pointwise vector norms are Euclidean. The norms C^j_a and C^j_x include derivatives through order j in the indicated variable, uniformly in both variables; $C^j(\mathsf{X})$ is the state-only norm and C^j on a joint field includes all derivatives through order j. The $L^2(\nu)$ norm uses the stated probability measure ν.

The state and action spaces are flat tori, $\mathsf{X} = \mathbb{T}^{d_x}$ and $\mathsf{A} = \mathbb{T}^{d_a}$.[1] The Brownian motions B^X_t and B^a_t are independent. The process and its generators are

$$\mathrm{d}X_t = b_X(X_t, A_t)\,\mathrm{d}t + \sigma_X(X_t, A_t)\,\mathrm{d}B^X_t, \qquad \mathrm{d}A_t = \Psi(X_t, A_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B^A_t, \tag{1}$$

$$\mathcal{L}^a_X f = b_X \cdot \nabla_x f + \tfrac{1}{2}\,\mathrm{Tr}(\sigma_X \sigma_X^\top \nabla^2_x f), \qquad \mathcal{A}^\Psi f = \Psi \cdot \nabla_a f + \Delta_a f. \tag{2}$$

For action-sampling speed $c > 0$ relative to physical time, define $\mathcal{G}^\Psi_c = \mathcal{L}^a_X + c\mathcal{A}^\Psi$ and

$$\ell^\Psi = r - \frac{\lambda_0}{2}\,\|\Psi\|^2, \qquad Q^\Psi_c = \mathbb{E}^{c,\Psi}\int_0^\infty e^{-\beta t}\ell^\Psi(X_t, A_t)\,\mathrm{d}t, \qquad (\beta - \mathcal{G}^\Psi_c)Q^\Psi_c = \ell^\Psi. \tag{3}$$

[1] The compact torus setting is used only for the general operator results, as it avoids boundary conditions and provides uniform invariant-measure and regularity estimates.

The penalty is charged per unit physical time. The case $c = 1$ recovers the synchronized state–action dynamics underlying CQSM (Hua et al., 2025), with the corresponding critic denoted by Q_1^Ψ.

Frozen-action averaging. At a fixed state, let S_τ^Ψ denote the action semigroup and $\pi^\Psi(\mathrm{d}a \mid x)$ its invariant law. Set $\Pi^\Psi f(x) = \int f(x,a)\pi^\Psi(\mathrm{d}a \mid x)$. We identify state-only functions with their extensions constant in a. In particular, $\Pi^\Psi S_\tau^\Psi = S_\tau^\Psi \Pi^\Psi = \Pi^\Psi$ and $\Pi^\Psi \mathcal{A}^\Psi = 0$. Moreover, every smooth centered function admits a unique centered Poisson solution. These standard ergodic and elliptic facts are stated precisely in Lemma A.3; see also Mattingly et al. (2010) for Poisson-equation regularity for diffusions on the torus.

Transition operators and their limits. For physical decision interval $h > 0$ and action-sampling horizon $\tau > 0$, define the transition operator by $\mathcal{K}_{h,\tau}^\Psi := P_h S_\tau^\Psi$, where P_h denotes the physical-state transition over horizon h with the action held fixed, and S_τ^Ψ denotes the frozen-state action semigroup over horizon τ. The corresponding stationary-resampling operator is $\mathcal{K}_{h,\mathrm{st}}^\Psi := P_h \Pi^\Psi$, where Π^Ψ projects onto the invariant action law $\pi^\Psi(\cdot \mid x)$.

We denote by $Q_{h,\tau}^\Psi$ and Q_h^Ψ the corresponding discounted critics, defined as the solutions of

$$Q_{h,\tau}^\Psi = h\ell^\Psi + e^{-\beta h}\mathcal{K}_{h,\tau}^\Psi Q_{h,\tau}^\Psi, \tag{4}$$

$$Q_h^\Psi = h\ell^\Psi + e^{-\beta h}\mathcal{K}_{h,\mathrm{st}}^\Psi Q_h^\Psi. \tag{5}$$

For an h-dependent action horizon $\tau_h > 0$, we write $c_h := \frac{\tau_h}{h}$ for the ratio of action-sampling time to physical time.

Proposition 3.1. *Under Assumption A.1, for any smooth f, as $h \to 0$,*

$$\frac{\mathcal{K}_{h,\tau_h}^\Psi f - f}{h} \to (\mathcal{L}_X^a + c\mathcal{A}^\Psi)f, \qquad \tau_h \to 0,\ c_h \to c < \infty, \tag{6}$$

$$\frac{\mathcal{K}_{h,\tau_h}^\Psi f - f}{\tau_h} \to \mathcal{A}^\Psi f, \qquad \tau_h \to 0,\ c_h \to \infty, \tag{7}$$

$$\mathcal{K}_{h,\tau_h}^\Psi f \to S_{\tau_0}^\Psi f, \quad \mathcal{K}_{h,\mathrm{st}}^\Psi f \to \Pi^\Psi f, \qquad \tau_h \to \tau_0 \in (0,\infty), \tag{8}$$

with all convergences in $\|\cdot\|_\infty$. Moreover, if $c_h \to c \in (0,\infty)$, then $Q_{h,\tau_h}^\Psi \to Q_c^\Psi$ also in $\|\cdot\|_\infty$.

The time-scale ratio $c_h = \tau_h/h$ measures the action-sampling horizon relative to the physical step, but does not determine the absolute amount of action mixing within one decision. In particular, $c_h \to \infty$ may occur either with $\tau_h \to 0$ or with $\tau_h \to \tau_0 > 0$, which correspond to different limiting action dynamics.

3.2 Stationary critic scaling

For a fixed policy field Ψ, define the averaged state generator and averaged running reward by $\bar{\mathcal{L}}^\Psi v := \Pi^\Psi \mathcal{L}_X^a v, \bar{\ell}^\Psi := \Pi^\Psi \ell^\Psi$. The corresponding averaged value V^Ψ is defined as the solution of

$$(\beta - \bar{\mathcal{L}}^\Psi)V^\Psi = \bar{\ell}^\Psi. \tag{9}$$

Associated with V^Ψ, define the centered advantage rate $q^\Psi := \ell^\Psi + \mathcal{L}_X^a V^\Psi - \beta V^\Psi$. Applying Π^Ψ and using Eq. (9) gives $\Pi^\Psi q^\Psi = 0$. Hence, by Lemma A.3, there exists a unique centered Poisson corrector U^Ψ satisfying

$$-\mathcal{A}^\Psi U^\Psi = q^\Psi, \qquad \Pi^\Psi U^\Psi = 0. \tag{10}$$

For the stationary-resampling critic Q_h^Ψ, define $V_h^\Psi := \Pi^\Psi Q_h^\Psi, T_h^\Psi := \Pi^\Psi P_h$. Then the discrete Bellman equation reduces exactly to

$$V_h^\Psi = h\bar{\ell}^\Psi + e^{-\beta h}T_h^\Psi V_h^\Psi, \tag{11}$$

while the full critic is reconstructed from

$$Q_h^\Psi = h\ell^\Psi + e^{-\beta h}P_h V_h^\Psi. \tag{12}$$

For the stationary-resampling critic, we quantify the magnitude of its action-dependent signal by $G_Q(h;\nu) := \left\|\nabla_a Q_h^\Psi\right\|_{L^2(\nu)}$, where ν is a fixed probability measure on $\mathsf{X} \times \mathsf{A}$. This quantity measures the $L^2(\nu)$ magnitude of the critic action gradient and will be used below to characterize its dependence on the physical step size h.

Theorem 3.2 (Stationary critic signal)**.** *Suppose Assumption A.1 holds and fix Ψ. Then, as $h \downarrow 0$, $\|V_h^\Psi - V^\Psi\|_\infty + \|Q_h^\Psi - V^\Psi\|_\infty = O(h)$. If, in addition, Assumption A.2 holds, then*

$$\left\| \frac{Q_h^\Psi - V_h^\Psi}{h} - q^\Psi \right\|_{C_a^1} \longrightarrow 0. \tag{13}$$

Consequently, for every fixed probability measure ν,

$$G_Q(h;\nu) = h \left\| \nabla_a q^\Psi \right\|_{L^2(\nu)} + o(h). \tag{14}$$

Remark 3.3. The centering by $V_h^\Psi = \Pi^\Psi Q_h^\Psi$ is essential. Indeed, $\Pi^\Psi(Q_h^\Psi - V_h^\Psi) = 0$, so $Q_h^\Psi - V_h^\Psi$ isolates the action-dependent component of the critic. Although $V_h^\Psi - V^\Psi = O(h)$, this state-only discrepancy may contribute a nonvanishing $O(1)$ term after division by h. Consequently, the expansion in Eq. (13) need not remain valid with V^Ψ in place of V_h^Ψ. If $\|\nabla_a q^\Psi\|_{L^2(\nu)} > 0$, Eq. (14) shows that the stationary critic action-gradient signal decays linearly with the physical step size h.

For a fixed action horizon $\tau > 0$, define the centered corrector by $(I - S_\tau^\Psi)\chi_\tau^\Psi = q^\Psi, \Pi^\Psi \chi_\tau^\Psi = 0$. Theorem A.5 gives $Q_{h,\tau}^\Psi - \Pi^\Psi Q_{h,\tau}^\Psi = h\chi_\tau^\Psi + O(h^2)$ in the uniform norm for fixed τ. Under its additional action-curvature bound, $h^{-1}\nabla_a Q_{h,\tau}^\Psi \to \nabla_a \chi_\tau^\Psi$. Moreover, $\chi_\tau^\Psi \to q^\Psi$ as $\tau \to \infty$, whereas $\tau\chi_\tau^\Psi \to U^\Psi$ as $\tau \downarrow 0$.

4 Policy improvement and calibration

The stationary critic identifies the leading action signal $\nabla_a q^\Psi$. Policy improvement, however, depends on the cumulative relaxation of this signal under the frozen-action dynamics. This motivates the Poisson improvement field

$$B^\Psi := \frac{1}{\lambda_0}\nabla_a U^\Psi = \frac{1}{\lambda_0}\nabla_a(-\mathcal{A}^\Psi)^{-1} q^\Psi, \tag{15}$$

where the inverse of $-\mathcal{A}^\Psi$ is understood on the centered subspace. The same construction gives an exact improvement rule at any fixed physical step $h > 0$. Define

$$q_h^\Psi := \frac{Q_h^\Psi - V_h^\Psi}{h}, \qquad -\mathcal{A}^\Psi U_h^\Psi = q_h^\Psi, \quad \Pi^\Psi U_h^\Psi = 0, \qquad B_h^\Psi := \frac{1}{\lambda_0}\nabla_a U_h^\Psi. \tag{16}$$

Since $\Pi^\Psi q_h^\Psi = 0$, the centered Poisson solution exists by Lemma A.3. For a unified statement, let $h = 0$ denote the averaged model and set $(V_0^\Psi, q_0^\Psi, U_0^\Psi, B_0^\Psi) = (V^\Psi, q^\Psi, U^\Psi, B^\Psi)$. For a candidate policy Φ, define the discounted state resolvent

$$\mathcal{R}_h^\Phi := \begin{cases} h(I - e^{-\beta h} T_h^\Phi)^{-1}, & h > 0, \\ (\beta - \bar{\mathcal{L}}^\Phi)^{-1}, & h = 0, \end{cases} \qquad T_h^\Phi = \Pi^\Phi P_h. \tag{17}$$

Theorem 4.1 (Policy improvement under stationary action sampling)**.** *Under Assumption A.1, fix $h \geq 0$ and a smooth policy field Ψ. For any smooth candidate field Φ, set $D_{h,\Psi}(\Phi) := \Pi^\Phi \left[q_h^\Psi - \frac{\lambda_0}{2}\left(\|\Phi\|^2 - \|\Psi\|^2\right)\right]$. Then*

$$D_{h,\Psi}(\Phi) = \frac{\lambda_0}{2}\Pi^\Phi \left[\left\|\Psi - B_h^\Psi\right\|^2 - \left\|\Phi - B_h^\Psi\right\|^2 \right], \tag{18}$$

$$V_h^\Phi - V_h^\Psi = \mathcal{R}_h^\Phi D_{h,\Psi}(\Phi). \tag{19}$$

Thus $D_{h,\Psi}(\Phi) \geq 0$ implies $V_h^\Phi \geq V_h^\Psi$. In particular, for $0 < \eta \leq 1$, the policy $\Phi_\eta = (1-\eta)\Psi + \eta B_h^\Psi$ satisfies

$$V_h^{\Phi_\eta} - V_h^\Psi = \frac{\lambda_0}{2}\eta(2-\eta)\mathcal{R}_h^{\Phi_\eta}\Pi^{\Phi_\eta}\left\|\Psi - B_h^\Psi\right\|^2 \geq 0. \tag{20}$$

The inequality is strict at every state whenever $\Psi \not\equiv B_h^\Psi$.

The projection uses the invariant action law of the candidate policy, so the identity accounts for the change in action occupancy. At $h = 0$, it recovers the stationary-averaged performance difference,

with the shorthand $D_\Psi(\Phi) := D_{0,\Psi}(\Phi) = \Pi^\Phi[\ell^\Phi + \mathcal{L}_X^a V^\Psi - \beta V^\Psi]$. As an immediate consequence of Theorem 4.1, the repeated updates $\Psi_{k+1} = (1-\eta_k)\Psi_k + \eta_k B_h^{\Psi_k}$, with $\eta_k \in (0,1]$, generate pointwise nondecreasing values $V_h^{\Psi_k}$, with strict improvement unless $\Psi_k = B_h^{\Psi_k}$. Since the discounted value is uniformly bounded above, the values converge pointwise to finite limits. Furthermore, any fixed policy $\Psi = B_h^\Psi$ is globally optimal among smooth policy fields by (19).

We next compare these Poisson updates with scalar normalization of the critic action gradient. For stationary resampling, consider the actor target $\mathcal{T}_h^\Psi := \frac{\nabla_a Q_h^\Psi}{\lambda_A(h)}, \lambda_A(h) > 0$. For a fixed finite action horizon $\tau > 0$, the corresponding target is obtained by replacing Q_h^Ψ with $Q_{h,\tau}^\Psi$. Under the critical scaling $\lambda_A(h) = \alpha h$, the limiting stationary target is $\nabla_a q^\Psi/\alpha$, whereas Theorem 4.1 identifies $B^\Psi = \frac{1}{\lambda_0}\nabla_a U^\Psi$ as a sufficient improvement benchmark. We therefore define the scalar calibration mismatch by

$$e_\alpha^\Psi := \left\| \frac{\nabla_a q^\Psi}{\alpha} - B^\Psi \right\|_\infty . \tag{21}$$

Proposition 4.2 (Exact scalar calibration)**.** *Under Assumption A.1, let $\alpha > 0$ and $q^\Psi \not\equiv 0$. Then $e_\alpha^\Psi = 0$ if and only if $-\mathcal{A}^\Psi q^\Psi = \frac{\alpha}{\lambda_0} q^\Psi$. Equivalently, exact scalar calibration is possible only when q^Ψ lies in a single active eigenspace of $-\mathcal{A}^\Psi$ with a common relaxation rate $\omega = \frac{\alpha}{\lambda_0}$. In that case, the stationary calibration is $\lambda_A(h) = \lambda_0 \omega h$. For fixed $\tau > 0$, the corresponding finite-horizon calibration is $\lambda_A(h) = \frac{\lambda_0 \omega h}{1 - e^{-\omega\tau}}$.*

Thus scalar normalization reproduces the Poisson improvement field exactly only in a single common action-relaxation mode. Multiple active relaxation rates generally prevent exact scalar matching, although they do not preclude the sufficient improvement condition in Theorem 4.1.

Corollary 4.3 (Approximate calibration and value-loss bound)**.** *Suppose Assumptions A.1 and A.2 hold, and fix a policy field Ψ and $\alpha > 0$. Let $\widehat{Q}_h$ be such that $\Phi_h := \frac{\nabla_a \widehat{Q}_h}{\alpha h}$ is a smooth admissible candidate field and $\epsilon_h := \left\| \frac{\nabla_a Q_h^\Psi}{h} - \nabla_a q^\Psi \right\|_\infty, \eta_h := \left\| \nabla_a (\widehat{Q}_h - Q_h^\Psi) \right\|_\infty$. Then $\epsilon_h \to 0$ as $h \downarrow 0$, and*

$$V^{\Phi_h} - V^\Psi \geq -\frac{\lambda_0}{2\beta} \left(e_\alpha^\Psi + \frac{\epsilon_h}{\alpha} + \frac{\eta_h}{\alpha h} \right)^2 \tag{22}$$

pointwise on X. In particular, if $e_\alpha^\Psi = 0, \eta_h = o(h)$, then

$$\liminf_{h \downarrow 0} \inf_{x \in \mathsf{X}} \left(V^{\Phi_h}(x) - V^\Psi(x) \right) \geq 0. \tag{23}$$

The bound separates three sources of error: scalar calibration mismatch e_α^Ψ, finite-step critic bias ϵ_h, and critic-gradient estimation error η_h. The critical $1/h$ normalization restores the scale of the vanishing critic signal, but amplifies the estimation error to η_h/h. Consequently, consistency of the normalized actor target requires the critic-gradient error to vanish faster than h.

5 LQ Analysis of Scalar Calibration

The LQ model makes this distinction explicit within a single affine policy family. Its action relaxation modes show that scalar calibration is generally inexact, while an invariant single-mode slice yields a sharper counterexample: a scalar can exactly match the Poisson improvement field and strictly improve the value at one policy, yet reusing the same scalar in subsequent updates can still drive the population iteration to divergence.

5.1 Action Relaxation Modes

We consider a scalar LQ model with state $X_t \in \mathbb{R}$ and action $a_t \in \mathbb{R}$. The physical state evolves as

$$dX_t = -X_t\, dt + a_t\, dB_t^X, \tag{24}$$

with running reward $r(x,a) = -(x^2 + xa + a^2 + x + 2a), \beta = 1, \lambda_0 = 0.1$. The action follows the affine Ornstein–Uhlenbeck dynamics

$$da_t = \Psi_\theta(X_t, a_t)\, dt + \sqrt{2}\, dB_t^a, \qquad \Psi_\theta(x,a) = -\kappa(a - px - d), \qquad \theta = (\kappa, p, d), \tag{25}$$

where $\kappa > 0$ controls the action relaxation rate and $px + d$ is the conditional stationary mean. We use the superscript θ as shorthand for Ψ_θ. The noncompact state and action spaces, polynomially growing coefficients, and degeneracy of $\sigma_X(x, a) = a$ place this model outside Assumption A.1. We establish the required conclusions directly within the affine policy family (see Appendix C). Let $\mathcal{D} = \{\kappa > 0,\ p^2 < 3,\ d \in \mathbb{R}\}$. With $\beta = 1, p^2 < 3$ ensures a finite discounted averaged value.

For the frozen-action Ornstein–Uhlenbeck dynamics, $A \mid X = x \sim N(\mu(x), \kappa^{-1}), \mu(x) = px + d$. Writing $V^\theta(x) = \frac{1}{2}vx^2 + wx + z$, coefficient matching in the averaged Bellman equation gives

$$v = -\frac{2(1 + p + p^2)}{3 - p^2}, \qquad R(p) := 2 - v = \frac{2(4 + p)}{3 - p^2}, \qquad H := R(p) + \lambda_0\kappa^2. \tag{26}$$

Set $y := a - \mu(x), b(x) := -R(p)\mu(x) - x - 2$. Substituting the quadratic value into $q^\theta = \ell^\theta + \mathcal{L}_X^a V^\theta - V^\theta$ and using $\Pi^\theta y = 0, \Pi^\theta y^2 = \kappa^{-1}$, gives

$$q^\theta = -\frac{H}{2}\left(y^2 - \kappa^{-1}\right) + b(x)y. \tag{27}$$

Moreover, $-\mathcal{A}^\theta y = \kappa y, -\mathcal{A}^\theta\left(y^2 - \kappa^{-1}\right) = 2\kappa\left(y^2 - \kappa^{-1}\right)$. Hence the centered Poisson equation $-\mathcal{A}^\theta U^\theta = q^\theta$ has the explicit solution

$$U^\theta = -\frac{H}{4\kappa}\left(y^2 - \kappa^{-1}\right) + \frac{b(x)}{\kappa}y. \tag{28}$$

The two centered action modes $y^2 - \kappa^{-1}$ and $b(x)y$ relax under the frozen-action Ornstein–Uhlenbeck dynamics at rates 2κ and κ, respectively. Hence the LQ model is generically multi-mode even though the action is one-dimensional. If $b \not\equiv 0$, no single scalar α can reproduce the full Poisson improvement field.

For the stationary-resampling model, the limiting actor target as $h \downarrow 0$ is $F_0(\Psi_\theta) = \frac{1}{\alpha}\partial_a q^\theta$. Since $\partial_a q^\theta = -Ha + (\lambda_0\kappa^2 p - 1)x + (\lambda_0\kappa^2 d - 2)$, the target remains in the affine policy family. Writing $F_0(\Psi_\theta) = -\kappa'(a - p'x - d')$ and matching the coefficients of a, x, and the constant term gives $\kappa' = \frac{H}{\alpha}, p' = \frac{\lambda_0\kappa^2 p - 1}{H}, d' = \frac{\lambda_0\kappa^2 d - 2}{H}$. Hence the limiting actor map restricted to the affine family is

$$\mathcal{F}_\alpha(\theta) = \left(\frac{H}{\alpha}, \frac{\lambda_0\kappa^2 p - 1}{H}, \frac{\lambda_0\kappa^2 d - 2}{H}\right). \tag{29}$$

Under full target replacement, the exact population iteration therefore reduces to the finite-dimensional recursion $\theta_{k+1} = \mathcal{F}_\alpha(\theta_k)$ with a fixed scalar α. The map preserves $\mathcal{D}$ and is the finite-dimensional representation of the limiting actor map F_0 on the affine policy family. Its preservation of $\mathcal{D}$ is verified in Appendix C.2.

5.2 An improving update can diverge under repeated reuse

The population map (29) also shows why exact one-step calibration does not imply that the same scalar can be reused throughout learning. Define $p_* = -4 + \sqrt{13}, d_* = 2p_*, R_* := R(p_*) = \frac{4 + \sqrt{13}}{3}$. These constants satisfy

$$R_* p_* = -1, \qquad R_* d_* = -2. \tag{30}$$

Hence the one-dimensional family $\mathcal{M} := \{(\kappa, p_*, d_*) : \kappa > 0\}$ is invariant under the population map $\mathcal{F}_\alpha$ for every $\alpha > 0$. Indeed, for $\theta \in \mathcal{M}$, $p' = \frac{\lambda_0\kappa^2 p_* - 1}{R_* + \lambda_0\kappa^2} = p_*, d' = \frac{\lambda_0\kappa^2 d_* - 2}{R_* + \lambda_0\kappa^2} = d_*$. Moreover, (30) gives $b \equiv 0$ on $\mathcal{M}$. Thus the linear action mode disappears and (27) reduces to

$$q^\theta = -\frac{R_* + \lambda_0\kappa^2}{2}\left(y^2 - \kappa^{-1}\right), \qquad \theta \in \mathcal{M}. \tag{31}$$

The only active action mode therefore relaxes at rate 2κ. By Proposition 4.2, the exact scalar calibration at the current policy is $\alpha = 2\lambda_0\kappa$. Under a fixed scalar α, however, the population iteration on $\mathcal{M}$ becomes

$$\kappa_{k+1} = \frac{R_* + \lambda_0\kappa_k^2}{\alpha}. \tag{32}$$

The next result shows that choosing α to be exactly correct at one policy does not make it a valid calibration for subsequent policies.

Proposition 5.1 (One-step improvement can precede divergence). *Let* $0 < \kappa_0 < \sqrt{\frac{R_*}{\lambda_0}}, \theta_0 = (\kappa_0, p_*, d_*), \alpha = 2\lambda_0\kappa_0$, *and iterate* $\theta_{k+1} = \mathcal{F}_\alpha(\theta_k)$ *with this fixed scalar. Then the first update coincides with the Poisson improvement field* B^{θ_0} *and strictly improves the stationary-averaged value:*

$$V^{\theta_1}(x) - V^{\theta_0}(x) = \frac{\left(R_* - \lambda_0\kappa_0^2\right)^2}{4\kappa_0\left(R_* + \lambda_0\kappa_0^2\right)} > 0. \tag{33}$$

Nevertheless, $\kappa_k \to \infty, V^{\theta_k}(x) \to -\infty$ *for every fixed* x.

Proposition 5.1 separates two distinct roles of scalar normalization. The choice $\alpha = 2\lambda_0\kappa_0$ exactly matches the Poisson geometry at the initial policy, so its first update is genuinely improving. After the policy changes, however, the active relaxation rate changes with κ, whereas the reused scalar remains fixed. Thus even exact calibration at one iterate need not remain calibrated along the learning trajectory.

6 Experiments

We evaluate two central predictions of the theory: the $O(h)$ attenuation of the critic action gradient and the different roles of step-size scaling and scalar calibration in actor updates. All experiments in the main text use the LQ benchmark with stationary action sampling and keep the objective penalty λ_0 fixed throughout.

Critic attenuation under stationary resampling. We first isolate the critic effect by freezing the reference affine policy. For each physical step h, we solve the exact quadratic population critic induced by the one-step stationary-resampling operator used in the implementation. The next action is drawn directly from the invariant conditional distribution rather than evolved for a finite horizon $\tau = h$. The full construction and derivation are given in Appendix D.

Writing $\partial_a Q_h(x,a) = \theta_{2,h}a + \theta_{4,h}x + \theta_{3,h}$, we measure the action-signal magnitude by $S_h = (\theta_{2,h}^2 + \theta_{4,h}^2 + \theta_{3,h}^2)^{1/2}$. For the fixed reference measure $\nu_0 = \mathcal{N}(0,1)_x \otimes \mathcal{N}(0,1)_a$, this coefficient norm satisfies $S_h = \|\partial_a Q_h\|_{L^2(\nu_0)} = G_Q(h;\nu_0)$. Figure 1(a) shows that S_h decays linearly with h, with a fitted log–log slope of 0.9892. Moreover, the error between the rescaled coefficients $h^{-1}(\theta_{2,h}, \theta_{4,h}, \theta_{3,h})$ and the limiting little-q coefficients decays with slope 0.9889 (Figure 1(b)). Both observations agree with the exact expansion $\partial_a Q_h^\Psi = h\,\partial_a q^\Psi + O(h^2)$.

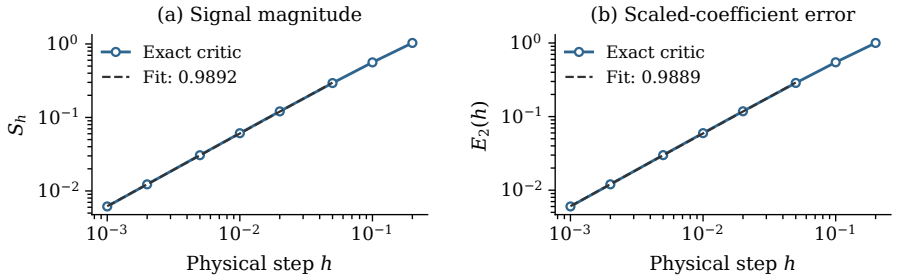

Figure 1: Exact stationary-resampling critic under the fixed reference policy. (a) Action-signal magnitude S_h. (b) Error of the $1/h$-rescaled gradient coefficients relative to the limiting little-q gradient. Dashed lines are log–log fits over $h \le 0.05$.

Actor normalization and calibration. We next restore joint actor–critic learning. The objective coefficient remains fixed at $\lambda_0 = 0.1$, while the actor uses $\Psi_{\mathrm{tar}}(x,a) = \frac{\partial_a Q(x,a)}{\lambda_A(h)}$. To align evaluation with the control objective, we freeze each learned final policy and compute its discounted value under the Euler state transition used during training,

$$V_{\mathrm{E},h}^\Psi(x_0) = \mathbb{E}_{x_0}^\Psi\left[\sum_{n=0}^{\infty} h e^{-nh}\left(r(X_n,A_n) - \frac{\lambda_0}{2}|\Psi(X_n,A_n)|^2\right)\right]. \tag{34}$$

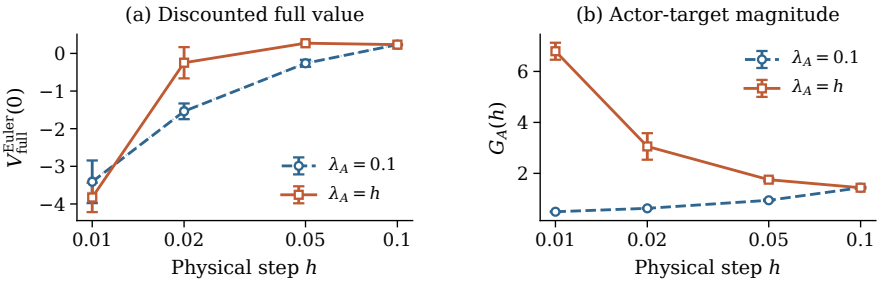

Figure 2: Actor normalization. (a) Discounted full value of the frozen final policy. (b) Actor-target magnitude averaged over the final 10% of training. Error bars show one standard deviation over ten seeds, except for $\lambda_A = 0.1$ at $h = 0.01$, where nine valid runs remain.

The LQ structure gives this value in closed form. Figure 2 separates these two effects. With fixed normalization, the actor-target magnitude decreases as h becomes smaller, consistent with the $O(h)$ attenuation of the critic gradient under stationary action resampling. Scaling $\lambda_A(h)$ with h restores this signal. The resulting full objective improves substantially at $h = 0.05$ and $h = 0.02$, but not at $h = 0.01$, where the stronger actor field incurs a large control cost. Thus, inverse-step normalization removes the leading attenuation of this actor target, but does not by itself ensure improvement of the full discounted objective.

Restoring the h-scale leaves a scalar calibration degree of freedom, so we further set $\lambda_A(h) = \alpha h$. At $h = 0.02$, increasing α from 1 to 1.25 reduces the discounted control cost from 0.817 to 0.332 and raises the full value from -0.248 to 0.215. At $h = 0.01$, the same effect is more pronounced: the full value increases from -3.828 at $\alpha = 1$ to -0.102 at $\alpha = 1.5$, as the control cost falls from 3.911 to 0.585. Larger α eventually suppresses the raw-value gain. These results support the theoretical distinction between signal scaling and policy calibration: h determines the leading-order signal scale, whereas α controls the resulting control-cost/performance tradeoff. Full decompositions and the original training metrics are reported in Appendix E.4.

Additional exact LQ mechanism diagnostics are reported in Appendix E.5. They show that $\tau_h/h \to \infty$ does not uniquely determine the critic-signal scaling, and that scalar normalization exactly recovers the Poisson improvement field only in the single-mode case. MuJoCo experiments in Appendix E.6 further show that performance is sensitive to the schedule of physical decision intervals.

7 Conclusion

In this work, we study how the physical and action-sampling time scales interact in continuous-time diffusion policy learning. We derive a finite-horizon action correction that connects partial action mixing to stationary resampling and the fast-action limit. Under stationary resampling, the critic action gradient decreases linearly with the physical decision interval, causing the actor signal to vanish as the interval approaches zero. An exact performance-difference identity further characterizes the action Poisson field as a value-improving direction. Normalizing the critic-gradient signal by the physical decision interval restores its magnitude, but recovers the Poisson field exactly only when the relevant action modes share a common relaxation rate. An LQ example shows that a fixed scalar calibration can improve the first update yet lead to divergent policy iterates when reused. Experiments support the predicted attenuation and the benefit of step-dependent normalization. These results highlight the distinction between correcting signal magnitude and recovering the policy-improvement direction. Future work includes learning the Poisson correction, controlling critic-gradient error under inverse-step normalization, and establishing convergence with approximate evaluation and policy updates.

AI USE STATEMENT

We used generative AI tools for language editing, polishing writing, and code debugging for numerical experiments. AI-assisted mathematical derivations and code were manually checked against the analytical results and tested for correctness. We reviewed all AI-assisted content and take responsibility for the final text, claims, proofs, and experimental results in this work.

ETHICS STATEMENT

This work is theoretical and uses LQ & Mujoco experiments. It does not involve human subjects, personal data, or sensitive datasets. We are not aware of any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Complete proofs and LQ derivations are provided in the appendices. We also report the exact evaluation formulas, training settings, hyperparameters, random-seed protocol, numerical safeguards, and additional experimental results.

REFERENCES

Vivek S Borkar and Vladimir Gaitsgory. Singular perturbations in ergodic control of diffusions. *SIAM journal on control and optimization*, 46(5):1562–1577, 2007.

Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. Model-free deep reinforcement learning for urban autonomous driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 2765–2771, 2019. doi: 10.1109/ITSC.2019.8917306.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1): 219–245, 2000.

Wendell H Fleming and H Mete Soner. *Controlled Markov processes and viscosity solutions*. Springer, 2006.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

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.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Chengxiu Hua, Jiawen Gu, and Yushun Tang. Continuous q-score matching: Diffusion guided reinforcement learning for continuous-time control. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=o8F5lNOTG6.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

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.

Yanwei Jia and Xunyu 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.

Yanwei Jia and Xunyu 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.

Jeongho Kim, Jaeuk Shin, and Insoon Yang. Hamilton-jacobi deep q-learning for deterministic continuous-time systems with lipschitz continuous controls. *Journal of Machine Learning Research*, 22(206):1–34, 2021.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022. doi: 10.1109/TITS.2021.3054625.

James-Michael Leahy, Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Convergence of policy gradient for entropy regularized mdps with neural network approximation in the mean-field regime. In *International Conference on Machine Learning*, pp. 12222–12252. PMLR, 2022.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(1):1334–1373, January 2016. ISSN 1532-4435.

Jonathan C. Mattingly, Andrew M. Stuart, and Michael V. Tretyakov. Convergence of numerical time-averaging and stationary measures via poisson equations. *SIAM Journal on Numerical Analysis*, 48(2):552–577, 2010. doi: 10.1137/090770527. URL https://doi.org/10.1137/090770527.

Rémi Munos. Policy gradient in continuous time. *Journal of Machine Learning Research*, 7:771–791, 2006.

Grigorios A Pavliotis and Andrew M Stuart. *Multiscale methods: averaging and homogenization*, volume 53. Springer New York, 2008.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. In *International Conference on Machine Learning*, pp. 41163–41182. PMLR, 2024.

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

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1):153–188, 2025.

Haoran Wang and Xunyu Zhou. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308, 2020.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

Yinuo Wang, Likun Wang, Yuxuan Jiang, Wenjun Zou, Tong Liu, Xujie Song, Wenxuan Wang, Liming Xiao, Jiang Wu, Jingliang Duan, et al. Diffusion actor-critic with entropy regulator. *Advances in Neural Information Processing Systems*, 37:54183–54204, 2024.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AHvFDPi-FA.

Bo Wu and Lingfei Li. Reinforcement learning for continuous-time mean-variance portfolio selection in a regime-switching market. *Journal of Economic Dynamics and Control*, 158:104787, 2024.

Fuke Wu and George Yin. Fast-slow-coupled stochastic functional differential equations. *Journal of Differential Equations*, 323:1–37, 2022.

Hanyang Zhao, Wenpin Tang, and David Yao. Policy optimization for continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 36:13637–13663, 2023.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David Yao, and Wenpin Tang. Score as action: Fine tuning diffusion generative models by continuous-time reinforcement learning. In *International Conference on Machine Learning*, pp. 77371–77389. PMLR, 2025.

A PROOFS FOR EVALUATION ACROSS PHYSICAL AND ACTION-SAMPLING TIME SCALES

We use the following standard assumptions throughout.

Assumption A.1 (Compact smooth model)**.** The coefficients r, b_X, σ_X and each policy field Ψ are smooth, $\sigma_X \sigma_X^\top$ is uniformly positive definite, and $\beta, \lambda_0 > 0$.

Assumption A.2 (Uniform regularity of the stationary value)**.** For a fixed policy field Ψ, there exist $h_0 > 0$ and $C < \infty$ such that $\sup_{0<h\le h_0} \left\|V_h^\Psi\right\|_{C^6(\mathsf{X})} \le C$.

A.1 FROZEN-ACTION REGULARITY AND POISSON SOLVABILITY

Lemma A.3 (Projection and Poisson equation)**.** *Under Assumption A.1, fix a smooth policy field Ψ. The unique invariant law of the frozen action process has density*

$$\pi^\Psi(\mathrm{d}a \mid x) = \rho^\Psi(x,a)\ \mathrm{d}a, \qquad \int_{\mathsf{A}} \rho^\Psi(x,a)\ \mathrm{d}a = 1, \tag{35}$$

where $\rho^\Psi \in C^\infty(\mathsf{X}\times\mathsf{A})$ is strictly positive. For every $\tau \ge 0$, $\Pi^\Psi S_\tau^\Psi = S_\tau^\Psi \Pi^\Psi = \Pi^\Psi$ on smooth functions, and $\Pi^\Psi \mathcal{A}^\Psi f = 0$ for every smooth f. Every $g \in C^\infty(\mathsf{X}\times\mathsf{A})$ with $\Pi^\Psi g = 0$ admits a unique $u \in C^\infty(\mathsf{X}\times\mathsf{A})$ satisfying $-\mathcal{A}^\Psi u = g$ and $\Pi^\Psi u = 0$.

Proof of Lemma A.3. Write $L_x = \mathcal{A}^{x,\Psi} = \Psi(x,\cdot)\cdot\nabla_a + \Delta_a$. Its formal adjoint with respect to volume measure $\mathrm{d}a$ is $L_x^* v = \Delta_a v - \nabla_a \cdot \big(\Psi(x,\cdot)v\big)$. For each fixed x, standard elliptic theory for the uniformly elliptic diffusion on the connected torus A gives a unique invariant probability density $\rho_x = \rho^\Psi(x,\cdot)$ satisfying $L_x^*\rho_x = 0$ and $\int_{\mathsf{A}} \rho_x\ \mathrm{d}a = 1$, and the kernels $\ker L_x = \mathrm{span}\{1\}$ and $\ker L_x^* = \mathrm{span}\{\rho_x\}$. Elliptic regularity makes ρ_x smooth in a. The strictly positive transition density of the uniformly elliptic diffusion on the connected torus and invariance imply $\rho_x > 0$.

Invariance gives $\Pi^\Psi S_\tau^\Psi = \Pi^\Psi$, while preservation of constants in a gives $S_\tau^\Psi \Pi^\Psi = \Pi^\Psi$. For smooth f, periodic integration by parts gives

$$\Pi^\Psi \mathcal{A}^\Psi f(x) = \int_{\mathsf{A}} L_x f(x,a)\rho_x(a)\ \mathrm{d}a = \int_{\mathsf{A}} f(x,a) L_x^*\rho_x(a)\ \mathrm{d}a = 0. \tag{36}$$

We verify smooth dependence on x on fixed function spaces. Fix x_0, $k \in \mathbb{N}_0$, and $0 < \gamma < 1$, and put $\rho_0 = \rho_{x_0}$. On periodic Hölder spaces, the Fredholm alternative and Schauder regularity give $C^{k+2,\gamma}$ solutions of $-L_x u = f$ and $-L_x^* v = h$ for $f, h \in C^{k,\gamma}$ precisely when $\int_{\mathsf{A}} f\rho_x\ \mathrm{d}a = 0$ and $\int_{\mathsf{A}} h\ \mathrm{d}a = 0$, respectively. Consider the bounded operators

$$F_x : C^{k+2,\gamma}(\mathsf{A}) \longrightarrow C^{k,\gamma}(\mathsf{A}), F_x v = -L_x^* v + \rho_0 \int_{\mathsf{A}} v\ \mathrm{d}a. \tag{37}$$

The operator F_{x_0} is bijective: since $L_{x_0}1 = 0$, $\int_{\mathsf{A}} L_{x_0}^* v\ \mathrm{d}a = 0$. Integrating $F_{x_0}v = f$ and using $\int_{\mathsf{A}} \rho_0\ \mathrm{d}a = 1$ therefore forces $\int v\ \mathrm{d}a = \int f\ \mathrm{d}a$, and the remaining equation is $-L_{x_0}^* v = f - \rho_0 \int_{\mathsf{A}} f\ \mathrm{d}a$. Its right-hand side has zero volume integral, which is precisely the adjoint Fredholm compatibility condition; its solutions differ by a multiple of ρ_0, fixed by the required integral. Elliptic regularity and the bounded inverse theorem give a bounded inverse on these spaces. Since the coefficients are jointly smooth, $x \mapsto F_x$ is smooth in operator norm. The inverse map on bounded invertible operators is smooth, so F_x^{-1} exists and is smooth locally near x_0. Set $\widetilde{\rho}_x = F_x^{-1}\rho_0$. Integrating this equation gives $\int \widetilde{\rho}_x\ \mathrm{d}a = 1$, and hence $L_x^*\widetilde{\rho}_x = 0$. The one-dimensional kernel and normalization identify $\widetilde{\rho}_x = \rho_x$. These local constructions agree by uniqueness. Since k is arbitrary, they give all mixed derivatives in (x,a) and hence joint smoothness of ρ^Ψ on $\mathsf{X}\times\mathsf{A}$.

Now write $\Pi_x v = \int_{\mathsf{A}} v(a)\rho_x(a)\,\mathrm{d}a$, interpreted as a constant function, and define

$$B_x = -L_x + \Pi_x : C^{k+2,\gamma}(\mathsf{A}) \longrightarrow C^{k,\gamma}(\mathsf{A}). \tag{38}$$

This operator is bijective. Indeed, $\Pi_x L_x = 0$ and $\Pi_x^2 = \Pi_x$ show that $B_x v = f$ is equivalent to $\Pi_x v = \Pi_x f$ and $-L_x v = f - \Pi_x f$, which has a unique solution with the prescribed invariant mean. The established regularity of ρ_x makes B_x a smooth operator family, so B_x^{-1} is locally smooth. For $g_x = g(x,\cdot)$ with $\Pi_x g_x = 0$, the function $u_x = B_x^{-1} g_x$ therefore depends smoothly on x in every such Hölder space. Averaging its equation gives $\Pi_x u_x = 0$ and then $-L_x u_x = g_x$, proving existence and joint smoothness. Finally, the difference of two solutions lies in $\ker L_x$ for each x, so it equals a state-only function $c(x)$. The centering condition forces $c(x) = 0$, proving uniqueness. □

A.2 EXECUTION LIMITS

Proof of Proposition 3.1. Smooth semigroup expansions on the compact domain yield

$$P_h S_\tau^\Psi f = f + h\mathcal{L}_X^a f + \tau \mathcal{A}^\Psi f + O(h^2 + h\tau + \tau^2), \tag{39}$$

uniformly for smooth fixed f. If $c_h = \tau_h/h$ remains bounded and $\tau_h \to 0$, then $\tau_h = O(h)$, and division by h gives

$$\frac{\mathcal{K}_{h,\tau_h}^\Psi f - f}{h} = \mathcal{L}_X^a f + c_h \mathcal{A}^\Psi f + O(h). \tag{40}$$

Hence, if $c_h \to c < \infty$, the first limit follows.

If instead $c_h \to \infty$ and $\tau_h \to 0$, then $h/\tau_h = 1/c_h \to 0$. Dividing by τ_h gives

$$\frac{\mathcal{K}_{h,\tau_h}^\Psi f - f}{\tau_h} = \frac{h}{\tau_h}\mathcal{L}_X^a f + \mathcal{A}^\Psi f + O\left(\frac{h^2}{\tau_h} + h + \tau_h\right). \tag{41}$$

Since $\tau_h/h \to \infty$, we have $\tau_h > h$ for all sufficiently small h. Hence $0 \le \frac{h^2}{\tau_h} \le h \to 0$, which implies $\frac{h^2}{\tau_h} = o(1)$. Together with $h \to 0$ and $\tau_h \to 0$, this yields $\frac{h^2}{\tau_h} + h + \tau_h = o(1)$. Thus the second limit follows.

If $\tau_h \to \tau_0 \in (0, \infty)$, the Markov contraction property and strong continuity of the semigroups give

$$\left\|P_h S_{\tau_h}^\Psi f - S_{\tau_0}^\Psi f\right\|_\infty \le \left\|S_{\tau_h}^\Psi f - S_{\tau_0}^\Psi f\right\|_\infty + \left\|P_h S_{\tau_0}^\Psi f - S_{\tau_0}^\Psi f\right\|_\infty \longrightarrow 0. \tag{42}$$

Similarly, since $\Pi^\Psi f$ is continuous, $\left\|P_h \Pi^\Psi f - \Pi^\Psi f\right\|_\infty \longrightarrow 0$. For any Markov kernel K, $\|K\|_{\infty\to\infty} \le 1$, and hence the Neumann series gives

$$(I - e^{-\beta h}K)^{-1} = \sum_{n=0}^{\infty} e^{-n\beta h} K^n, \qquad \left\|(I - e^{-\beta h}K)^{-1}\right\|_{\infty\to\infty} \le \frac{1}{1 - e^{-\beta h}} = O(h^{-1}). \tag{43}$$

Finally, suppose $c_h \to c \in (0, \infty)$ and define the Bellman residual of Q_c^Ψ by

$$R_h := Q_c^\Psi - h\ell^\Psi - e^{-\beta h}\mathcal{K}_{h,\tau_h}^\Psi Q_c^\Psi. \tag{44}$$

Since $\tau_h = c_h h = O(h)$,

$$e^{-\beta h}\mathcal{K}_{h,\tau_h}^\Psi Q_c^\Psi = Q_c^\Psi + h(\mathcal{L}_X^a + c_h \mathcal{A}^\Psi - \beta)Q_c^\Psi + O(h^2). \tag{45}$$

Using $(\beta - \mathcal{L}_X^a - c\mathcal{A}^\Psi)Q_c^\Psi = \ell^\Psi$, we obtain $R_h = h(c - c_h)\mathcal{A}^\Psi Q_c^\Psi + O(h^2) = o(h)$. Since Q_{h,τ_h}^Ψ satisfies the discrete Bellman equation and Q_c^Ψ has Bellman residual R_h, subtraction gives

$$\left(I - e^{-\beta h}\mathcal{K}_{h,\tau_h}^\Psi\right)\left(Q_{h,\tau_h}^\Psi - Q_c^\Psi\right) = -R_h. \tag{46}$$

Hence, by Eq. (43),

$$\left\|Q_{h,\tau_h}^\Psi - Q_c^\Psi\right\|_\infty \le \left\|\left(I - e^{-\beta h}\mathcal{K}_{h,\tau_h}^\Psi\right)^{-1}\right\|_{\infty\to\infty} \|R_h\|_\infty = O(h^{-1})\, o(h) = o(1). \tag{47}$$

Therefore, $Q_{h,\tau_h}^\Psi \longrightarrow Q_c^\Psi$ uniformly as $h \downarrow 0$. □

A.3 STATIONARY ACTION GRADIENTS

Proof of Theorem 3.2. For a smooth state-only function v, the physical semigroup satisfies $P_h v = v + h\mathcal{L}_X^a v + O(h^2)$ uniformly. Hence, $T_h^\Psi V^\Psi = V^\Psi + h\bar{\mathcal{L}}^\Psi V^\Psi + O(h^2)$, and, using $e^{-\beta h} = 1 - \beta h + O(h^2)$ together with $(\beta - \bar{\mathcal{L}}^\Psi)V^\Psi = \bar{\ell}^\Psi$, we obtain $[I - e^{-\beta h}T_h^\Psi]V^\Psi = h\bar{\ell}^\Psi + O(h^2)$. Since V_h^Ψ satisfies $[I - e^{-\beta h}T_h^\Psi]V_h^\Psi = h\bar{\ell}^\Psi$, subtraction and Eq. (43) yield $\left\|V_h^\Psi - V^\Psi\right\|_\infty = O(h^{-1})O(h^2) = O(h)$.

Using the reconstruction $Q_h^\Psi = h\ell^\Psi + e^{-\beta h}P_h V_h^\Psi$, the Markov contraction property, the preceding estimate, and the weak expansion of $P_h V^\Psi$ give $\left\|Q_h^\Psi - V^\Psi\right\|_\infty = O(h)$.

We next establish the first-order action-dependent expansion. For every smooth state-only v, the parameter-dependent weak expansion satisfies

$$\|P_h v - v - h\mathcal{L}_X^a v\|_{C_a^1} \le Ch^2 \|v\|_{C^6(\mathsf{X})}. \tag{48}$$

This follows from the iterated Dynkin formula together with differentiation of the frozen-action stochastic flow with respect to its action parameter.

Under Assumption A.2, the uniform C^6 bound, together with $\left\|V_h^\Psi - V^\Psi\right\|_\infty = O(h)$ and interpolation, implies $V_h^\Psi \longrightarrow V^\Psi$ in $C^2(\mathsf{X})$. Subtracting V_h^Ψ from the reconstruction formula gives

$$\frac{Q_h^\Psi - V_h^\Psi}{h} = \ell^\Psi + e^{-\beta h}\frac{P_h V_h^\Psi - V_h^\Psi}{h} - \frac{1-e^{-\beta h}}{h}V_h^\Psi. \tag{49}$$

By Eq. (48),

$$\frac{P_h V_h^\Psi - V_h^\Psi}{h} = \mathcal{L}_X^a V_h^\Psi + O_{C_a^1}(h). \tag{50}$$

Since $V_h^\Psi \to V^\Psi$ in C^2, smoothness of the coefficients implies $\mathcal{L}_X^a V_h^\Psi \longrightarrow \mathcal{L}_X^a V^\Psi$ in C_a^1. Together with $e^{-\beta h} \to 1$, $\frac{1-e^{-\beta h}}{h} \to \beta$, this yields

$$\frac{Q_h^\Psi - V_h^\Psi}{h} \longrightarrow \ell^\Psi + \mathcal{L}_X^a V^\Psi - \beta V^\Psi = q^\Psi \tag{51}$$

in C_a^1, which proves Eq. (13). □

A.4 The finite-horizon corrector

Lemma A.4 (Uniform mixing and smoothing of the frozen action process)**.** *Under Assumption A.1, fix a smooth policy field Ψ. There exists $\gamma = \gamma(\Psi) > 0$ such that, for every integer $j \ge 0$, there is $C_j = C_j(\Psi) < \infty$, independent of x, t, and g, for which every smooth g with $\Pi^\Psi g = 0$ satisfies*

$$\left\|S_t^\Psi g(x,\cdot)\right\|_{C^j(\mathsf{A})} \le C_j e^{-\gamma t}\|g(x,\cdot)\|_\infty, \qquad x \in \mathsf{X}, \quad t \ge 1. \tag{52}$$

Taking the supremum over x gives

$$\left\|S_t^\Psi g\right\|_{C_a^j} \le C_j e^{-\gamma t}\|g\|_\infty, \qquad t \ge 1. \tag{53}$$

Proof. Let P_t^x be the action semigroup generated by $L_x = \Psi(x,\cdot)\cdot\nabla_a + \Delta_a$, and write $\pi_x = \pi^\Psi(\cdot \mid x)$. Let $p_x(t,a,a')$ be its transition density with respect to normalized Lebesgue measure m on A. Joint smoothness of Ψ on the compact product $\mathsf{X}\times\mathsf{A}$ gives uniform bounds on all coefficient derivatives, and the second-order part is uniformly elliptic. Standard parabolic regularity with parameters implies that, for each $t > 0$, $p_x(t,a,a')$ and all its derivatives in a are jointly continuous in (x,a,a'). The heat kernel is strictly positive because A is connected. Thus, by compactness,

$$\delta := \tfrac{1}{2}\inf_{x,a,a'} p_x(1,a,a') \in (0,1), \qquad P_1^x(a,B) \ge \delta m(B) \tag{54}$$

for every Borel set $B \subseteq \mathsf{A}$.

Write $P_1^x = \delta m + (1-\delta)R_x$, where R_x is a Markov kernel. For $\operatorname{osc}(f) := \sup f - \inf f$, this gives $\operatorname{osc}(P_1^x f) \le (1-\delta)\operatorname{osc}(f)$. Iteration and invariance of π_x yield

$$\|P_n^x f - \pi_x f\|_\infty \le \operatorname{osc}(P_n^x f) \le 2(1-\delta)^n \|f\|_\infty, \qquad n \in \mathbb{N}_0. \tag{55}$$

For $t = n + s$, with $n = \lfloor t \rfloor$ and $s \in [0,1)$, the supremum-norm contraction of P_s^x therefore gives

$$\|P_t^x f - \pi_x f\|_\infty \le 2(1-\delta)^{\lfloor t \rfloor}\|f\|_\infty \le Ce^{-\gamma t}\|f\|_\infty, \qquad t \ge 0, \tag{56}$$

where $\gamma = -\log(1-\delta) > 0$ and $C = 2e^\gamma$ are independent of x.

The heat-kernel regularity and compactness also give, for every $j \ge 0$,

$$K_j := \sum_{|\alpha|\le j}\sup_{x,a}\int_{\mathsf{A}} |\partial_a^\alpha p_x(1,a,a')|\, m(\mathrm{d}a') < \infty.$$

Differentiating the kernel representation in the initial action a then yields

$$\|P_1^x f\|_{C^j(\mathsf{A})} \le K_j \|f\|_\infty . \tag{57}$$

For $g_x := g(x,\cdot)$ with $\pi_x g_x = 0$, the semigroup property and (56)–(57) imply

$$\|P_t^x g_x\|_{C^j(\mathsf{A})} \le K_j \left\|P_{t-1}^x g_x\right\|_\infty \le K_j C e^{-\gamma(t-1)} \|g_x\|_\infty , \qquad t \ge 1. \tag{58}$$

Taking $C_j = K_j C e^\gamma$ proves (52). □

By Lemma A.4, for every fixed $\tau > 0$ and smooth centered g, the series

$$\sum_{n=0}^{\infty} S_{n\tau}^\Psi g \tag{59}$$

converges in C_a^j for every finite j: its tail decays geometrically, and the finitely many terms with $n\tau < 1$ are smooth. Moreover,

$$(I - S_\tau^\Psi) \sum_{n=0}^{N} S_{n\tau}^\Psi g = g - S_{(N+1)\tau}^\Psi g, \tag{60}$$

and Eq. (52) implies that the last term vanishes as $N \to \infty$. Hence

$$(I - S_\tau^\Psi)^{-1} g = \sum_{n=0}^{\infty} S_{n\tau}^\Psi g \tag{61}$$

on the centered subspace.

A.5 Finite-horizon correctors

Theorem A.5 (Finite-horizon action corrector)**.** *Under Assumption A.1, fix a policy field Ψ and an action horizon $\tau > 0$. For $h > 0$, let $V_{h,\tau}^\Psi := \Pi^\Psi Q_{h,\tau}^\Psi$. Then there exists a unique smooth centered function χ_τ^Ψ satisfying*

$$(I - S_\tau^\Psi)\chi_\tau^\Psi = q^\Psi, \qquad \Pi^\Psi \chi_\tau^\Psi = 0. \tag{62}$$

For fixed $\tau > 0$, as $h \downarrow 0$,

$$\left\|V_{h,\tau}^\Psi - V^\Psi\right\|_\infty = O(h), \tag{63}$$

$$\left\|Q_{h,\tau}^\Psi - V_{h,\tau}^\Psi - h\chi_\tau^\Psi\right\|_\infty = O(h^2). \tag{64}$$

If, in addition,

$$\sup_{0<h\le h_0} \left\|Q_{h,\tau}^\Psi\right\|_{C_a^2} < \infty, \tag{65}$$

then

$$\left\| \frac{Q_{h,\tau}^\Psi - V_{h,\tau}^\Psi}{h} - \chi_\tau^\Psi \right\|_{C_a^1} = O(h^{1/2}). \tag{66}$$

Moreover, the corrector satisfies

$$\left\|\chi_\tau^\Psi - q^\Psi\right\|_{C_a^1} \longrightarrow 0, \tau \to \infty, \tag{67}$$

$$\left\|\tau\chi_\tau^\Psi - U^\Psi\right\|_{C_a^1} \longrightarrow 0, \tau \downarrow 0. \tag{68}$$

For fixed τ, Eqs. (63)–(66) identify χ_τ^Ψ as the leading-order action-dependent component of the finite-horizon critic. The additional bound (65) controls the action curvature required to upgrade the uniform expansion to a C_a^1 estimate. The constants in the h-remainders are allowed to depend on τ. Consequently, Eqs. (67) and (68) describe the limiting behavior of the corrector itself, but do not imply a joint limit for $Q_{h,\tau}^\Psi$ or its action gradient when h and τ vanish simultaneously. Such a conclusion would require estimates uniform in τ together with a prescribed relation between h and τ.

Proof of Theorem A.5. Suppress the dependence on Ψ and τ, and write $S = S_\tau^\Psi, \Pi = \Pi^\Psi, D = \beta - \mathcal{L}_X^a, \bar{D} = \beta - \bar{\mathcal{L}}^\Psi$. Since P_h is the Markov semigroup generated by $\mathcal{L}_X^a$, $P_h = e^{h\mathcal{L}_X^a}$. Hence, $e^{-\beta h}P_h = e^{-hD}$. Let $\chi = (I-S)^{-1}q, \Pi\chi = 0$. Since $q = \ell - DV$, the identity $(I-S)\chi = q$ determines the leading centered correction.

We seek an approximate solution of the form

$$\widetilde{Q}_h = V + h(W_1 + \chi) + h^2(W_2 + \chi_2) + h^3\chi_3, \tag{69}$$

where W_1, W_2 are state-only and χ, χ_2, χ_3 are centered. Expanding $(I - e^{-hD}S)\widetilde{Q}_h$ in powers of h, the first-order identity follows from $(I-S)\chi = q = \ell - DV$. At each higher order, we first choose the state-only correction so that the remaining right-hand side has zero Π-mean, and then solve the corresponding discrete Poisson equation on the centered subspace. Matching the coefficients of h^2 and h^3 gives

$$\bar{D}W_1 = \Pi\left[-DS\chi + \frac{1}{2}D^2V\right], \tag{70}$$

$$(I-S)\chi_2 = -DS\chi - DW_1 + \frac{1}{2}D^2V, \tag{71}$$

$$\bar{D}W_2 = \Pi\left[-DS\chi_2 + \frac{1}{2}D^2S(\chi + W_1) - \frac{1}{6}D^3V\right], \tag{72}$$

$$(I-S)\chi_3 = -DS\chi_2 - DW_2 + \frac{1}{2}D^2S(\chi + W_1) - \frac{1}{6}D^3V. \tag{73}$$

The equations for W_1 and W_2 enforce the centering conditions required for solvability of the corresponding discrete Poisson equations. Since $\beta > 0$, the averaged operator $\bar{D}$ is invertible, and hence all four equations are well posed.

Because $e^{-\beta h}P_h = e^{-hD}$, expansion through third order gives

$$\left(I - e^{-\beta h}P_hS\right)\widetilde{Q}_h - h\ell = O(h^4), \tag{74}$$

where the coefficients of h, h^2, and h^3 vanish by Eqs. (71)–(73). Since the exact critic satisfies $\left(I - e^{-\beta h}P_hS\right)Q_{h,\tau} = h\ell$, Eq. (43) yields

$$\left\|Q_{h,\tau} - \widetilde{Q}_h\right\|_\infty = O(h^{-1})O(h^4) = O(h^3). \tag{75}$$

Since χ, χ_2, χ_3 are centered, $V_{h,\tau} = \Pi Q_{h,\tau} = V + hW_1 + O(h^2)$, and therefore $Q_{h,\tau} - V_{h,\tau} = h\chi + h^2\chi_2 + O(h^3)$. This proves $\|V_{h,\tau} - V\|_\infty = O(h)$ and $\|Q_{h,\tau} - V_{h,\tau} - h\chi\|_\infty = O(h^2)$.

Assume now that $\sup_{0<h\le h_0}\|Q_{h,\tau}\|_{C_a^2} < \infty$. Let $E_h := Q_{h,\tau} - \widetilde{Q}_h$. The preceding estimate gives $\|E_h\|_\infty = O(h^3)$, while the curvature assumption and smoothness of $\widetilde{Q}_h$ imply a uniform C_a^2 bound on E_h. Interpolation in the action variable therefore gives

$$\|\nabla_a E_h\|_\infty \le C\left(\|E_h\|_\infty + \|E_h\|_\infty^{1/2}\|E_h\|_{C_a^2}^{1/2}\right) = O(h^{3/2}). \tag{76}$$

Since $\nabla_a(h^2\chi_2 + h^3\chi_3) = O(h^2)$, subtracting the action average and dividing by h yields

$$\left\|\frac{Q_{h,\tau} - V_{h,\tau}}{h} - \chi\right\|_{C_a^1} = O(h^{1/2}), \tag{77}$$

which proves Eq. (66).

Finally, Eq. (59) gives $\chi_\tau = \sum_{n=0}^\infty S_{n\tau}q$. As $\tau \to \infty$, $\chi_\tau - q = \sum_{n\ge1} S_{n\tau}q \longrightarrow 0$ in C_a^1 by exponential decay on centered functions. As $\tau \downarrow 0$, $\tau\chi_\tau = \tau\sum_{n=0}^\infty S_{n\tau}q$ is a Riemann sum for $\int_0^\infty S_tq\,\mathrm{d}t$. This integral equals the centered Poisson solution U satisfying $-\mathcal{A}U = q$. Hence $\chi_\tau \longrightarrow q$ as $\tau \to \infty$ and $\tau\chi_\tau \longrightarrow U$ as $\tau \downarrow 0$ in C_a^1. □

The expansion above also shows why the uncentered approximation $Q_{h,\tau} = V + h\chi_\tau + o(h)$ is generally invalid: the term hW_1 produces a state-only first-order bias. This term disappears after centering and therefore does not affect the action gradient. The estimates are uniform for τ in compact subsets of $(0,\infty)$ when the curvature bound is uniform there; no uniform remainder as $\tau \downarrow 0$ is asserted.

A.6 CONTINUOUS FAST-ACTION EVALUATION

Proposition A.6 (Continuous fast-action corrector)**.** *Under Assumption A.1, fix a policy field Ψ. Then, as $c \to \infty$,*

$$\left\| Q_c^\Psi - \Pi^\Psi Q_c^\Psi - \frac{1}{c} U^\Psi \right\|_\infty = O(c^{-2}), \tag{78}$$

where U^Ψ is the centered Poisson corrector defined in Eq. (10)*. Equivalently,*

$$\left\| c\big(Q_c^\Psi - \Pi^\Psi Q_c^\Psi\big) - U^\Psi \right\|_\infty = O(c^{-1}), \tag{79}$$

and therefore $c\big(Q_c^\Psi - \Pi^\Psi Q_c^\Psi\big) \longrightarrow U^\Psi$ uniformly as $c \to \infty$.

Proposition A.6 identifies U^Ψ as the leading-order centered correction to the continuous-time critic in the fast-action regime. Together with Theorem A.5, this gives $\chi_\tau^\Psi \to q^\Psi$ as $\tau \to \infty$ and $\tau\chi_\tau^\Psi \to U^\Psi$ as $\tau \downarrow 0$.

The estimate in Eq. (78) is only in the uniform norm and therefore does not imply a corresponding expansion for $\nabla_a Q_c^\Psi$. Likewise, the fixed-τ gradient estimate in Theorem A.5 is not uniform as $\tau \downarrow 0$. Thus no general simultaneous limit for $h, \tau \downarrow 0$ is claimed. For the LQ model, Appendix C.4 verifies directly that $\frac{\tau}{h}\partial_a Q_{h,\tau} \longrightarrow \partial_a U$ when $\tau \downarrow 0$ and $h/\tau \to 0$.

Proof of Proposition A.6. Suppress the dependence on Ψ and set $\varepsilon = c^{-1}$. Write $D := \beta - \mathcal{L}_X^a$, $\bar{D} := \beta - \bar{\mathcal{L}}$. Then Q_c satisfies $(D - \varepsilon^{-1}\mathcal{A})Q_c = \ell$.

We construct an approximation through second order in ε. Since $q = \ell - DV, -\mathcal{A}U = q$, the leading correction to the averaged value V is U. The first-order residual DU, however, need not be centered. We therefore introduce a state-only correction V_1 satisfying $\bar{D}V_1 = \Pi\mathcal{L}_X^a U$. Then $F_1 := D(U + V_1)$ is centered, since $\Pi F_1 = -\Pi\mathcal{L}_X^a U + \bar{D}V_1 = 0$. Hence the centered Poisson equation $\mathcal{A}W_2 = F_1, \Pi W_2 = 0$, has a unique smooth solution.

Define $\widetilde{Q}_\varepsilon := V + \varepsilon(U + V_1) + \varepsilon^2 W_2$. Since V and V_1 are state-only, $\mathcal{A}V = \mathcal{A}V_1 = 0$. Therefore

$$(D - \varepsilon^{-1}\mathcal{A})\widetilde{Q}_\varepsilon = DV - \mathcal{A}U + \varepsilon\left[D(U + V_1) - \mathcal{A}W_2\right] + \varepsilon^2 DW_2. \tag{80}$$

Using $q = \ell - DV$, $-\mathcal{A}U = q$, and $\mathcal{A}W_2 = D(U + V_1)$, the first two orders cancel, and hence $(D - \varepsilon^{-1}\mathcal{A})\widetilde{Q}_\varepsilon - \ell = \varepsilon^2 DW_2$.

Since $D - \varepsilon^{-1}\mathcal{A} = \beta - (\mathcal{L}_X^a + c\mathcal{A}) = \beta - \mathcal{G}_c$, the discounted Markov resolvent satisfies $\left\|(\beta - \mathcal{G}_c)^{-1}\right\|_{\infty\to\infty} \le \frac{1}{\beta}$. Subtracting the equations for Q_c and $\widetilde{Q}_\varepsilon$ therefore yields $\left\| Q_c - \widetilde{Q}_\varepsilon \right\|_\infty = O(\varepsilon^2)$.

Finally, V and V_1 are state-only, while U and W_2 are centered. Hence $(I - \Pi)\widetilde{Q}_\varepsilon = \varepsilon U + \varepsilon^2 W_2$. Applying $I - \Pi$ to the preceding error estimate gives $Q_c - \Pi Q_c = \varepsilon U + O(\varepsilon^2)$. Since $\varepsilon = c^{-1}$, this proves Eq. (78). □

B PROOFS FOR POLICY-IMPROVEMENT GEOMETRY

Proof of Theorem 4.1. For fixed $h > 0$, the operator $T_h^\Psi = \Pi^\Psi P_h$ maps bounded state functions to smooth functions, by uniform ellipticity of the physical diffusion and smoothness of the invariant action density. The reduced Bellman equation (11) therefore makes V_h^Ψ smooth, and (12) makes Q_h^Ψ smooth. Since $\Pi^\Psi(Q_h^\Psi - V_h^\Psi) = 0$, Lemma A.3 provides a smooth centered U_h^Ψ. The corresponding regularity at $h = 0$ follows from (9) and (10). Thus B_h^Ψ and the damped candidate fields are smooth for every $h \ge 0$.

Performance difference. The Poisson equation and invariance under the candidate policy give

$$\Pi^\Phi q_h^\Psi = -\Pi^\Phi \mathcal{A}^\Psi U_h^\Psi = \Pi^\Phi(\mathcal{A}^\Phi - \mathcal{A}^\Psi)U_h^\Psi = \Pi^\Phi(\Phi - \Psi)\cdot\nabla_a U_h^\Psi. \tag{81}$$

Consequently,

$$D_{h,\Psi}(\Phi) = \Pi^\Phi\left[(\Phi - \Psi)\cdot\nabla_a U_h^\Psi - \frac{\lambda_0}{2}(\|\Phi\|^2 - \|\Psi\|^2)\right]. \tag{82}$$

Completing the square with $B_h^\Psi = \nabla_a U_h^\Psi / \lambda_0$ proves (18).

For $h > 0$, the reconstruction formula (12) gives

$$q_h^\Psi = \ell^\Psi + \frac{e^{-\beta h} P_h V_h^\Psi - V_h^\Psi}{h}. \tag{83}$$

Since $\ell^\Phi - \ell^\Psi = -\lambda_0(\|\Phi\|^2 - \|\Psi\|^2)/2$, averaging under Φ yields

$$h D_{h,\Psi}(\Phi) = h\bar{\ell}^\Phi + e^{-\beta h} T_h^\Phi V_h^\Psi - V_h^\Psi. \tag{84}$$

Subtracting this expression from the reduced Bellman equation for V_h^Φ gives

$$(I - e^{-\beta h} T_h^\Phi)(V_h^\Phi - V_h^\Psi) = h D_{h,\Psi}(\Phi). \tag{85}$$

For $h = 0$, subtraction of the averaged Bellman equations similarly gives

$$(\beta - \bar{\mathcal{L}}^\Phi)(V^\Phi - V^\Psi) = D_{0,\Psi}(\Phi). \tag{86}$$

These identities prove (19). Both resolvents preserve nonnegativity: for $h > 0$,

$$\mathcal{R}_h^\Phi f = h \sum_{n=0}^{\infty} e^{-\beta h n} (T_h^\Phi)^n f, \tag{87}$$

and for $h = 0$ the analogous representation uses the discounted semigroup of $\bar{\mathcal{L}}^\Phi$. Thus a nonnegative gap implies improvement.

Poisson update. For $\Phi_\eta = (1-\eta)\Psi + \eta B_h^\Psi$, $\left\|\Psi - B_h^\Psi\right\|^2 - \left\|\Phi_\eta - B_h^\Psi\right\|^2 = \eta(2-\eta)\left\|\Psi - B_h^\Psi\right\|^2$. Substitution proves (20). If $\Psi \not\equiv B_h^\Psi$, strict positivity of the invariant action density makes $\Pi^{\Phi_\eta}\left\|\Psi - B_h^\Psi\right\|^2$ nonnegative and nonzero. For $h > 0$, $T_h^{\Phi_\eta}$ has a strictly positive kernel on the connected state torus, by uniform ellipticity of the physical diffusion. For $h = 0$, the averaged generator is uniformly elliptic and has a strictly positive transition density. Hence either resolvent maps this nonzero nonnegative function to a strictly positive function, proving strict improvement at every state. □

Proof of Proposition 4.2. By definition, $e_\alpha^\Psi = 0 \iff \frac{\nabla_a q^\Psi}{\alpha} = \frac{\nabla_a U^\Psi}{\lambda_0}$. Hence $\nabla_a\left(\frac{q^\Psi}{\alpha} - \frac{U^\Psi}{\lambda_0}\right) = 0$. Since the action space is connected, the quantity in parentheses is independent of a. Both q^Ψ and U^Ψ are centered under Π^Ψ, so this state-only constant vanishes. Therefore $\frac{q^\Psi}{\alpha} = \frac{U^\Psi}{\lambda_0}$. Applying $-\mathcal{A}^\Psi$ and using $-\mathcal{A}^\Psi U^\Psi = q^\Psi$ gives $-\mathcal{A}^\Psi q^\Psi = \frac{\alpha}{\lambda_0} q^\Psi$. This proves necessity.

Conversely, suppose $-\mathcal{A}^\Psi q^\Psi = \frac{\alpha}{\lambda_0} q^\Psi$. Then $\widetilde{U}^\Psi := \frac{\lambda_0}{\alpha} q^\Psi$ is centered and satisfies $-\mathcal{A}^\Psi \widetilde{U}^\Psi = q^\Psi$. By uniqueness of the centered Poisson solution, $U^\Psi = \frac{\lambda_0}{\alpha} q^\Psi$, and hence $e_\alpha^\Psi = 0$.

Writing $\omega := \frac{\alpha}{\lambda_0}$, the stationary calibration is therefore $\lambda_A(h) = \lambda_0 \omega h$. Moreover, $S_\tau^\Psi q^\Psi = e^{-\omega\tau} q^\Psi$, so the discrete Poisson equation $(I - S_\tau^\Psi)\chi_\tau^\Psi = q^\Psi$ gives $\chi_\tau^\Psi = \frac{q^\Psi}{1 - e^{-\omega\tau}}$. Matching the limiting finite-horizon actor target $h\nabla_a \chi_\tau^\Psi / \lambda_A(h)$ to $B^\Psi = \nabla_a q^\Psi / (\lambda_0 \omega)$ yields the finite-horizon calibration in Proposition 4.2. □

Proof of Corollary 4.3. By Theorem 3.2,

$$\epsilon_h = \left\| \frac{\nabla_a Q_h^\Psi}{h} - \nabla_a q^\Psi \right\|_\infty \longrightarrow 0. \tag{88}$$

Using the definition of e_α^Ψ and the triangle inequality,

$$\left\|\Phi_h - B^\Psi\right\|_\infty \le \left\| \frac{\nabla_a \widehat{Q}_h - \nabla_a Q_h^\Psi}{\alpha h} \right\|_\infty + \left\| \frac{\nabla_a Q_h^\Psi}{\alpha h} - \frac{\nabla_a q^\Psi}{\alpha} \right\|_\infty + \left\| \frac{\nabla_a q^\Psi}{\alpha} - B^\Psi \right\|_\infty \tag{89}$$

$$\le \frac{\eta_h}{\alpha h} + \frac{\epsilon_h}{\alpha} + e_\alpha^\Psi. \tag{90}$$

By the $h = 0$ case of Eq. (18),

$$D_\Psi(\Phi_h) \geq -\frac{\lambda_0}{2}\Pi^{\Phi_h}\left\|\Phi_h - B^\Psi\right\|^2 \geq -\frac{\lambda_0}{2}\left\|\Phi_h - B^\Psi\right\|_\infty^2. \tag{91}$$

Using the preceding bound on $\left\|\Phi_h - B^\Psi\right\|_\infty$ gives

$$D_\Psi(\Phi_h) \geq -\frac{\lambda_0}{2}\left(e_\alpha^\Psi + \frac{\epsilon_h}{\alpha} + \frac{\eta_h}{\alpha h}\right)^2. \tag{92}$$

Since the discounted resolvent $(\beta - \bar{\mathcal{L}}^{\Phi_h})^{-1}$ is positivity preserving and $(\beta - \bar{\mathcal{L}}^{\Phi_h})^{-1}1 = \frac{1}{\beta}$, Eq. (19) yields Eq. (22). If $e_\alpha^\Psi = 0$ and $\eta_h = o(h)$, then $\epsilon_h \to 0$ and $\eta_h/h \to 0$, so the right-hand side of Eq. (22) converges uniformly to zero. This proves Eq. (23). □

C Direct verification for the LQ family

This appendix establishes the LQ conclusions of Section 5 directly on $\mathbb{R}^2$, with $\beta = 1$ and $\theta \in \mathcal{D}$. Polynomial estimates use the coefficient norm $\|\cdot\|_{\mathrm{pol}}$ on $\mathcal{P}_2 = \mathrm{span}\{1, x, a, x^2, xa, a^2\}$, and its induced operator norms. Uniformity refers to compact parameter sets $K \Subset \mathcal{D}$; the physical state and action spaces remain unbounded. Appendix C.1 proves the moment bounds, vanishing discounted terminal terms, and the performance-difference identity on $\mathbb{R}$. Appendix C.2 verifies the polynomial action equations and preservation of $\mathcal{D}$. Appendices C.3 and C.4 justify the finite-step and finite-action-horizon limits in polynomial coefficient norms, uniformly on compact subsets of $\mathcal{D}$.

C.1 Averaged value and performance comparison

Averaging the physical generator under the Gaussian law in Section 5 gives

$$\bar{\mathcal{L}}^\theta f = -xf'(x) + \frac{1}{2}\left[(px+d)^2 + \kappa^{-1}\right]f''(x). \tag{93}$$

The diffusion coefficient $\sqrt{(px+d)^2 + \kappa^{-1}}$ is globally Lipschitz with linear growth. Thus the averaged process is nonexplosive and has finite moments of every order on finite time intervals. Its first two moments satisfy

$$\mathbb{E}_x^\theta X_t = xe^{-t}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_x^\theta X_t^2 = (p^2-2)\mathbb{E}_x^\theta X_t^2 + 2pdxe^{-t} + d^2 + \kappa^{-1}. \tag{94}$$

Variation of constants bounds the second moment by $C_\theta(1+x^2)(1+t)e^{(p^2-2)_+ t}$, where $z_+ = \max\{z, 0\}$. Consequently, $p^2 < 3$ implies

$$\int_0^\infty e^{-t}\mathbb{E}_x^\theta(1+X_t^2)\,\mathrm{d}t < \infty, \qquad \lim_{T\to\infty} e^{-T}\mathbb{E}_x^\theta(1+X_T^2) = 0. \tag{95}$$

If $p^2 < 2$, the same moment equation gives $\sup_{t\geq 0}\mathbb{E}_x^\theta X_t^2 < \infty$.

Coefficient matching in $(1-\bar{\mathcal{L}}^\theta)V^\theta = \bar{\ell}^\theta$ gives $V^\theta = vx^2/2 + wx + z$, with v from Eq. (26) and

$$w = \tfrac{1}{2}[-(d + 2pd + 1 + 2p) + vpd], \tag{96}$$

$$z = -(d^2 + \kappa^{-1} + 2d) - \tfrac{1}{2}\lambda_0\kappa + \tfrac{v}{2}(d^2 + \kappa^{-1}). \tag{97}$$

Both this candidate and $\bar{\ell}^\theta$ have at most quadratic growth. The discounted Dynkin formula gives

$$V^\theta(x) = \mathbb{E}_x^\theta \int_0^T e^{-t}\bar{\ell}^\theta(X_t)\,\mathrm{d}t + e^{-T}\mathbb{E}_x^\theta V^\theta(X_T). \tag{98}$$

By Eq. (95), the terminal term vanishes and the integral converges absolutely as $T \to \infty$. Thus the polynomial candidate is the actual discounted averaged value.

For two policies $\theta, \vartheta \in \mathcal{D}$, set

$$D_\theta(\Psi_\vartheta) = \bar{\ell}^\vartheta + \bar{\mathcal{L}}^\vartheta V^\theta - V^\theta. \tag{99}$$

Applying the same formula to V^θ along the process under ϑ, and using Eq. (95) for ϑ, yields

$$V^\vartheta(x) - V^\theta(x) = \mathbb{E}_x^\vartheta \int_0^\infty e^{-t}D_\theta(\Psi_\vartheta)(X_t)\,\mathrm{d}t. \tag{100}$$

All terms have at most quadratic growth. This establishes the performance-difference identity on $\mathbb{R}$ for the affine family; in particular, a nonnegative gap implies value improvement.

C.2 Polynomial action equations and the affine actor map

With $\mu = px + d$ and $y = a - \mu$, Gaussian projection maps $\mathcal{P}_2$ onto $\mathrm{span}\{1, x, x^2\}$, and

$$\ker \Pi^\theta \cap \mathcal{P}_2 = \mathrm{span}\{y, xy, y^2 - \kappa^{-1}\}. \tag{101}$$

The action generator acts on these centered modes at rates $\kappa, \kappa, 2\kappa$, as in the main-text calculation. Thus $-\mathcal{A}^\theta$ and $I - S_\tau^\theta$ are invertible on this subspace for every $\tau > 0$, and $\Pi^\theta S_\tau^\theta = S_\tau^\theta \Pi^\theta = \Pi^\theta$. These polynomial identities supply the action resolvents needed below.

Substituting the verified value into $\ell^\theta + \mathcal{L}_X^a V^\theta - V^\theta$ gives q^θ in Eq. (27). Dividing its centered mode coefficients by their action rates gives the unique centered polynomial solution U^θ in Eq. (28). Differentiating q^θ with respect to a and dividing by α gives the actor map (29). For $\alpha > 0$, $H > 0$ implies $\kappa' > 0$, and

$$p' = \frac{\lambda_0 \kappa^2}{H} p + \frac{R}{H}(-1/R), \qquad R(p) - R_* = \frac{R_*(p - p_*)^2}{3 - p^2} \geq 0. \tag{102}$$

Since $-1/R \in [p_*, 0)$, this convex combination preserves $p^2 < 3$ and also preserves $p^2 < 2$ whenever the latter holds initially. Hence $\mathcal{F}_\alpha$ maps $\mathcal{D}$ into itself.

C.3 Exact finite-step evaluation

For the physical step write

$$e_h = e^{-h}, \quad t_h = (1 - e_h^2)/2, \quad \zeta_h = \frac{1 - e_h^3}{e_h t_h}, \quad R_h(p) = \frac{2(\zeta_h + 1 + p)}{\zeta_h - p^2}. \tag{103}$$

Here ζ_h is a scalar arising from the discrete value equation; it is distinct from the time-scale ratio $c_h = \tau_h/h$. The frozen-action physical moments are $\mathbb{E}[X_h \mid x, a] = e_h x$ and $\mathbb{E}[X_h^2 \mid x, a] = e_h^2 x^2 + t_h a^2$. For the state chain under stationary action resampling, let $m_{j,n} = \mathbb{E}_x X_n^j$. Then

$$m_{1,n} = e_h^n x, \qquad m_{2,n+1} = (e_h^2 + t_h p^2) m_{2,n} + 2 t_h p d\, m_{1,n} + t_h(d^2 + \kappa^{-1}). \tag{104}$$

Discounted second-moment summability follows when $e_h(e_h^2 + t_h p^2) < 1$, or $p^2 < \zeta_h$. Since

$$\zeta_h - 3 = \frac{(1 - e_h)(e_h + 2)}{e_h(1 + e_h)} > 0, \tag{105}$$

this holds on $\mathcal{D}$ for every $h > 0$. In particular, $e_h^n \mathbb{E}_x(1 + X_n^2) \to 0$. The exact averaged critic $V_h = v_h x^2/2 + w_h x + z_h$ has

$$\begin{aligned} v_h &= -\frac{2h(1 + p + p^2)}{1 - e_h^3 - e_h t_h p^2}, \\ w_h &= \frac{-h(d + 2pd + 1 + 2p) + e_h t_h v_h p d}{1 - e_h^2}, \\ z_h &= \frac{-h(d^2 + \kappa^{-1} + 2d + \lambda_0 \kappa/2) + (e_h t_h v_h/2)(d^2 + \kappa^{-1})}{1 - e_h}. \end{aligned} \tag{106}$$

These solve $V_h = h\bar{\ell} + e_h \Pi P_h V_h$. Iterating this equation and using the discounted moment bounds identifies V_h with the infinite discounted payoff. Reconstructing $Q_h = h\ell + e_h P_h V_h$ and forming $\partial_a Q_h/(\alpha h)$ gives Eq. (29) with R replaced by R_h.

C.4 Finite action horizons and a simultaneous limit

The operators P_h, S_τ, Π preserve $\mathcal{P}_2$. On the centered subspace, the action eigenvalues are $e^{-\kappa\tau}$ (twice) and $e^{-2\kappa\tau}$. Therefore

$$\chi_\tau^\theta = -\frac{H}{2(1 - e^{-2\kappa\tau})}(y^2 - \kappa^{-1}) + \frac{b(x)}{1 - e^{-\kappa\tau}} y. \tag{107}$$

Suppress θ and write $T = e^{-h}P_hS_\tau$, $J = I - \Pi$, and $Q_{h,\tau} = V_{h,\tau} + w_{h,\tau}$, with $V_{h,\tau} = \Pi Q_{h,\tau}$ and $\Pi w_{h,\tau} = 0$. The polynomial Bellman equation is the block system

$$\begin{pmatrix} I - \Pi T\Pi & -\Pi TJ \\ -JT\Pi & I - JTJ \end{pmatrix} \begin{pmatrix} V_{h,\tau} \\ w_{h,\tau} \end{pmatrix} = h \begin{pmatrix} \Pi\ell \\ J\ell \end{pmatrix}. \tag{108}$$

For fixed $\tau > 0$, the centered block at $h = 0$ is $(I - S_\tau)|_{\ker \Pi}$ and is invertible. The off-diagonal blocks are $O(h)$. After eliminating $w_{h,\tau}$ and dividing the projected equation by h, the state block tends to $1 - \bar{\mathcal{L}}$, whose diagonal on $\mathrm{span}\{1, x, x^2\}$ is $1, 2, 3 - p^2$. This gives

$$V_{h,\tau} = V + O_{\mathrm{pol}}(h), \qquad w_{h,\tau} = h\chi_\tau + O_{\mathrm{pol}}(h^2). \tag{109}$$

Action differentiation is a bounded map on $\mathcal{P}_2$, so the expansion also holds after applying ∂_a.

To identify this polynomial solution with the discounted payoff, write $e = e^{-h}$, $t_h = (1 - e^2)/2$, $r = e^{-\kappa\tau}$, and $s = 1 - r$. One executed transition has the exact representation

$$X' = ex + \sqrt{t_h}\, aZ, \qquad A' = ra + s(pX' + d) + \sqrt{(1 - r^2)/\kappa}\, Z', \tag{110}$$

where Z, Z' are independent standard normal variables. The first-moment matrix is $\left(\begin{smallmatrix} e & 0 \\ spe & r \end{smallmatrix}\right)$. For the second moments $(\mathbb{E}X^2, \mathbb{E}XA, \mathbb{E}A^2)$, the homogeneous block is

$$M_{h,\tau} = \begin{pmatrix} e^2 & 0 & t_h \\ spe^2 & re & spt_h \\ s^2p^2e^2 & 2rspe & r^2 + s^2p^2t_h \end{pmatrix}; \tag{111}$$

the remaining terms involve only first moments and constants. At $h = 0$ its eigenvalues are $1, r, r^2$. The eigenvalue of $e^{-h}M_{h,\tau}$ issuing from 1 is $1 - (3 - p^2)h + O(h^2)$; the other two tend to $r, r^2 < 1$. The discounted constant and first-moment blocks have eigenvalues e, e^2, er. Hence all discounted moment eigenvalues lie inside the unit circle for sufficiently small h, uniformly on compact subsets of $\mathcal{D}$ for fixed $\tau > 0$. It follows that $\sum_{n\geq 0} e^{-nh}\mathbb{E}_{x,a}(1 + X_n^2 + A_n^2) < \infty$. Since $|\ell(x, a)| \leq C_\theta(1 + x^2 + a^2)$, the Bellman inverse is the actual infinite discounted payoff.

For $\tau \downarrow 0$, uniformly on compact subsets of $\mathcal{D}$, $\left\|(I - S_\tau)^{-1}\right\|_{\ker \Pi} = O(\tau^{-1})$. When $h/\tau \to 0$, the centered block perturbation has relative size $O(h/\tau)$, while the projected block after elimination and division by h differs from $1 - \bar{\mathcal{L}}$ by $O(h + h/\tau)$. Consequently,

$$V_{h,\tau} - V = O_{\mathrm{pol}}(h + h/\tau), \qquad w_{h,\tau} = h\chi_\tau + O_{\mathrm{pol}}(h^2/\tau^2 + h^2/\tau). \tag{112}$$

The moment argument remains valid: the fast eigenvalues have gaps of order τ and perturbations of order h, while the slow quadratic eigenvalue is $1 - (3 - p^2)h + O(h^2/\tau)$. Thus discounted moments remain summable when h/τ is sufficiently small. Finally, Eq. (107) gives $\chi_\tau = \tau^{-1}U + q/2 + O_{\mathrm{pol}}(\tau)$, hence

$$\left\| \frac{\tau}{h}\partial_a Q_{h,\tau} - \partial_a U \right\|_{\mathrm{pol}} = O(\tau + h/\tau) \tag{113}$$

as $\tau \to 0$ and $h/\tau \to 0$, uniformly on compact admissible parameter sets. This assertion concerns coefficient norms and does not assert a global unweighted uniform bound on $\mathbb{R}^2$.

When $b \not\equiv 0$, exact matching of $\partial_a\chi_\tau/\alpha$ to $\partial_a U/\lambda_0$ requires both $\alpha = \lambda_0\kappa/(1 - e^{-\kappa\tau})$ and $\alpha = 2\lambda_0\kappa/(1 - e^{-2\kappa\tau})$. Their ratio is $2/(1 + e^{-\kappa\tau}) > 1$, verifying the finite-horizon obstruction to a common scalar.

Proof of Proposition 5.1. By (30), the slice $\mathcal{M}$ is invariant under $\mathcal{F}_\alpha$, and $b \equiv 0$ on $\mathcal{M}$. Hence only the quadratic action mode is active, with relaxation rate 2κ. Proposition 4.2 therefore implies that $\alpha = 2\lambda_0\kappa_0$ gives exact Poisson calibration at θ_0: $F_0(\Psi_{\theta_0}) = B^{\theta_0}$.

Using (32), the first precision update is

$$\kappa_1 = \frac{R_* + \lambda_0\kappa_0^2}{2\lambda_0\kappa_0} = \frac{1}{2}\left(\kappa_0 + \frac{R_*}{\lambda_0\kappa_0}\right). \tag{114}$$

Along $\mathcal{M}$, the only κ-dependent part of the averaged value is $-\frac{R_*}{2\kappa} - \frac{\lambda_0\kappa}{2}$. Substituting (114) and subtracting the value at κ_0 gives

$$V^{\theta_1}(x) - V^{\theta_0}(x) = \frac{\left(R_* - \lambda_0\kappa_0^2\right)^2}{4\kappa_0\left(R_* + \lambda_0\kappa_0^2\right)}. \tag{115}$$

Since $\kappa_0 < \sqrt{R_*/\lambda_0}$, the numerator is nonzero, proving the strict first-step improvement.

It remains to show that repeated reuse of the same scalar diverges. From (32),

$$\kappa_{k+1} - \kappa_k = \frac{\lambda_0\kappa_k^2 - \alpha\kappa_k + R_*}{\alpha} = \frac{\lambda_0(\kappa_k - \kappa_0)^2 + R_* - \lambda_0\kappa_0^2}{\alpha}, \tag{116}$$

where the second equality uses $\alpha = 2\lambda_0\kappa_0$. Set $\delta := R_* - \lambda_0\kappa_0^2 > 0$. Then $\kappa_{k+1} - \kappa_k \geq \frac{\delta}{\alpha} > 0$, and therefore $\kappa_k \geq \kappa_0 + k\frac{\delta}{\alpha} \longrightarrow \infty$.

Finally, along $\mathcal{M}$ the averaged value contains the term $-\lambda_0\kappa/2$, while all remaining κ-dependent terms are bounded above as $\kappa \to \infty$. Hence $V^{\theta_k}(x) \longrightarrow -\infty$ for every fixed x. □

D EXACT STATIONARY-SAMPLING CRITIC FOR THE LQ IMPLEMENTATION

This appendix gives the transition, population equation, and coefficient diagnostics underlying Figure 1 and Table 1. All calculations use the Euler state transition and one-step martingale residual.

D.1 REFERENCE POLICY AND SAMPLING SCHEME

Fixed reference policy. The reference actor is

$$\Psi(x, a) = -\kappa(a - px - d), \tag{117}$$

with

$$\begin{aligned} \kappa &= e^{1.52913155} \approx 4.614168, \\ p = \frac{-1.5119060}{\kappa} \approx -0.327666, &\qquad d = \frac{-3.5624157}{\kappa} \approx -0.772060. \end{aligned} \tag{118}$$

These parameters originate from the previously reported $c = 1$ actor, but here they are used only to define a fixed evaluation policy. In particular, the corresponding $c = 1$ critic is not used as the theoretical critic for stationary sampling.

The running reward is

$$r(x, a) = -\left(x^2 + xa + a^2 + x + 2a\right), \tag{119}$$

and the penalized running payoff is

$$\ell^\Psi(x, a) = r(x, a) - \frac{\lambda_0}{2}\Psi(x, a)^2, \qquad \lambda_0 = 0.1. \tag{120}$$

The discount rate in the martingale residual is $\beta = 1$.

One-step transition under stationary sampling. The physical state dynamics used in the experiment have

$$A = -1, \qquad B = 0, \qquad C = 0, \qquad D = 1. \tag{121}$$

Thus one Euler step of length h from (x, a) is

$$X' = (1 - h)x + a\sqrt{h}\,\xi, \qquad \xi \sim N(0, 1). \tag{122}$$

Hence

$$m_1(x, a) := \mathbb{E}[X' \mid x, a] = (1 - h)x, \tag{123}$$

$$m_2(x, a) := \mathbb{E}[(X')^2 \mid x, a] = (1 - h)^2x^2 + ha^2. \tag{124}$$

After the state update, the next action is sampled directly from the conditional stationary distribution of the fixed affine policy:

$$A' \mid X' = x' \sim N\left(px' + d,\, \kappa^{-1}\right). \tag{125}$$

The draw is conditionally independent of the previous action given X'. In the $c = 1$ execution regime, by contrast, the frozen action diffusion runs only for a horizon $\tau = h$ and generally retains dependence on the current action. Stationary resampling replaces that finite-time transition by the invariant projection Π^Ψ. Although this conditional invariant law does not depend on h, the Euler state transition and martingale residual do, so the population critic must be solved separately for each h.

Consequently,

$$\mathbb{E}[A' \mid x, a] = p\, m_1 + d, \tag{126}$$

$$\mathbb{E}[(A')^2 \mid x, a] = p^2 m_2 + 2pd\, m_1 + d^2 + \kappa^{-1}, \tag{127}$$

$$\mathbb{E}[X'A' \mid x, a] = p\, m_2 + d\, m_1. \tag{128}$$

D.2 Exact population critic

Population martingale equation. We seek a quadratic critic of the form

$$Q_h(x, a) = \frac{1}{2}\theta_{0,h}x^2 + \theta_{1,h}x + \frac{1}{2}\theta_{2,h}a^2 + \theta_{3,h}a + \theta_{4,h}xa + \theta_{5,h}. \tag{129}$$

For compactness, write $\boldsymbol{\theta}_h = (\theta_{0,h}, \theta_{1,h}, \theta_{2,h}, \theta_{3,h}, \theta_{4,h}, \theta_{5,h})^\top$. Using Eqs. (123)–(128), the conditional expectation of the next critic is

$$\begin{aligned}\mathbb{E}[Q_h(X', A') \mid x, a] = \frac{1}{2}\theta_{0,h}m_2 + \theta_{1,h}m_1 \\ + \frac{1}{2}\theta_{2,h}\left(p^2 m_2 + 2pd\, m_1 + d^2 + \kappa^{-1}\right) \\ + \theta_{3,h}(pm_1 + d) + \theta_{4,h}(pm_2 + dm_1) + \theta_{5,h}.\end{aligned} \tag{130}$$

The implementation uses the one-step martingale residual

$$\delta_h = Q_h(X', A') - Q_h(x, a) + h\left(\ell^\Psi(x, a) - \beta Q_h(x, a)\right). \tag{131}$$

The exact population critic is therefore characterized by

$$\mathbb{E}[\delta_h \mid x, a] = 0 \qquad \text{for all } (x, a). \tag{132}$$

Equivalently,

$$\left[(1 + \beta h)I - \mathcal{K}_h^\Psi\right] Q_h = h\ell^\Psi, \tag{133}$$

where $\mathcal{K}_h^\Psi$ denotes one Euler state step followed by conditional stationary action sampling.

Coefficient system. Substituting Eq. (130) into Eq. (132) and matching the coefficients of $x^2, x, a^2, a, xa, 1$ gives the linear system

$$A_h^{\mathrm{st}}\boldsymbol{\theta}_h = b_h, \tag{134}$$

with

$$A_h^{\mathrm{st}} = \begin{pmatrix} \frac{h^2}{2} - \frac{h\rho}{2} - h & 0 & \frac{h^2p^2}{2} - hp^2 + \frac{p^2}{2} & 0 & h^2p - 2hp + p & 0 \\ 0 & -h\beta - h & -dhp + dp & -hp + p & -dh + d & 0 \\ \frac{h}{2} & 0 & \frac{hp^2}{2} - \frac{h\beta}{2} - \frac{1}{2} & 0 & hp & 0 \\ 0 & 0 & 0 & -h\beta - 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -h\beta - 1 & 0 \\ 0 & 0 & \frac{d^2}{2} + \frac{1}{2\kappa} & d & 0 & -h\beta \end{pmatrix}, \tag{135}$$

and

$$b_h = \begin{pmatrix} \frac{h\kappa^2\lambda_0 p^2}{2} + h \\ dh\kappa^2\lambda_0 p + h \\ \frac{h\kappa^2\lambda_0}{2} + h \\ -dh\kappa^2\lambda_0 + 2h \\ -h\kappa^2\lambda_0 p + h \\ \frac{d^2 h\kappa^2\lambda_0}{2} \end{pmatrix}. \tag{136}$$

Thus the theoretical coefficients in Table 1 are obtained directly from

$$\boldsymbol{\theta}_h = (A_h^{\mathrm{st}})^{-1} b_h. \tag{137}$$

D.3 SMALL-STEP EXPANSION AND NUMERICAL DIAGNOSTICS

Small-step action-gradient limit. The fourth and fifth rows of Eq. (134) can be solved explicitly:

$$\theta_{3,h} = \frac{h(\lambda_0\kappa^2 d - 2)}{1 + \beta h}, \tag{138}$$

$$\theta_{4,h} = \frac{h(\lambda_0\kappa^2 p - 1)}{1 + \beta h}. \tag{139}$$

Hence

$$\frac{\theta_{3,h}}{h} = \lambda_0\kappa^2 d - 2 + O(h), \tag{140}$$

$$\frac{\theta_{4,h}}{h} = \lambda_0\kappa^2 p - 1 + O(h). \tag{141}$$

The third row gives

$$\frac{\theta_{2,h}}{h} = \frac{\theta_{0,h} - \lambda_0\kappa^2 - 2 + 2p\theta_{4,h}}{1 + h(\beta - p^2)}. \tag{142}$$

The state-only coefficient equations imply $\theta_{0,h} = v + O(h)$. Eqs. (139) and (142) yield

$$\frac{\theta_{2,h}}{h} = -H + O(h). \tag{143}$$

Therefore

$$\frac{1}{h}\partial_a Q_h(x,a) = -Ha + (\lambda_0\kappa^2 p - 1)x + (\lambda_0\kappa^2 d - 2) + O(h), \tag{144}$$

Equivalently, for each fixed (x, a),

$$\partial_a Q_h^{\Psi}(x,a) = h\,\partial_a q^{\Psi}(x,a) + O(h^2). \tag{145}$$

For the fixed reference policy,

$$H = 4.668147, \qquad \lambda_0\kappa^2 p - 1 = -1.697619, \qquad \lambda_0\kappa^2 d - 2 = -3.643758, \tag{146}$$

and hence

$$\partial_a q^{\Psi}(x,a) = -4.668147\,a - 1.697619\,x - 3.643758. \tag{147}$$

Action-gradient scaling. The quadratic critic has action gradient $\partial_a Q_h = \theta_{2,h}a + \theta_{4,h}x + \theta_{3,h}$. We measure its scale by the coefficient norm

$$S_h = \left(\theta_{2,h}^2 + \theta_{4,h}^2 + \theta_{3,h}^2\right)^{1/2}. \tag{148}$$

Equivalently, under the fixed reference measure $\nu_0 = \mathcal{N}(0,1)_x \otimes \mathcal{N}(0,1)_a$,

$$S_h = \|\partial_a Q_h\|_{L^2(\nu_0)} = G_Q(h;\nu_0). \tag{149}$$

The error of the rescaled coefficients relative to Eq. (147) is

$$E_2(h) := \left[\left(\frac{\theta_{2,h}}{h} + 4.668147\right)^2 + \left(\frac{\theta_{4,h}}{h} + 1.697619\right)^2 + \left(\frac{\theta_{3,h}}{h} + 3.643758\right)^2\right]^{1/2}. \tag{150}$$

Table 1 gives the coefficients and errors underlying Figure 1. The error decreases from 0.5469 at $h = 0.1$ to 0.0300 at $h = 0.005$, and reaches 6.01×10^{-3} at $h = 0.001$. The log–log fits over $h \le 0.05$ have slopes 0.9892 for S_h and 0.9889 for $E_2(h)$, consistent with first-order decay of both quantities. Thus a fixed actor denominator leaves the leading action signal vanishing as $h \downarrow 0$, whereas a denominator proportional to h restores its leading-order scale.

Table 1: Exact population critic under stationary action sampling for the fixed reference policy. The action-gradient coefficients are asymptotically linear in the physical step h, while their $1/h$ rescaling converges to the coefficients of $\partial_a q^{\Psi}$.

h	$\theta_{2,h}$	$\theta_{4,h}$	$\theta_{3,h}$	$\theta_{2,h}/h$	$\theta_{4,h}/h$	$\theta_{3,h}/h$	$E_2(h)$
0.1	−0.426132	−0.154329	−0.331251	−4.261319	−1.543290	−3.312508	0.546858
0.05	−0.222744	−0.080839	−0.173512	−4.454885	−1.616780	−3.470246	0.286570
0.02	−0.091606	−0.033287	−0.071446	−4.580295	−1.664332	−3.572312	0.118028
0.01	−0.046238	−0.016808	−0.036077	−4.623779	−1.680811	−3.607682	0.059603
0.005	−0.023229	−0.008446	−0.018128	−4.645851	−1.689173	−3.625630	0.029951
$h \downarrow 0$	–	–	–	−4.668147	−1.697619	−3.643758	0

E Experimental details and additional results

This appendix reports the complete numerical results for the LQ experiments and the MuJoCo control tasks. The LQ experiments use stationary action sampling, whereas the MuJoCo experiments compare QSM under uniform and irregular decision intervals.

Implementation details and numerical safeguards. All joint actor–critic experiments use the same LQ implementation and keep the objective coefficient fixed at $\lambda_0 = 0.1$. The critic and actor are updated once per environment step with base step sizes $\eta_Q = \eta_\Psi = 0.01$ and the common schedule

$$\eta_k = \frac{0.01}{\max\{\log((k+1)h),\, 1\}}. \tag{151}$$

Each parameter-update component is clipped to $[-20, 20]$ before applying the step. All reported experiments use a fixed physical training horizon $T = 10^4$ and ten random seeds. Within each fixed h, compared methods use paired random streams. Clipping is infrequent in all experiments, with the maximum recorded clipping rate below 0.7%.

A run is marked invalid when any simulated state or action, reward, martingale residual, critic parameter, or actor parameter becomes non-finite. The failure step is recorded explicitly. Among the fixed-versus-scaled runs, one configuration ($h = 0.01$, fixed normalization, seed 5) becomes invalid at step 283544; all remaining runs are finite. The α-sensitivity experiment contains no invalid runs.

E.1 Learned critic under small physical steps

In this section, we train the martingale critic with the actor frozen. We examine whether the stochastic CQSM critic can recover the shrinking action signal. We fix the reference policy, disable actor updates, and train only the martingale critic under $h \in \{0.1, 0.05, 0.02, 0.01\}$. For each h, the exact finite-step critic derived in Appendix D is used as the reference.

Let

$$g_h^{\mathrm{exact}} = (\theta_{2,h}^{\mathrm{exact}}, \theta_{4,h}^{\mathrm{exact}}, \theta_{3,h}^{\mathrm{exact}}), \qquad \widehat{g}_h = (\widehat{\theta}_{2,h}, \widehat{\theta}_{4,h}, \widehat{\theta}_{3,h}), \tag{152}$$

and define

$$S_h^{\mathrm{exact}} = \|g_h^{\mathrm{exact}}\|_2, \qquad S_h^{\mathrm{learn}} = \|\widehat{g}_h\|_2, \tag{153}$$

together with the relative coefficient error

$$E_{\mathrm{rel}}(h) = \frac{\|\widehat{g}_h - g_h^{\mathrm{exact}}\|_2}{\|g_h^{\mathrm{exact}}\|_2}. \tag{154}$$

As shown in Table 2, the learned critic tracks the shrinking signal magnitude closely: the ratio $S_h^{\mathrm{learn}}/S_h^{\mathrm{exact}}$ improves from 0.87 at $h = 0.1$ to approximately one for $h \le 0.05$. The relative error is smallest for the intermediate steps, but rises again from 7.2% at $h = 0.02$ to 11.9% at $h = 0.01$. Thus the critic can recover the $O(h)$ signal magnitude, while maintaining proportional accuracy becomes more difficult at the smallest tested step. This distinction is relevant when the actor subsequently rescales the critic gradient by a factor of order $1/h$.

Table 2: Learned critic accuracy under the fixed reference policy. S_h^{exact} is computed from the exact population critic, while the learned quantities are mean $\pm$ standard deviation over ten seeds.

h	S_h^{exact}	S_h^{learn}	$E_{\mathrm{rel}}(h)$
0.10	0.561367	0.488460 ± 0.007772	0.143559 ± 0.019209
0.05	0.293695	0.286503 ± 0.007255	0.062064 ± 0.022280
0.02	0.120848	0.119904 ± 0.005232	0.072372 ± 0.024614
0.01	0.061008	0.062063 ± 0.003838	0.119432 ± 0.053145

E.2 Actor normalization under small physical steps

We use joint actor–critic learning and examine whether the small-h degradation can be mitigated by correcting the scale of the actor target. Importantly, we separate the regularization coefficient in the control objective from the actor-side normalization. The former is kept fixed at $\lambda_0 = 0.1$ in the martingale critic for all experiments, while the actor update uses

$$\Psi_{\mathrm{tar}}(x, a) = \frac{\partial_a Q(x, a)}{\lambda_A(h)}. \tag{155}$$

We compare the original fixed normalization $\lambda_A(h) = 0.1$ with the step-scaled choice $\lambda_A(h) = h$. We use $h \in \{0.1, 0.05, 0.02, 0.01\}$, and report results over ten seeds.

We evaluate the environment reward over the final 10% of the physical training horizon. To diagnose the actor signal directly, we also report the corresponding actor-target magnitude

$$G_A(h) := \mathbb{E}\left[\left|\frac{\partial_a Q(X_t, A_t)}{\lambda_A(h)}\right|\right], \tag{156}$$

averaged over the same final window.

Figure 2 compares the two normalizations; Table 3 reports the numerical values. Under the original fixed denominator, the environment reward deteriorates rapidly as h decreases, from 0.227 at $h = 0.1$ to -4.167 at $h = 0.01$. At the same time, the actor-target magnitude decreases from 1.437 to 0.495, consistent with the $O(h)$ attenuation of $\partial_a Q_h$ identified above.

In contrast, the scaled choice $\lambda_A(h) = h$ substantially improves the environment reward at all smaller steps. In particular, the final reward changes from -0.399 to 0.375 at $h = 0.05$ and from -1.958 to 0.521 at $h = 0.02$. At $h = 0.01$, the scaled update yields a higher online environment reward, but its frozen-policy full discounted value is lower, as shown in Table 5.

The actor-target measurements in Figure 2(b) show a corresponding change in update magnitude. While the fixed normalization increasingly suppresses the actor signal, the scaled normalization restores and eventually amplifies it: $G_A(h)$ increases from 1.437 at $h = 0.1$ to 6.793 at $h = 0.01$. Thus, the h-scaling corrects the vanishing-signal problem, but the resulting repeated actor–critic interaction can become overly aggressive at very small h. This supports the distinction between restoring the correct signal scale and guaranteeing stable repeated learning, and motivates the more general calibration $\lambda_A(h) = \alpha h$ studied next.

Table 3: Effect of actor-side normalization under different physical steps. Reward is the environment reward averaged over the final 10% of the physical training horizon, and G_A is the corresponding actor-target magnitude in Eq. (156). Results are mean $\pm$ standard deviation over ten seeds, except for the fixed-normalization result at $h = 0.01$, where one run became invalid and the reported finite statistics use the remaining nine runs.

	Final environment reward		$G_A(h)$	
h	$\lambda_A = 0.1$	$\lambda_A = h$	$\lambda_A = 0.1$	$\lambda_A = h$
0.10	0.227 ± 0.024	0.227 ± 0.024	1.437 ± 0.030	1.437 ± 0.030
0.05	-0.399 ± 0.051	0.375 ± 0.029	0.944 ± 0.029	1.753 ± 0.064
0.02	-1.958 ± 0.057	0.521 ± 0.080	0.629 ± 0.014	3.057 ± 0.521
0.01	-4.167 ± 0.114	0.086 ± 0.011	0.495 ± 0.015	6.793 ± 0.333

E.3 Sensitivity to the calibration coefficient

The preceding experiment shows that the step-scaled normalization $\lambda_A(h) = h$ restores the actor signal, but can become overly aggressive at the smallest physical step. We therefore consider the more general calibration $\lambda_A(h) = \alpha h$, where $\alpha > 0$ controls the strength of the normalized actor update without changing the $O(h)$ scaling itself.

We test $\alpha \in \{1, 1.25, 1.5, 2\}$ at $h \in \{0.02, 0.01\}$, while keeping the critic objective coefficient fixed at $\lambda_0 = 0.1$. Each configuration is evaluated over ten paired seeds for a fixed physical training horizon $T = 10^4$. As before, we report the environment reward over the final 10% of training and the actor-target magnitude

$$G_A(h, \alpha) = \mathbb{E}\left[\left|\frac{\partial_a Q(X_t, A_t)}{\alpha h}\right|\right]. \tag{157}$$

Table 4 reveals two distinct regimes. At $h = 0.02$, the unmodified scaling $\alpha = 1$ already provides a sufficient actor signal: increasing α monotonically reduces G_A from 3.06 to 1.14 and simultaneously lowers the final environment reward from 0.521 to -0.174. At $h = 0.02$, increasing α reduces the final online environment reward over the tested range. However, the frozen-policy evaluation in Table 6 shows that increasing α from 1 to 1.25 improves the full discounted value from -0.248 to 0.215; calibration should therefore be assessed using the penalized objective rather than environment reward alone.

At the smaller step $h = 0.01$, however, $\alpha = 1$ produces a much stronger actor target, with $G_A = 6.79$, and only achieves a final reward of 0.086. Moderate calibration substantially improves the result: $\alpha = 1.25$ and $\alpha = 1.5$ reduce the actor-target magnitude to 4.19 and 2.57, respectively, while increasing the reward to 0.323 and 0.370. Further increasing the coefficient to $\alpha = 2$ suppresses the actor target to 1.68, but the reward falls again to 0.085.

These results indicate that the h factor and the coefficient α play different roles. The scaling $\lambda_A(h) \propto h$ corrects the leading-order attenuation of the critic action signal, whereas α controls the strength of the resulting repeated actor–critic interaction. Too little damping can produce an increasingly aggressive actor update at very small h, while excessive damping again weakens learning. Hence the scalar coefficient should be interpreted as a calibration parameter rather than as part of the asymptotic h-scaling law.

Calibration dynamics. Figure 3 supplements Table 4 with training trajectories at $h = 0.01$. Each point is averaged over the interval since the preceding checkpoint. In particular, the point at $T = 10^4$ summarizes $[7500, 10000]$, whereas Table 4 uses the final-10% window $[9000, 10000]$. The initialization placeholders at $t = 0$ are omitted.

Table 4: Sensitivity to the actor-side calibration coefficient in $\lambda_A(h) = \alpha h$. Final reward denotes the environment reward averaged over the final 10% of the physical training horizon, and G_A is the corresponding actor-target magnitude. Results are mean $\pm$ standard deviation over ten seeds.

h	α	Final environment reward	G_A
0.02	1.00	$\mathbf{0.521 \pm 0.080}$	3.057 ± 0.521
	1.25	0.409 ± 0.088	2.000 ± 0.330
	1.50	0.157 ± 0.108	1.467 ± 0.181
	2.00	-0.174 ± 0.092	1.141 ± 0.079
0.01	1.00	0.086 ± 0.011	6.793 ± 0.333
	1.25	0.323 ± 0.118	4.188 ± 1.131
	1.50	$\mathbf{0.370 \pm 0.123}$	2.568 ± 0.992
	2.00	0.085 ± 0.195	1.682 ± 0.550

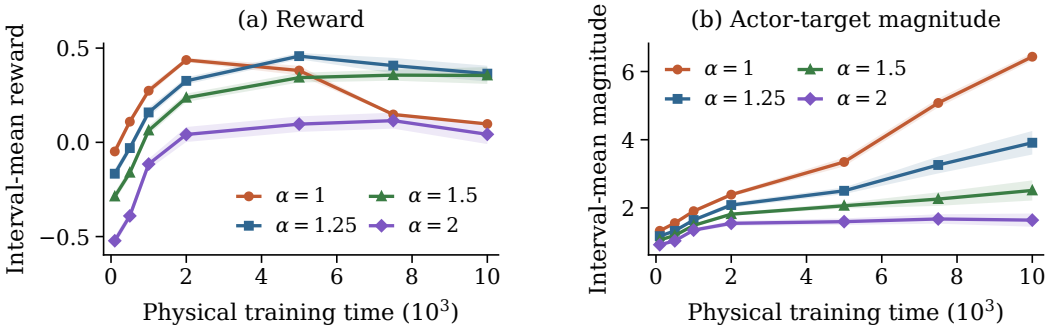

Figure 3: Calibration dynamics at $h = 0.01$ with $\lambda_A = \alpha h$ and $\lambda_0 = 0.1$. (a) Environment reward. (b) Actor-target magnitude. Points show interval averages, further averaged over ten paired seeds; shaded bands indicate one standard error across seeds.

E.4 Objective-consistent evaluation of learned policies

The online reward reported during training is not itself the discounted control objective studied in the theory. We therefore perform a post-processing evaluation of every learned final policy.

Recovering the learned affine policy. The actor parameterization used in the experiments is

$$\Psi^\theta(x,a) = -\kappa a + v_1 x + v_2 = -\kappa(a - px - d), \tag{158}$$

where $\kappa = e^{v_0}, p = \frac{v_1}{\kappa}, d = \frac{v_2}{\kappa}$. Accordingly, stationary action resampling gives $A_n \mid X_n = x \sim \mathcal{N}(px + d, \kappa^{-1})$. The reconstruction (158) is verified numerically for every retained run.

Discounted value under the implemented Euler dynamics. The learning experiments use the Euler state transition

$$X_{n+1} = (1-h)X_n + \sqrt{h}\, A_n \xi_n, \qquad \xi_n \sim \mathcal{N}(0,1), \tag{159}$$

with a fresh stationary action draw at each step. To evaluate the same objective as in the theoretical analysis, we define

$$V^\theta_{\mathrm{E},h}(x) := \mathbb{E}^\theta_x \left[\sum_{n=0}^{\infty} h e^{-nh} \left(r(X_n, A_n) - \frac{\lambda_0}{2} |\Psi^\theta(X_n, A_n)|^2 \right) \right]. \tag{160}$$

Here $\beta = 1$ and $\lambda_0 = 0.1$ in all experiments. Let $\gamma_h = e^{-h}, g_h = 1 - h, A_h = g_h^2 + hp^2$. For the LQ reward used in the experiments, the raw discounted value has the quadratic form

$$V^\theta_{\mathrm{E},h,\mathrm{raw}}(x) = \frac{1}{2} v_{\mathrm{E},h} x^2 + w_{\mathrm{E},h} x + z^{\mathrm{raw}}_{\mathrm{E},h}, \tag{161}$$

with

$$v_{\mathrm{E},h} = -\frac{2h(1 + p + p^2)}{1 - \gamma_h A_h}, \tag{162}$$

$$w_{\mathrm{E},h} = \frac{-h(d + 2pd + 1 + 2p) + \gamma_h h v_{\mathrm{E},h} pd}{1 - \gamma_h g_h}, \tag{163}$$

$$z^{\mathrm{raw}}_{\mathrm{E},h} = \frac{-h(d^2 + \kappa^{-1} + 2d) + \frac{1}{2}\gamma_h h v_{\mathrm{E},h}(d^2 + \kappa^{-1})}{1 - \gamma_h}. \tag{164}$$

Since $\mathbb{E}[|\Psi^\theta(x, A)|^2 \mid x] = \kappa$, the discounted control-cost contribution is independent of the state:

$$C^\theta_{\mathrm{E},h} = \frac{h\lambda_0\kappa}{2(1 - e^{-h})}. \tag{165}$$

Hence $V^\theta_{\mathrm{E},h}(x) = V^\theta_{\mathrm{E},h,\mathrm{raw}}(x) - C^\theta_{\mathrm{E},h}$. All reported values use $x_0 = 0$.

Continuous-time reference. For comparison, we also evaluate the continuous averaged value derived in Eq. (97). Its raw component is

$$V^\theta_{\mathrm{raw}}(x) = \frac{1}{2} v x^2 + w x + z_{\mathrm{raw}}, \tag{166}$$

where

$$v = -\frac{2(1 + p + p^2)}{3 - p^2}, \tag{167}$$

$$w = \frac{1}{2}\left[-(d + 2pd + 1 + 2p) + vpd\right], \tag{168}$$

$$z_{\mathrm{raw}} = -(d^2 + \kappa^{-1} + 2d) + \frac{v}{2}(d^2 + \kappa^{-1}). \tag{169}$$

The corresponding control cost is

$$C^\theta = \frac{\lambda_0\kappa}{2}, \qquad V^\theta = V^\theta_{\mathrm{raw}} - C^\theta. \tag{170}$$

We additionally evaluate the exact physical-transition formula in Eq. (106) as a diagnostic. It is not used as the primary learning metric. Both finite-step evaluations converge to (170) as $h \to 0$.

h	Normalization	n	$V^{\mathrm{E}}_{\mathrm{raw}}(0)$	C^{E}	$V^{\mathrm{E}}_{\mathrm{full}}(0)$
0.01	$\lambda_A = 0.1$	9	-3.3944 ± 0.5691	0.0163 ± 0.0029	-3.4107 ± 0.5662
0.01	$\lambda_A = h$	10	0.0838 ± 0.0124	3.9115 ± 0.3789	-3.8276 ± 0.3888
0.02	$\lambda_A = 0.1$	10	-1.5091 ± 0.2120	0.0294 ± 0.0031	-1.5385 ± 0.2090
0.02	$\lambda_A = h$	10	0.5695 ± 0.1134	0.8174 ± 0.3126	-0.2479 ± 0.4153
0.05	$\lambda_A = 0.1$	10	-0.1917 ± 0.1010	0.0709 ± 0.0085	-0.2626 ± 0.0931
0.05	$\lambda_A = h$	10	0.5217 ± 0.0216	0.2499 ± 0.0258	0.2718 ± 0.0117
0.10	$\lambda_A = 0.1$	10	0.4061 ± 0.0257	0.1729 ± 0.0173	0.2331 ± 0.0112
0.10	$\lambda_A = h$	10	0.4061 ± 0.0257	0.1729 ± 0.0173	0.2331 ± 0.0112

Table 5: Frozen final-policy discounted values at $x_0 = 0$, $\beta = 1$, and $\lambda_0 = 0.1$, evaluated under the Euler state transition used in learning. The positive control cost C^{E} is subtracted from the raw value. Entries are mean $\pm$ population standard deviation.

h	α	n	$V^{\mathrm{E}}_{\mathrm{raw}}(0)$	C^{E}	$V^{\mathrm{E}}_{\mathrm{full}}(0)$
0.01	1.00	10	0.0838 ± 0.0124	3.9115 ± 0.3789	-3.8276 ± 0.3888
0.01	1.25	10	0.3250 ± 0.1601	1.6250 ± 0.8345	-1.2999 ± 0.9703
0.01	1.50	10	0.4834 ± 0.1139	0.5852 ± 0.5495	-0.1018 ± 0.6097
0.01	2.00	10	0.1241 ± 0.4227	0.2327 ± 0.1617	-0.1086 ± 0.3368
0.02	1.00	10	0.5695 ± 0.1134	0.8174 ± 0.3126	-0.2479 ± 0.4153
0.02	1.25	10	0.5470 ± 0.0835	0.3322 ± 0.1163	0.2148 ± 0.0634
0.02	1.50	10	0.3370 ± 0.0999	0.1757 ± 0.0427	0.1613 ± 0.0761
0.02	2.00	10	0.0283 ± 0.0810	0.1138 ± 0.0286	-0.0855 ± 0.0851

Table 6: Frozen final-policy value decomposition for $\lambda_A = \alpha h$. The positive control cost C^{E} is subtracted, so $V^{\mathrm{E}}_{\mathrm{full}} = V^{\mathrm{E}}_{\mathrm{raw}} - C^{\mathrm{E}}$. Entries are mean $\pm$ population standard deviation over ten seeds. The tested values illustrate the calibration tradeoff and do not imply a universally optimal α.

Fixed versus step-scaled normalization. Table 5 reports the frozen final-policy values for the fixed and step-scaled actor normalizations. At $h = 0.05$ and $h = 0.02$, scaling the actor denominator with h substantially increases the full discounted value. At $h = 0.01$, however, the raw discounted value increases from -3.394 to 0.084, while the control cost rises from 0.016 to 3.911. Consequently, the full value decreases from -3.411 to -3.828. This provides a direct empirical distinction between restoring the critic-signal scale and calibrating the resulting policy update.

Sensitivity to scalar calibration. Table 6 decomposes the value under $\lambda_A = \alpha h$. The results show that α affects both the raw discounted value and the cost of the learned drift. At $h = 0.02$, the full value changes from -0.248 at $\alpha = 1$ to 0.215 at $\alpha = 1.25$; at $h = 0.01$, it changes from -3.828 at $\alpha = 1$ to approximately -0.102 at $\alpha = 1.5$. Increasing α further reduces the control cost but can also reduce the raw-value gain. Thus these experiments illustrate a calibration tradeoff rather than a universally optimal scalar coefficient.

E.5 Exact mechanism validation

We provide two exact LQ diagnostics for the mechanisms underlying the main results. The first varies the action-sampling horizon to distinguish different time-scale regimes, while the second compares scalar-normalized little-q fields with the exact Poisson improvement field. Both experiments use closed-form LQ quantities and involve no actor or critic training.

Finite action-sampling horizons. We first examine how the action-sampling horizon affects the leading action-dependent component of the finite-step critic. We use the fixed affine reference policy from Appendix D, $\kappa \approx 4.614168, p \approx -0.327666, d \approx -0.772060$, and set $x_{\mathrm{ref}} = 0$. For this policy, $H \approx 4.668147, b(x_{\mathrm{ref}}) \approx -0.039668$, so both centered action modes are active.

From Eq. (107), the finite-horizon corrector is

$$\chi^{\theta}_{\tau} = -\frac{H}{2(1 - e^{-2\kappa\tau})}\left(y^2 - \kappa^{-1}\right) + \frac{b(x)}{1 - e^{-\kappa\tau}}y, \qquad y = a - (px + d). \tag{171}$$

Action horizon	Predicted order	Fitted slope
$\tau_h = h$	$O(1)$	0.008
$\tau_h = \sqrt{h}$	$O(h^{1/2})$	0.578
$\tau_h = 0.1$	$O(h)$	1.000
stationary	$O(h)$	1.000

Table 7: Scaling of the exact finite-horizon corrector signal $h\chi^\theta_{\tau_h}$. Slopes are fitted on the log–log scale over $h \le 10^{-2}$.

At $x = x_{\mathrm{ref}}$, we measure the scale of the leading action-dependent critic component by the coefficient norm

$$\left\|h\chi^\theta_{\tau_h}\right\|_{\mathrm{pol}} := \left[\left(\frac{hH}{2(1-e^{-2\kappa\tau_h})}\right)^2 + \left(\frac{h\,b(x_{\mathrm{ref}})}{1-e^{-\kappa\tau_h}}\right)^2\right]^{1/2}. \tag{172}$$

For stationary resampling, $\chi^\theta_\tau \to q^\theta$ as $\tau \to \infty$, giving

$$\left\|hq^\theta\right\|_{\mathrm{pol}} = h\left[\left(\frac{H}{2}\right)^2 + b(x_{\mathrm{ref}})^2\right]^{1/2}. \tag{173}$$

We evaluate four action-sampling regimes, $\tau_h = h, \tau_h = \sqrt{h}, \tau_h = 0.1$, and stationary resampling over a logarithmic grid $h \in [10^{-4}, 10^{-1}]$. The predicted asymptotic orders are

$$\left\|h\chi^\theta_h\right\|_{\mathrm{pol}} = O(1), \left\|h\chi^\theta_{\sqrt{h}}\right\|_{\mathrm{pol}} = O(h^{1/2}), \left\|h\chi^\theta_{0.1}\right\|_{\mathrm{pol}} = O(h), \left\|hq^\theta\right\|_{\mathrm{pol}} = O(h). \tag{174}$$

The fitted log–log slopes over $h \le 10^{-2}$, reported in Table 7, are consistent with these predictions.

Importantly, both $\tau_h = \sqrt{h}$ and $\tau_h = 0.1$ satisfy $\tau_h/h \to \infty$ as $h \to 0$, yet the former also satisfies $\tau_h \to 0$ and yields an $O(h^{1/2})$ signal, whereas the latter has a fixed positive horizon and yields an $O(h)$ signal. Hence the relative time-scale ratio alone does not determine the effective action-sampling regime; the absolute behavior of τ_h also matters. This distinction is illustrated in Figure 4(a).

Scalar versus Poisson calibration. We next compare the scalar-normalized little-q field with the exact Poisson improvement field. From Eqs. (27)–(28),

$$F^\theta_\alpha(x,a) := \frac{1}{\alpha}\partial_a q^\theta(x,a) = -\frac{H}{\alpha}y + \frac{b(x)}{\alpha}, \tag{175}$$

$$B^\theta(x,a) := \frac{1}{\lambda_0}\partial_a U^\theta(x,a) = -\frac{H}{2\lambda_0\kappa}y + \frac{b(x)}{\lambda_0\kappa}. \tag{176}$$

At a fixed state x, the stationary action law gives $y \sim \mathcal{N}(0, \kappa^{-1})$. We therefore measure the relative field mismatch by

$$E_{\mathrm{field}}(\alpha) := \frac{\|F^\theta_\alpha - B^\theta\|_{L^2}}{\|B^\theta\|_{L^2}}, \tag{177}$$

where the L^2 norm is taken with respect to this conditional stationary action law. The numerator has the closed form

$$\|F^\theta_\alpha - B^\theta\|^2_{L^2} = \frac{1}{\kappa}\left(-\frac{H}{\alpha} + \frac{H}{2\lambda_0\kappa}\right)^2 + \left(\frac{b(x)}{\alpha} - \frac{b(x)}{\lambda_0\kappa}\right)^2. \tag{178}$$

For comparison across policies, we use the dimensionless scalar $z := \frac{\alpha}{\lambda_0\kappa}$. For the single-mode case, we take the invariant slice $p = p_* = -4 + \sqrt{13}, d = d_* = 2p_*$, with the same κ as above. On this slice, $b(x) \equiv 0$, so only the quadratic action mode with relaxation rate 2κ remains. Consequently, $E_{\mathrm{field}} = 0$ at $z = 2(\alpha = 2\lambda_0\kappa)$, in agreement with Proposition 4.2.

For the two-mode case, we use the reference policy above at $x_{\mathrm{ref}} = 0$, where both H and $b(x_{\mathrm{ref}})$ are nonzero. The quadratic and linear modes require, respectively, $z = 2$ and $z = 1$ for exact matching.

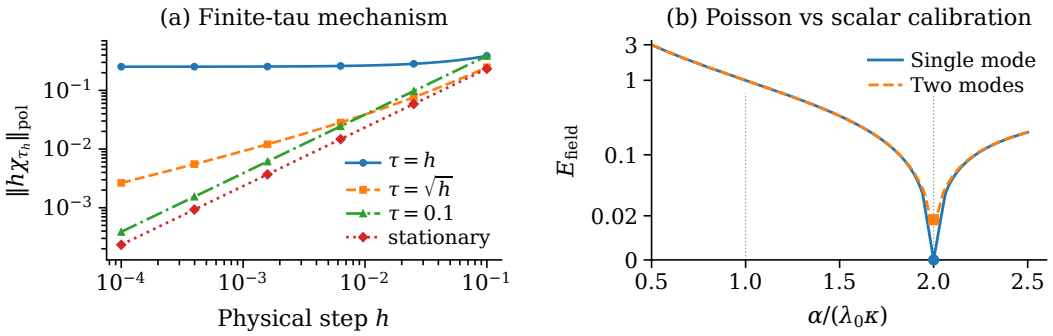

Figure 4: Exact LQ mechanism diagnostics. (a) Scaling of the finite-horizon corrector signal under different action-sampling horizons. Regimes satisfying $\tau_h/h \to \infty$ can still exhibit different asymptotic scales. (b) Relative error between the scalar-normalized little-q field and the Poisson improvement field. Exact matching occurs in the single-mode case, whereas a nonzero mismatch remains when two relaxation modes are active.

Hence no single scalar can reproduce the full Poisson field. On the numerical grid, the relative error is minimized near $z \approx 1.999$, but remains strictly positive, $\min_z E_{\mathrm{field}}(z) \approx 1.82 \times 10^{-2}$. The small residual reflects the relatively weak linear mode for this particular reference policy at $x_{\mathrm{ref}} = 0$; the nonzero minimum nevertheless verifies the multi-mode obstruction. The analytic expression in Eq. (178) also agrees with direct Gaussian integration to numerical precision. Figure 4(b) visualizes the resulting single- and two-mode errors.

E.6 Additional empirical sensitivity under DDPM sampling

We examine whether the temporal sampling schedule affects QSM (Psenka et al., 2024) on Humanoid, Cheetah, and Walker. Let h_k denote the physical time for which the k-th action is held, and let h be the default uniform decision interval. We use the normalized holding time $d_k := \frac{h_k}{h}$. Uniform sampling corresponds to $d_k \equiv 1$, whereas the irregular setting draws different d_k over an episode. We denote the mean normalized holding time of an irregular schedule by $\bar{d}_{\mathrm{irr}} := \mathbb{E}[d_k]$.

Implementation Details Following QSM (Psenka et al., 2024), we generate actions using a five-step DDPM sampler with a variance-preserving noise schedule. This finite-step procedure aims to approximate the conditional stationary action distribution associated with the learned score. The denoising network is parameterized as an MLP with hidden dimensions $[400, 300]$ and ReLU activations, while the twin Q-networks share the same architecture. Fourier features are used to encode the diffusion timestep, and actions are generated via the standard DDPM reverse diffusion process, followed by rescaling to match the environment action bounds. Across all MuJoCo tasks, we use a single set of hyperparameters, including $\lambda^{-1} = 200$ for Cheetah and Humanoid, and $\lambda^{-1} = 80$ for Walker, $\sigma_A = 5.0$, and $\tau = 0.99$. Optimization is performed using Adam with learning rate 3×10^{-4} and batch size 256. The discount factor is set to $\gamma = 0.98$ for Cheetah, Walker, and $\gamma = 0.99$ for Humanoid.The hyperparameters of QSM are chosen following Psenka et al. (2024). For each task, we report the minimum and maximum decision increments $(h_{\min}, h_{\max})$ and the maximum episode duration $T_{\max}$, along with the fractions of small, large, and average steps and the range of number of decision steps per episode. Control-task time is measured in seconds.

Table 8: Irregular-time sampling configuration per task.

Task	$h_{\min}$	$h_{\max}$	% small	% large	% avg	$T_{\max}$	# of steps range
Cheetah	0.002	0.030	89.1%	9.9%	1.0%	10	1200–2000
Walker	0.005	0.075	89.1%	9.9%	1.0%	25	1200–2000
Humanoid	0.010	0.040	40.0%	40.0%	20.0%	25	800–1000

Table 9: The final return of QSM under irregular and uniform decision schedules. Values are mean $\pm$ standard deviation.

Task	$\bar{d}_{\text{irr}}$	Irregular	Uniform
Humanoid	1.000	217.48 ± 191.43	$\mathbf{413.52 \pm 11.77}$
Cheetah	0.491	$\mathbf{775.61 \pm 39.03}$	668.69 ± 54.30
Walker	0.491	$\mathbf{877.00 \pm 6.71}$	636.96 ± 19.03

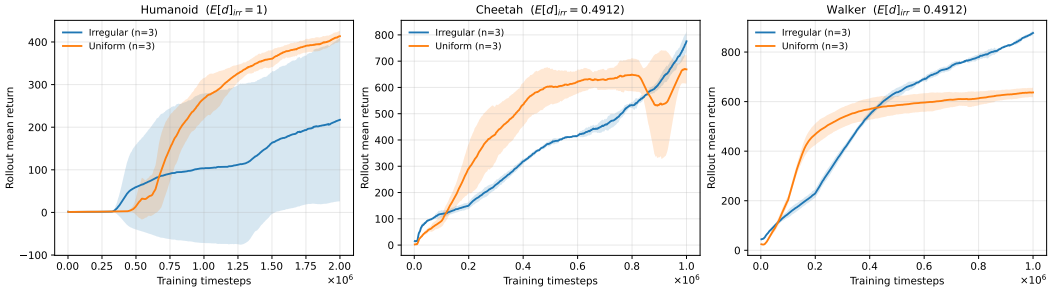

Figure 5: Rollout-mean return during training for QSM with irregular and uniform decision intervals.

We compare irregular and uniform QSM using three seeds, and report the final logged rollout-mean return. The results are summarized in Table 9 and the training curves are shown in Figure 5. The effect of irregular timing is strongly schedule dependent. For Humanoid, the irregular schedule is mean preserving ($\bar{d}_{\text{irr}} = 1$), so it changes the variability of the holding times without changing their mean. In this case, irregular sampling does not improve QSM and performs substantially worse than uniform sampling. In contrast, the Cheetah and Walker schedules are strongly biased toward short holding times, with $\bar{d}_{\text{irr}} \simeq 0.49$. Both tasks obtain substantially higher returns under these irregular schedules. Thus, the experiments do not suggest that irregularity alone is beneficial. Rather, they show that QSM can be highly sensitive to how control decisions are distributed in physical time, and that short-step-dominated multi-scale schedules can be advantageous in some tasks. For Cheetah and Walker, however, $\bar{d}_{\text{irr}} < 1$ also implies a higher average control frequency than under the uniform baseline. Consequently, these experiments compare complete temporal schedules and do not isolate the effect of timing variability from that of the mean holding time. We therefore use them as empirical evidence of time-sampling sensitivity rather than as a causal ablation of irregularity itself.